Reflecting upon Perceptual Speed Tests in Information Retrieval: Limitations, Challenges, and Recommendations

Olivia Foulds
olivia.foulds@strath.ac.uk
Department of Computer and Information Sciences
University of Strathclyde
Glasgow, Scotland

Leif Azzopardi
leif.azzopardi@strath.ac.uk
Department of Computer and Information Sciences
University of Strathclyde
Glasgow, Scotland

Martin Halvey
martin.halvey@strath.ac.uk
Department of Computer and Information Sciences
University of Strathclyde
Glasgow, Scotland

ABSTRACT

Perceptual Speed (PS) is a cognitive ability defined by an individual's accuracy and speed to scan information while completing visual search tasks. Prior studies using PS tests have demonstrated that PS affects multiple factors in Information Retrieval (IR), such as a user's search performance, interaction with the system, time spent completing tasks, and subjective impression of their workload. With greater knowledge of PS, systems could be designed that accommodate users with low PS to improve their overall search experience. However, in this perspectives paper, we analyse how PS tests have been used in IR, and identify multiple uncertainties regarding PS content, administration, analysis, and reporting of findings. Consequently, we aim to stir discussion between IR researchers by drawing awareness to these issues. As a result, we further discuss challenges involved in advancing how future PS tests are used in IR. Finally, we propose recommendations that have the potential for enhancing the reliability and validity of current PS tests.

CCS CONCEPTS

• **Information systems → Users and interactive retrieval**; • **Human-centered computing → HCI design and evaluation methods**.

KEYWORDS

perceptual speed; cognitive ability; individual differences; information retrieval

ACM Reference Format:
Olivia Foulds, Leif Azzopardi, and Martin Halvey. 2020. Reflecting upon Perceptual Speed Tests in Information Retrieval: Limitations, Challenges, and Recommendations. In *2020 Conference on Human Information Interaction and Retrieval (CHIIR '20), March 14–18, 2020, Vancouver, BC, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343413.3377982

CHIIR '20, March 14–18, 2020, Vancouver, BC, Canada

ACM ISBN 978-1-4503-6892-6/20/03…$15.00
https://doi.org/10.1145/3343413.3377982

1 INTRODUCTION

In the field of Information Retrieval (IR), it has long been advised that researchers must look beyond the system and consider how a user's individual capabilities and expertise impact their search behaviour, performance and experience [2, 42]. Out of the many different factors that influence people's success when undertaking information seeking tasks is their cognitive abilities – where it has been observed that people with higher cognitive abilities tend to perform searches better than those with lower levels [8]. Specifically, *Perceptual Speed* has been regarded as one of the most important cognitive abilities that affects information-seeking [2, 3, 29].

Perceptual Speed (PS) is defined by an individual's accuracy and speed to view, scan, and compare information during visual search tasks [4]. With PS being a type of cognitive ability, it's underlying neural mechanisms are thought to be automatic and fairly stable throughout an individual's life [29]. As PS varies between individuals, multiple PS tests have been developed that attempt to detect this cognitive ability such as: the Minnesota Clerical Test (1965) [18], Ekstrom's (1976) Kit of Factor-Referenced Cognitive Tests[15]; Wechsler's (1981) Digit Symbol Substitution Test (cited in [33]); and Salthouse & Coon's (1994) Letter Comparison Test [16, 32]. Irrespective of the exact test used, they all follow a similar format that involves scanning a list of stimuli and identifying certain targets against a set time period. People who are most accurate at identifying targets in the fastest amount of time are said to have 'high' PS, while people who make more mistakes and take longer are considered to have 'lower' PS levels [14].

While PS tests have been used for over 50 years across a variety of domains, less than forty IR studies have been conducted using such tests as part of their experimental process. Yet in most of these studies, researchers have found that there have been significant differences in terms of behaviour between participants with low PS and high PS. For example, in a recent experiment on the influence of working memory and perceptual speed when interacting with aggegrated search result pages [6], they identified that individuals with lower PS found it more difficult to identify relevant results.

Considering that PS has been shown to significantly influence people's search performance and how well they complete information seeking tasks, it appears to be a very useful indicator and instrument to use in future studies. This is because if there is a strong link between PS and information seeking performance, then designers and developers can focus attention towards developing interfaces and interventions that aid and support people's varying cognitive abilities [10]. For instance, if perceptual speed involves scanning and finding some kind of target in a visual display, then

people who struggle with this and are therefore said to have low PS may benefit from alternative interface layouts that contain additional tools such as highlighting capabilities or hierarchical headings. These adaptations could theoretically allow the user to better navigate, and keep track of, what's in front of them so that less visual scanning is required [3, 8, 20, 39]

With the promising opportunities that PS measurement could provide for future system development, and with a growing number of works using PS tests, then it is timely to examine how such works have employed PS testing and to consider whether the instruments are valid and reliable in this IR context. Thus the aim of this paper is to examine how PS tests have been used within IR studies, what issues are associated with using such tests, and provide a guide for those wishing to use them. To this end, we review the currently available IR literature to inform our discussion and recommendations.

Therefore in this perspectives paper, three main contributions to the IR community are explored:

- Firstly, following the advice of [27], we aim to facilitate dialogue amongst IR researchers by quantifying and making others aware of the methodological, reliability, and validity issues associated with PS testing administration, analysis, and reporting.
- Secondly, having considered the limitations of PS testing, we discuss the current challenges that the IR community needs to address regarding PS testing.
- Finally, we provide a series of recommendations for enhancing the quality of PS testing in IR.

The rest of this paper is organised as follows: Section 2 describes one of the most commonly used PS tests and further elaborates upon the significant results that studies have found in the areas of Information Retrieval and Information Visualisation. Section 3 details the approach taken to find and analyse PS papers relevant to IR. Section 4 presents the main problematic themes that occur throughout the PS literature. Finally, the paper concludes in Sections 5 and 6 by acknowledging the challenges and making recommendations for improving future PS tests.

2 PERCEPTUAL SPEED TESTING IN IR

A variety of studies concerning IR have identified various parts of the search process that are significantly affected by PS. These studies have predominately used tests drawn from or based on *Ekstrom's Kit of Factor Referenced Cognitive Tests* [14]. The kit comprises of three different PS tests that researchers may choose to use. However, Ekstrom suggest that in order to fully deduce a cognitive factor, at least two tests should be administered [14]. The three Ekstrom PS tests to choose from involve numbers, words, or symbols and are shown in Figure 1 and described below:

- *Finding A's*: Participants must effectively scan columns of words and select any that contain a letter "a".
- *Number Comparison*: Participants are given pairs of numbers, and are required to indicate whether the numbers are the same or different by placing a cross on non-identical pairs.
- *Identical Pictures*: Participants are given a symbol and must select the identical image against a choice of five.

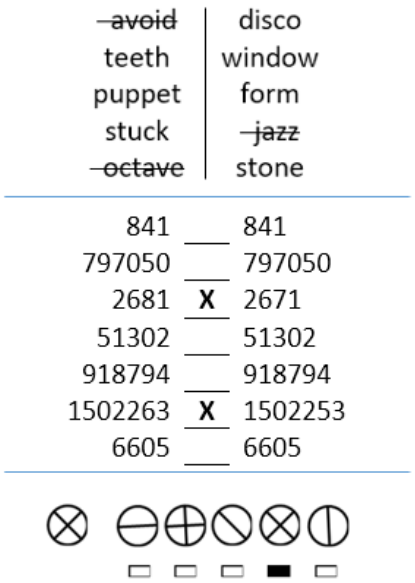

Figure 1: Sample PS Tests based on Ekstrom's Kit [14]. Top: *Finding A's* Test. Middle: *Number Comparison* Test. Bottom: *Identical Pictures* Test.

As previously mentioned, in IR, various parts of the search process are significantly affected by PS. People with higher levels of PS have: achieved better search performance and learned more vocabulary whilst being overall faster [4]; engaged in more search activity such as issuing longer queries, viewed more URLS, and clicked on more search engine result page (SERP) links, all while experiencing less self-reported workload [8]; and spent significantly less time finding relevant documents during TREC search tasks [2]. In comparison, those with low PS tend to spend more time examining SERPs [8]; have reported blended interfaces (when the display contains verticals such as videos/images) as less usable; and were less satisfied with their search performance [40]. In the study presented in [40], the participants with lower levels of PS who significantly rated the blended interface as less usable attributed factors such as distraction and confusing as the reason for worse user engagement. Other research has indicated why lower PS users may find interfaces more confusing: individuals with high PS have a higher eye fixation rate, and are thus able to scan what's in front of them more quickly compared to people with low PS [36, 39]. Therefore, greater understanding of PS could hypothetically enable search systems to be developed that can adapt and help users with lower PS levels retrieve relevant information to the same extent as high PS users.

Additionally, PS has been found to not only impact information retrieval, but also related tasks in the field of information visualisation. For example, PS identifies what kind of visualisation is most effective for a given user such that individuals with high PS appear to do better at tasks when observing data via coloured boxes, while low PS individuals excel when data is presented on a radar graph [13]. These results concerning how PS affects complex interactive

visualisations extend older work that only focused on the impact of PS on static visualisations: PS impacts an individual's suitability for a particular visualisation [35, 36, 41, 43]. Furthermore, similar to findings from information retrieval studies, experiments have also identified that PS interferes with search performance when different visualisations are manipulated such that a users performance, as measured by time on task, negatively correlates to particular visualisations depending on PS levels [11, 17, 26]. Consequently, PS potentially has the ability to inform interfaces that can adapt to the most ideal presentation for the individual user's needs.

Overall, it has been demonstrated that PS is an important factor to consider with the potential to lead to useful insights for future research and real-life application. However, although one review exists that analysed over 2100 articles on IR, and "perceptual speed" was used as one of the search terms, there was no explicit discussion or results of PS tests [26]. Instead, an amalgamation of cognitive abilities were merged together to conclude that as a result of issues around measurement and generalisability, it was unknown how these individual differences truly affected search outcomes [26]. This appears surprising when Ekstrom's PS tests have been widely used since their development in 1976. With such a long time period of use, the reliability of these tests would be thought to be high. However, as [27] pointed out, items in many studies lack validity and reliability evaluation. Thus, the present paper aims to evaluate the literature concerning PS and IR, in order to make researchers aware of any current limitations, and suggest future recommendations for improving PS usage.

3 REVIEW PROCESS

To provide the basis for our analysis and discussion, we performed library searches to identify studies which had used PS tests in the context of IR. Thus, we defined our search criteria as follows.

Firstly, we used the Association for Computing Machinery (ACM) Digital Library (DL) – which contains references to core IR resources, conferences and journals. Our initial search for *perceptual speed* returned 19,451 results. Subsequently, we added inverted commas to our query to ensure papers were returned that were not dealing with *perceptual* and *speed* as separate entities. This returned only 12 results, all of which have been used for analysis in this paper.

We further repeated this process and performed searches using the same query of *"perceptual speed"* in our university library, which encompasses a huge selection of many databases and returned a much larger result of 6,064 entries. Brief manual scanning revealed that many of these results were predominantly coming from the medical industry. Therefore, to maintain our focus on IR, we changed our search query to *"perceptual speed" AND "information retrieval"*, with the filter of peer-reviewed items only, which brought back a more manageable 69 results. To ensure that PS was one of the main focuses of the paper, our inclusion criteria involved manually reviewing each of the 69 abstracts to eliminate any that did not indicate the use of PS tests in the context of an IR study. This left 11 papers.

Finally, 16 more papers were discovered through reference crawling of the 23 already found papers. Although seven of these papers were not directly IR, but rather originated from a psychological background, they were still included to explain the psychological principles behind the fundamental PS tests.

With an overall corpus comprising of 39 papers published between 1965 and 2019, we began reviewing these papers in search of main themes. In this approach, data analysis is not conducted with pre-specified questions that need answering, but rather themes emerge from the data itself [31]. Consequently, through a reiterative process of paper reading, themes began to emerge regarding PS test content, administration, analysis, and how results were reported. Rather than quantitatively coding all possible themes, instead, we followed a more qualitative approach to accompany this perspectives paper. This involved reporting the main themes that with others awareness, we believe would help improve PS testing for future studies.

Of the 32 papers that used PS tests in IR studies, 30 used one of Ekstrom's test, while the other two used the Minnesota Clerical test. For the purposes of discussion we will focus on PS tests in light of Ekstrom's tests.

4 MAIN THEMES OF PS TESTS

As a result of letting themes emerge from the literature on PS in Information Retrieval, many uncertainties regarding PS test content, administration, analysis, and how results were reported have been identified and split into six main themes below.

4.1 No Standardised Thresholds

One of the most notable uncertainties with PS tests is that despite being over 40 years old, and many papers have used them and referred to "high" and "low" PS levels, there are no standardised thresholds for what defines a high/low PS. Rather, only a few papers have even explained how they categorised high/low PS: participants were assigned to a low or high group, based on a median split of perceptual speed scores[2, 35, 36, 40]. The problem with reporting high/low based on a median split without providing the scores is that is it not possible to compare across studies, nor can one know what is high or low, or whether there is any statistical difference between the groups.

Table 1 presents a summary of the IR papers that report the median score from the PS test used. With further examination, a huge discrepancy in results can be noted. In [40], participants were classified as having low PS if they scored between 34 and 51, and a high PS if they ranged between 51 and 74. On the contrary, [6] filtered low PS individuals as those scoring between 44 and 63. Therefore, despite the same identical tests being administered, if a participant scored within the range 51-63, one study would classify the participant as having high PS, whereas the other study would categorise the participant as having a low PS. With such discrepancy in analysis depending on the individual sample of participants tested, this greatly reduces the comparability of results across studies.

4.2 Inconsistent Reporting of Results

Out of the papers reviewed, only six, or 15.4%, reported exact figures for their PS test results (See Table 1). Instead, the majority of existing literature concerning PS tends to only report explicit figures that refer to the significant effects PS have had on another part of an

Table 1: Perceptual Speed Results Reported in the selected studies.

Study	PS Test	Possible Range	Mean (SD)	Median	Min, Max
EKM, cited in Turpin et al. [40]	Finding A's	-	47 (14.9) = Males, 54 (14.9) = Females	-	-
Turpin et al. [40]	Finding A's	0-200	51.94 (10.41)	51	34, 74
Arguello and Choi [6]	Finding A's	0-200	64.16 (12.00)	63	44, 90
USAF, cited in Brennan et al. [8]	Number Comparison	-	47.94 (12.32)	-	-
Brennan et al. [8]	Number Comparison	0-96	44.38 (10.58)	44	25, 73
Crabb and Hanson [11]	Number Comparison	-	46.63 (6.04) = Young 45.08 (6.94) = Old	-	-
Allen [3]	Number Comparison	37	30.1 (8.8)	-	-
Toker et al. [39]	Identical Pictures	-	85.70 (11.64)	-	54, 96
Allen [3]	Identical Pictures	42.5	80.9 (11.4)	-	-

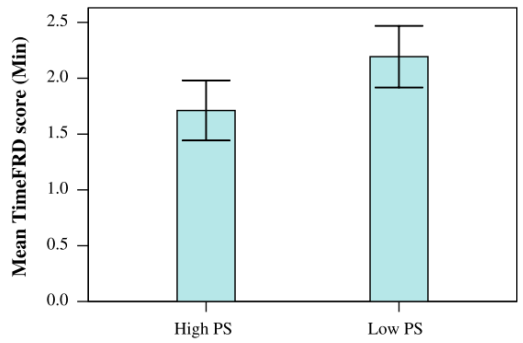

Figure 2: Example of how previous studies only report PS effects, and don't define what is high and low. Source: [2]

experiment. For example, in a study [2] that examined whether PS affected how long it took for a user to retrieve a relevant document, the only PS figures reported were that users had been grouped into high and low based on an unknown median split, and graphs that detailed how these categories impacted a users time on task were illustrated (See Figure 2). Therefore, apart from the six studies mentioned in Table 1, in the remaining 85.6% of papers examined, it is not possible to know the PS scores. Consequently, this lack of reporting figures makes it difficult for other researchers to compare and assess the reliability of any results found, which ultimately reduces the academic rigour of many PS studies.

Additionally, even in the studies that have reported PS figures, it is questionable whether the PS scores are truly valid. For instance, Ekstrom themselves originally stated that: "*It is strongly recommended that researchers use more than one of these tests in any exploratory endeavour that aims at identifying a factor*" [14]. Yet, many PS studies that have been discussed and claimed to find significant results only administered one of Ekstrom's tests [6, 8, 39–41].

Furthermore, even in the papers that did use more than one PS test, an explanation for how to merge scores from multiple tests is lacking. It is therefore unknown if test scores were weighted equally with an average of the two taken, or if precedence was given to one test over another and if so, which one? For example, although authors stated that two of Ekstrom's PS tests achieved a moderate Cronbach reliability rating, no explanation for how this was deduced was given [4]. Instead they claimed that the two PS tests were assessing different aspects of PS and thus analysed them as separate entities [4]. Similarly, another study claimed that one of Ekstrom's PS tests was too similar to a different cognitive ability test, and therefore they excluded these from their analyses [5]. With so many unknowns with calculating an overall PS score, this reduces the consistency with which PS tests can be analysed throughout the literature.

Lastly, from all of the papers reviewed, only one mentioned that in order to enhance the reliability of their results, participants repeated the PS test approximately 5 days later after their initial test [16]. However, in the work presented in [16], no explanation was provided for how they then calculated the overall PS score. For example, did they take an average between the two separate sessions, or just randomly decide to report only the results from one? Regardless, it is surprising that more PS tests are not repeated across multiple sessions considering PS is thought to be relatively stable in individuals [5]. Because of the stable nature of PS, if a participant was not gaining a similar PS score on both sessions, then this would imply that the PS test was not truly measuring PS [12].

4.3 Unclear Marking Instructions

If a researcher wishes to administer a PS test in their study, then they must subsequently be able to analyse the test results correctly to compute a PS score. However, the original marking advice for each individual test lacks clarity and may lead to some confusions by participants completing the tests, but also researchers scoring the tests. Unfortunately, as Ekstrom's PS tests are over 40 years old, the original references that are discussed by Ekstrom are very old and inaccessible, potentially because they have not been digitised. It is therefore unknown how the tests were exactly developed, and which points in the test are the most important factors that need to be considered to deduce the overall PS score. For example, the *Number Comparison* test instructs participants to cross any pair of numbers that are not identical. Results are then calculated by the 'number marked correctly minus number marked incorrectly' in a given time period [14]. As this test is meant to monitor how many pairs of numbers a participant can scan through in a set time period, yet participants are only indicating the numbers which

are non-identical, it is unknown how many pairs of numbers they have successfully acknowledged as identical. Thus, the current advice for scoring this test does not fully correspond to the original instructions given.

Furthermore, the *Finding A's* test also encompasses issues. To reiterate, this test instructs participants to score any words that contain the letter 'a' in them and emphasises that each column has 5 words containing the letter 'a'. Participants are also told: "Your score on this test will be the number of words marked correctly. Work as quickly as you can without sacrificing accuracy" [15]. However, there is no explanation given as to how to score a participant's answers if each column is not completed. For example, if a participant was aware that they hadn't identified 5 words containing 'a' in a column, should they delay their time by continuing to repeat a visual scan of the same column or skip to the next column? This lack of understanding in instruction creates a huge gap in deducing overall PS. Participants are different, and their scanning abilities will undoubtedly vary. Thus, one person may get an accuracy score of say 15, but they only completed the first 3 columns thoroughly with no mistakes on page 1. Whereas another person may get the same accuracy score of 15, but rather than finding all correct answers that were immediately in front of them, they got this score from briefly scanning rows across 9 columns in 2 pages. With such opposing possibilities of results, it seems unusual that there is no explanation for how to score these differences, and what these results may mean for a person's true PS level. If PS involves accuracy and scanning of what's visible [4], then surely there's a difference in PS levels depending on whether a participant can efficiently identify everything that's visible without making any mistakes, compared to finding some correct answers over multiple pages whilst simultaneously missing many others.

4.4 Different Formats

Similar to how there are unknowns in whether there is a difference in PS depending on whether systematic or random scattering of answers is employed, it is also unknown how PS tests were exactly formulated, again presumably due to the lack of accessible old references available. For example, Ekstrom's *Number Comparison* presents 24 rows of numbers in 2 columns [14]. Alternatively, another kind of PS test following the same principles, the 'Minnesota Clerical test', incorporates a number comparison test of 4 columns, each with 50 rows [18]. Yet, neither tests explain why these exact numbers were chosen. Thus, would a test for PS equally measure PS if there were 6 columns, as opposed to 3, visible at any one time?

Additionally, the same study [18] showed that some number comparison tests possibly contain confounding elements because the index of number change was never equally distributed. Likewise, we personally calculated the exact indexes for change in Ekstrom's tests, and found a couple of number pairs in the *Number Comparison* that had more than one difference in them. With no formal explanation as to how these numbers, indexes for change, and columns were formulated, it is unknown whether these are fundamental mistakes in the original design or whether or not these different variations matter for the validity of PS. However, as other research exists that demonstrates how visual perception changes depending on layout, it would make sense that the format of the PS test is important. For example, one eye-fixation can process 24 letters in a vertical position, compared to 12 letters in a horizontal position [22, 28]. Yet, Michalski and Grobelny (2015, cited in [19]) found that individuals better perceive horizontal layouts more than vertical ones.

Furthermore, although many studies stated that they used Ekstrom's PS tests, officially these tests require a licence to use [15], and yet none of the papers reviewed mentioned how, or even if, they obtained licensing. Therefore, this may suggest that researchers have instead used Ekstrom's PS tests as a guide to make their own test. Although this point is just speculation, if it is the case, then the exact format of how the PS test was visibly administered in many tests is unknown. This again makes the comparability of PS studies challenging.

4.5 Limited Linguistic Reasoning

In the *Finding A's* PS test, each page contains 5 columns, with 41 words per column, and thus 205 words per page [14]. With 4 pages per part, that totals 820 words. As there are 2 equivalent parts, this means there were 1640 words overall, out of which 200 (or 12.195%) contained the letter 'a'. After an analysis of the words used, we observed that the number of letters in all the words appeared to be quite equal, ranging from equivalent words that have 4,5,6,7, and 8 letters in length, and that there was a fairly equal balance between 1 or 2 syllables used. It was also noted that some of the words were repeated in Ekstrom's *Finding A's*.

Yet despite such a large set of word stimuli, there is no explanation for how the words were selected for the test. Additionally, it is not documented whether the words: contain the same frequency in the English language; elicit similar sentiment; were positioned in any particular order; are processed differently by a Native English speaker; and other factors that are important in linguistics such as the distribution of nouns and verbs [21, 25, 30]. These points do not necessarily reduce the reliability and validity of the PS tests to date, as they have been successfully used over many years to elicit significant results. However, it is worth being aware of these factors for any future researcher that may wish to expand upon and develop their own PS test to ensure that the stimuli chosen contain the same components that produce valid results for PS. For example, studies in neuropsychology have long recognised that emotional words are perceived stronger than non-emotional words [37]. Thus, if all the words that contained the letter 'a' were more common or emotionally sentimental in the English language compared to the words that didn't contain a letter 'a', then perhaps individuals would automatically identify them, regardless of their PS levels.

4.6 Outdated Administration and Content

If a researcher wanted to administer Ekstrom's PS tests, then a licence must first be sought from ETS Research [15], who then distribute PDF copies of the specific tests requested. Yet, the administration of the tests remains the same as 40 years ago when they were first devised: a paper-pen version, which ultimately requires manual scoring. In fact, psychological researchers have described how scoring PS tests takes longer than the participant completing the actual test, described in [1] as: "*the scoring process turns into somewhat of a PS test for the individual scorer, as he or she attempts*

to count correct, missed, and incorrect responses using a template to match to the examinee's responses". However, with many experiments now run online, it makes it impractical to use paper based surveys. This motivates the question, how do we computerise the PS tests such that they are reliable and valid instruments?

Furthermore, caution must be taken when using a cognitive test that dates back so many decades because over time, attention evolves. For example individuals now "*have to fight to stay focused on long pieces of writing*" as a result of information technology [9]. Likewise, a recent study by [23] reaffirm that individuals currently have a limited capacity for attentional resources, and that this is not helped by current information workers experiencing increasing levels of distractions. In relation to PS, attention is fundamental to cognition, and PS is a type of cognitive ability. Keeping this in mind, the authors of this paper conducted a pilot study of the *Finding A's* test, and discovered that it took over 10 minutes for some participants to complete the test. Yet, the original *Finding A's* was meant to only take 2 minutes to complete 4 pages with 820 words. Therefore, it is worthwhile making new researchers aware of these differences, to ensure that the current PS tests contain the right amount of stimuli and time necessary for current states of individual attention.

5 DISCUSSION

Although PS testing has provided promising results in many IR studies, as a result of thoroughly analysing these studies, many uncertainties have been described that provide interesting debate for current researchers to consider. Consequently, the above themes identified have provoked challenges and recommendations for future administration and analysis of PS tests.

5.1 Challenges

The key challenges researchers face with PS testing appear to concern the content and administration.

Regarding administration mainly from the above theme of *Outdated Administration and Content*, an obvious next step for furthering PS testing may seem to be converting the old paper-pen format into a modern, computerised test. This would then theoretically resolve the problems identified of being old-fashioned and difficult for researchers to score and analyse, which may have even put some researchers off from considering using PS tests. Consequently, if the PS test was administered online, then it might be easier for researchers to integrate into their studies where the main part is already administered online, and thus the hassle of switching between paper and computer would be eliminated. With a more effortless form of administration and automatic scoring from a computer, more researchers might be encouraged to involve PS testing into their research, which would in turn increase the reliability of results if more studies were able to be compared. Although these points are just hypothetical, other researchers have stipulated the benefits of computerising PS tests with the main reason being that software could be dynamically used to adapt the screen to counter the negative effects for low PS users [13].

However, it is not as simple as taking the same paper PS tests and converting them to an online format for many reasons. Firstly, there's a difference between how stimuli are perceived depending on whether they are viewed on paper or a computer. For example, completing 41 words on a column on A4 paper may differ to how many words you can physically see at once in a column on a different sized computer screen. This difference was reaffirmed by [34] who compared participant's responses to a PS test conducted on paper, a video display terminal (VDT), and a combination of switching between both. The exact PS test used was not one of Ekstrom's, but similarly involved 200 number comparisons taken from the Minnesota Clerical Test. As can be seen in Table 2, people score a lot less when conducting the tests online compared to paper. Therefore, a lot more further research is needed that explains these differences, in order to develop an online PS test that is truly measuring PS.

Table 2: PS differences depending on administration type in [34]

	Mean	Standard Deviation
Paper	119.29	32.42
Online	85.24	21.15
Both	67.32	15.98

Secondly, a few studies have attempted computerised PS tests, but with no explanation as to what measures were taken to account for the above problems surrounding converting PS tests from paper to online. For example, [44] took 60 numbers from Ekstrom's *Number Comparison* and administered it online within a 90 second time period. This is in comparison to Ekstrom's original 96 number comparisons over three minutes [14]. Yet, [44] provided no explanation for: why only 60/90 items were taken; how they chose those particular 60 items over the remaining 30 that were not picked; why the time limit was halved; or how the content was visually divided and presented on a screen in columns or rows. Similarly, [34] and [16] attempted computerised PS tests, but again, no justification for their content or explanation for how they were presented was given.

Additionally, from the literature reviewed on computerised PS tests, many other factors were also not discussed that may influence the validity of PS tests. These include: how participants physically select the answers on a screen such as whether selected items are scored out or change colour; whether all stimuli are presented in individual boxes, grid-lines, or blank backgrounds; if words/numbers are aligned to the left, middle, or right of the screen; what font is used; and what is the inter-letter spacing or spacing between items. This list is not exhaustive, and of course it may be that these factors are incidental in affecting a PS score. However, although not specifically examining PS, other psychological research has identified that inter-letter spacing is a perceptual factor that modulates visual word recognition performance: decreased spacing resulted in slower identification thereby confirming the interference between close proximity of stimuli and visual perception [24]. Thus, if interletting spacing affects perception in reading, it may also affect how PS tests are designed. Consequently, the above list of factors described may affect PS online test validity. Yet with so many variables apparent, much more research is clearly needed that investigates and accounts for these components before a precise and valid PS test can be assured.

If time was invested into developing a new computerised PS test, then it would appear worthwhile for researchers to consider, and account for, some of the other themes that this paper identified regarding the content of current PS tests. Namely, the different formats, linguistic reasoning, and attentional structure all ignite discussion for researchers to consider.

As one of the themes in this present paper identified that there is variation between different PS tests concerning the format of stimuli, such that Ekstrom's *Number Comparison* presented stimuli in 2 columns of 24 rows [14] while the Minnesota Number Comparison presented 4 columns of 50 rows [18], it is unknown what the optimal layout for PS tests should be. Moreover, details about how the original PS tests were developed have never been specified, causing unresolved questions as to whether different formats of visual presentation were even tested on people to gauge any possible differences in PS response. With other research having identified that visual perception is influenced by horizontal/vertical layouts [19, 22, 28], and calls for computerising PS tests have determined the need for reconsideration of PS test displays [1, 34], further PS development is needed. Experiments should manipulate multiple different ways at physically viewing the PS stimuli such as different variations of columns and rows. Although time consuming to design and test, these manipulations are necessary to ensure that any new computerised PS tests are still valid and effectively measuring PS.

Before research can consider the layout of stimuli, the correct kind of stimuli that will equally elicit valid PS results must first be deduced. Our theme of *limited linguistic reasoning* discussed how the meaning and structure behind the stimuli chosen for the PS test that contained words was unknown. Therefore, further research is required to make sure there are no confounding variables, such as certain words containing too highly emotional meanings and thus making perception easier [21, 25, 30], negatively influencing PS results. Thus, when selecting word stimuli, new researchers may wish to make use of databases such as The English Lexicon Project [7], where words can be chosen, filtered and equalled for specific lexical characteristics.

Furthermore, as PS tests are effectively measuring how accurate and fast an individual is at identifying some kind of perceptual change [14], such as a word that contains an 'a' or a number that doesn't equate with it's pair, more investigation is required as to where the index of change is positioned, and how many changes there are. For example, in the *Finding A's* PS test, there are 41 rows of words where 5 contain a letter 'a'. Firstly, questions to consider include whether it is necessary that there are always exactly 5 changes to be identified, as opposed to another specified or random number. Secondly, the spacing between target answers requires deliberation. For instance, does it matter how close together the words containing 'a's are? Are they all clustered together in the centre of the column, equally distributed throughout, or randomly dispersed such that some end up close together while other columns are sparse? Again, these questions aim to stir discussion with researchers who wish to develop new PS tests to ensure the structure is still reliably measuring PS.

Finally, this current paper identified a main theme which involved the current PS tests being outdated. Beyond the outdated administration of paper/pen formatting, the notion of human attention changing over 50 years was discussed. As PS is a type of cognitive ability, and attention is a key component of cognition, it is crucial that future PS tests do not overload people's limited attentional capacities. Thus, perhaps new PS tests may need to be shorter, contain fewer overall stimuli, or the length of time to complete the test should be extended. Before these revisions can be achieved, all components of the PS test that may affect attention need re-examination. This is essential to ensure future PS tests are still validly measuring PS, whilst simultaneously accounting for the fact that attention may have evolved over time. Lastly, reconsidering attention limits of participants is necessary to ensure that they are not overtired as a result of PS testing, as this may adversely confound any results found in subsequent tests they complete in the main studies.

5.2 Recommendations

In the *challenges* section of this review, many areas requiring a lot of further research have been discussed. Yet practically speaking, it would take a considerable amount of time before any of the results from this research could be implemented in future PS tests, where reliability and validity of PS is still guaranteed. Nonetheless, although we have explored the need for PS tests to be revised and computerised, the original paper/pen format has still proven to be quite useful, with many studies finding significant results. However, the themes identified in this current paper have ignited some recommendations for currently available PS tests that researchers might wish to follow in order to improve their administration and overall reliability and validity of results found.

Firstly, the theme of *unclear marking instructions* identified that the *Number Comparison* PS test score is calculated as a result of the items participants marked correctly *and* incorrectly. Yet, participants are only instructed to cross out non-identical number pairs, which leaves it unknown how many identical pairs they have correctly scanned through. Thus, a perhaps better way of administering this PS test would be for participants to 'tick' for same, and 'cross' for different pairs of numbers. That way, the researcher would be able to exactly quantify how many pairs a participant is efficiently scanning through. As PS concerns an individual's accuracy and speed to view, scan, and compare information during visual search tasks [4], we would hypothesise that having a more robust way of quantifying how many items a participant is processing would return a more valid measure of PS.

Another recommendation that we propose would increase the PS test validity regards how many times the PS test is administered on the same participants. Realistically, it may be difficult to recruit participants on multiple occasions. However, as PS is meant to be a stable cognitive ability [29], if a participant wasn't getting a similar score on the same test at different times, this would reduce the validity of results [12]. Thus, if a researcher wanted to reaffirm that the PS test they were administering was truly measuring PS, we would advise taking a small sample of participants and administering the PS on two separate occasions to ensure similar results were being obtained.

Additionally, another theme established a breach in PS validity as many studies claiming to have assessed PS only used one PS

test, which contradicts original guidelines that stated more than one PS test was required to fully identify a cognitive ability [14]. Accordingly, we encourage future researchers to avoid this problem by always administering at least two PS tests. Unfortunately, there are no explicit guidelines on how to merge multiple test results together. However, [3] utilised Cronbach reliability testing between 2 of Ekstrom's tests. This measures the internal consistency, otherwise known as how closely related a set of items are as a group [38]. Thus, we also encourage researchers who use more than one PS test to run reliability analyses between their PS tests to increase the reliability of their overall PS measure.

The theme of *inconsistent reporting of results* identified a consistent trend which involved how many previous IR studies involving PS failed to report the exact results or distribution of the PS tests used. Therefore, we strongly recommend that all researchers should avoid this unknown and instead report as many exact results as possible such as: the median; mean; standard deviation; and a graph that contained all possible PS scores with how many participants achieved each score. If unknown figures were made to be known and explicit in future PS tests, then we would predict that this would improve the reliability of results obtained and make it easier for other researchers to compare their studies to. Additionally, if there was then a large sample of multiple studies who had used and reported their PS scores, an analysis would be possible that could compute average standardised thresholds. Having an exact threshold for what was considered 'high' and 'low' would then benefit future studies to ensure a consistency in results, regardless of the sample of participants used.

Finally, beyond unknown exact figures, there are other factors that many studies failed to report, which if they had, would have increased the robustness of results obtained. For instance, in [6], they state that the *Finding A's* PS test incorporated a possible range of 0-200. Yet, no units were given for these figures or explanation for what those figures exactly meant. It is therefore unknown where these numbers came from which leads to new researchers being left unable to compare these figures into their own work. Furthermore, many studies never expressed the format of the PS test used: although they quoted that the PS test originated from Ekstrom, which as we know is paper-based, only few studies explicitly state whether the administration of the test was done on paper. Hence, there is no guarantee that other studies have all administered their PS test in paper/pen format, and perhaps instead taken the Ekstrom stimuli as a guide and computerised it. If this was the case, then this would interfere with the comparability of PS testing between studies. Consequently, researchers of future PS tests should ensure that all details and aspects of their PS test are always reported, to allow for easier reviewing and reliability assessments of the overall test usage to be made by others.

6 CONCLUSIONS

Overall, although PS has shown promising and significant results across the literature in IR, the current paper has identified many areas that could be improved upon to make the tests even more reliable and valid. Regarding the content of the tests, more understanding is needed for: the linguistic structure of words used as stimuli; where changes are positioned, and how many there are; a reconsideration of current human attention and how this may affect how many stimuli are visible; and a further exploration for how the format of stimuli should be visually presented. Concerning analysis of the tests, further research is needed for clearer marking instructions and setting standardised thresholds. Additionally, the administration of PS tests needs some refinement such that they: should be computerised; more than one PS test should be administered; and the same tests should be completed by the same participant on two separate occasions. However, challenges were noted that explained the difficulty of converting a paper test into an online format with appropriate stimuli. Finally, to increase the comparability of PS studies, researchers should report actual figures and specific details about their test so that transparency is increased and comparisons between different research samples is enabled.

All of these recommendations and challenges summed together provide many avenues and questions for exciting future research that could have real-life application. It has already been established that PS does affect IR, such that users with lower PS are: engaged in less search activity; report more workload; subjectively perceive interfaces as less usable; and spend significantly longer finding relevant documents [2, 8, 40]. Thus, if PS could be more easily tested, such as through a short computerised test, then this could drive adaptive systems to be developed. This would be advantageous for people identified as having lower levels of PS who as it stands, struggle with visually processing information that is presented in a cluttered manner on screens [40].

Additionally, if measures of PS were more reliable and valid, and further studies reinforce that low PS participants really do struggle with certain IR components, PS tests could potentially be used as a screening test for certain jobs. For example, in industrial applications such as selecting sonar operators that must visually scan and retrieve a huge amount of information, an ideal job candidate would be someone who is the best, in terms of both speed and accuracy, at processing visual stimuli in these high-pressured information environments. However, before these applications could be developed, it is clear that firstly, PS tests have a lot of refinement and further research needed. Consequently, it is hoped that this perspectives paper will make IR researchers aware of the limitations in order to stir discussion and ignite debate to advance future PS test usage.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their helpful comments. This work was part funded by BAE Systems Maritime and EPSRC as part of an Industrial Cooperative Award in Science & Technology (CASE) Studentship (EP/S513908/1).

REFERENCES

[1] Phillip L. Ackerman and Margaret E. Beier. 2007. Further Explorations of Perceptual Speed Abilities in the Context of Assessment Methods, Cognitive Abilities, and Individual Differences During Skill Acquisition. *Journal of Experimental Psychology: Applied* 13, 4 (2007), 249–272. https://doi.org/10.1037/1076-898X.13.4.249

[2] Azzah Al-Maskari and Mark Sanderson. 2011. The effect of user characteristics on search effectiveness in information retrieval. *Information Processing and Management* 47, 5 (2011), 719–729. https://doi.org/10.1016/j.ipm.2011.03.002

[3] Bryce Allen. 1992. Cognitive differences in end user searching of a CD-ROM index. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 298–309.

[4] Bryce Allen. 1994. Cognitive abilities and information system usability. *Information Processing and Management* 30, 2 (1994), 177–191. https://doi.org/10.1016/0306-4573(94)90063-9

[5] Bryce Allen. 2000. Individual differences and the conundrums of user-centered design: Two experiments. *Journal of the American Society for Information Science and Technology* 51, 6 (2000), 508–520. https://doi.org/10.1002/(SICI)1097-4571(2000)51:6<508::AID-ASI3>3.0.CO;2-Q

[6] Jaime Arguello and Bogeum Choi. 2019. The effects of working memory, perceptual speed, and inhibition in aggregated search. *ACM Transactions on Information Systems* 37, 3 (2019). https://doi.org/10.1145/3322128

[7] David A. Balota, Melvtn J. Yap, Michael J. Cortese, Keith A. Hutchison, Brett Kessler, Bjorn Loftis, James H. Neely, Douglas L. Nelson, Greg B. Simpson, and Rebecca Treiman. 2007. The english lexicon project. , 445–459 pages. https://doi.org/10.3758/BF03193014

[8] Kathy Brennan, Diane Kelly, and Jaime Arguello. 2014. The effect of cognitive abilities on information search for tasks of varying levels of complexity. In *Proceedings of the 5th Information Interaction in Context Symposium on - IIiX '14*. ACM Press, New York, New York, USA, 165–174. https://doi.org/10.1145/2637002.2637022

[9] Nicholas Carr. 2008. Is Google Making Us Stupid? *Yearbook of the National Society for the Study of Education* 107, 2 (oct 2008), 89–94. https://doi.org/10.1111/j.1744-7984.2008.00172.x

[10] Xiuli Chen, Gilles Bailly, Duncan P. Brumby, Antti Oulasvirta, and Andrew Howes. 2015. The Emergence of Interactive Behavior. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*. ACM Press, New York, New York, USA, 4217–4226. https://doi.org/10.1145/2702123.2702483

[11] Michael Crabb and Vicki L Hanson. 2016. An analysis of age, technology usage, and cognitive characteristics within information retrieval tasks. *ACM Transactions on Accessible Computing* 8, 3 (2016). https://doi.org/10.1145/2856046

[12] Victoria Crisp. 2017. Exploring the relationship between validity and comparability in assessment. *London Review of Education* 15, 3 (2017), 523–535. https://doi.org/10.18546/LRE.15.3.13

[13] Scott C. Douglas and Mark J. Martinko. 2001. Exploring the role of individual differences in the prediction of workplace aggression. *Journal of Applied Psychology* 86, 4 (2001), 547–559. https://doi.org/10.1037/0021-9010.86.4.547

[14] Ruth B Ekstrom, Diran Dermen, and Harry Horace Harman. 1976. *Manual for kit of factor-referenced cognitive tests*. Vol. 102. Educational testing service Princeton, NJ.

[15] Ruth B Ekstrom, John W French, and Harry H Harman. 1976. *Kit of Factor-Referenced Cognitive Tests*. Technical Report. https://www.ets.org/Media/Research/pdf/Kit{_}of{_}Factor-Referenced{_}Cognitive{_}Tests.pdf

[16] John E Fisk and Peter Warr. 1996. Age and working memory: The role of perceptual speed, the central executive, and the phonological loop. *Psychology and Aging* 11, 2 (jun 1996), 316–323. https://doi.org/10.1037/0882-7974.11.2.316

[17] Kyung Sun Kim and Bryce Allen. 2002. Cognitive and task influences on Web searching behavior. *Journal of the American Society for Information Science and Technology* 53, 2 (2002), 109–119. https://doi.org/10.1002/asi.10014

[18] D Kirkpatrick. 1965. The Minnesota Clerical Test. *The British Journal of Psychiatry* 111, 479 (1965), 1009–1010. https://doi.org/10.1192/bjp.111.479.1009-a

[19] M Kurosu. 2015. *Human-Computer Interaction. Users and Contexts. Part III*. 575 pages.

[20] Sébastien Lallé, Cristina Conati, and Giuseppe Carenini. 2017. Impact of individual differences on user experience with a visualization Interface for Public Engagement. *UMAP 2017 - Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization* (2017), 247–252. https://doi.org/10.1145/3099023.3099055

[21] Ayelet N. Landau, Lisa Aziz-Zadeh, and Richard B. Ivry. 2010. The influence of language on perception: listening to sentences about faces affects the perception of faces. *Journal of Neuroscience* 30, 45 (2010), 15254–15261. https://doi.org/10.1523/JNEUROSCI.2046-10.2010

[22] Laure Léger and Aline Chevalier. 2017. Location and orientation of panel on the screen as a structural visual element to highlight text displayed. *New Review of Hypermedia and Multimedia* 23, 3 (2017), 207–227. https://doi.org/10.1080/13614568.2017.1399468

[23] Gloria Mark, Mary Czerwinski, and Shamsi T. Iqbal. 2018. Effects of individual differences in blocking workplace distractions. *Conference on Human Factors in Computing Systems - Proceedings* 2018-April (2018). https://doi.org/10.1145/3173574.3173666

[24] Veronica Montani, Andrea Facoetti, and Marco Zorzi. 2015. The effect of decreased interletter spacing on orthographic processing. *Psychonomic Bulletin and Review* 22, 3 (2015), 824–832. https://doi.org/10.3758/s13423-014-0728-9

[25] Richard E. Nisbett and Yuri Miyamoto. 2005. The influence of culture: Holistic versus analytic perception. *Trends in Cognitive Sciences* 9, 10 (2005), 467–473. https://doi.org/10.1016/j.tics.2005.08.004

[26] Heather L. O'Brien, Rebecca Dickinson, and Nicole Askin. 2017. A scoping review of individual differences in information seeking behavior and retrieval research between 2000 and 2015. *Library and Information Science Research* 39, 3 (2017), 244–254. https://doi.org/10.1016/j.lisr.2017.07.007

[27] Heather L. O'Brien and Lori McCay-Peet. 2017. Asking "good" questions: Questionnaire design and analysis in interactive information retrieval research. *CHIIR 2017 - Proceedings of the 2017 Conference Human Information Interaction and Retrieval* (2017), 27–36. https://doi.org/10.1145/3020165.3020167

[28] Helena Ojanpää, Risto Näsänen, and Ilpo Kojo. 2002. Eye movements in the visual search of word lists. *Vision Research* 42, 12 (2002), 1499–1512. https://doi.org/10.1016/S0042-6989(02)00077-9

[29] Ruth A Palmquist and Kyung Sun Kim. 2000. Cognitive style and on-line database search experience as predictors of Web search performance. *Journal of the American Society for Information Science and Technology* 51, 6 (2000), 558–566. https://doi.org/10.1002/(SICI)1097-4571(2000)51:6<558::AID-ASI7>3.0.CO;2-9

[30] Keith Rayner. 1977. Visual attention in reading: Eye movements reflect cognitive processes. *Memory & Cognition* 5, 4 (1977), 443–448. https://doi.org/10.3758/BF03197383

[31] Jan Recker. 2012. *Scientific research in information systems: a beginner's guide*. Springer Science & Business Media.

[32] Timothy A Salthouse and Vicky E Coon. 1994. Interpretation of Differential Deficits: The Case of Aging and Mental Arithmetic. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20, 5 (1994), 1172–1182. https://doi.org/10.1037/0278-7393.20.5.1172

[33] Joseph Sharit, Mario A. Hernández, Sara J Czaja, and Peter Pirolli. 2008. Investigating the roles of knowledge and cognitive abilities in older adult information seeking on the Web. *ACM Transactions on Computer-Human Interaction* 15, 1 (2008). https://doi.org/10.1145/1352782.1352785

[34] Edward M Silver and Corwin Bennett. 1987. Modification of the Minnesota Clerical Test to predict performance on video display terminals. *Journal of applied psychology* 72, 1 (1987), 153.

[35] Ben Steichen, Giuseppe Carenini, and Cristina Conati. 2013. User-adaptive information visualization: using eye gaze data to infer visualization tasks and user cognitive abilities. In *Proceedings of the 2013 international conference on Intelligent user interfaces*. ACM, 317–328.

[36] Ben Steichen, Cristina Conati, and Giuseppe Carenini. 2014. Inferring visualization task properties, user performance, and user cognitive abilities from eye gaze data. *ACM Transactions on Interactive Intelligent Systems* 4, 2 (2014). https://doi.org/10.1145/2633043

[37] Esther Strauss. 1983. Perception of emotional words. *Neuropsychologia* 21, 1 (1983), 99–103. https://doi.org/10.1016/0028-3932(83)90104-5

[38] Keith S. Taber. 2018. The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education* 48, 6 (2018), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

[39] Dereck Toker, Cristina Conati, Ben Steichen, and Giuseppe Carenini. 2013. Individual user characteristics and information visualization: Connecting the dots through eye tracking. In *Conference on Human Factors in Computing Systems - Proceedings*. ACM Press, New York, New York, USA, 295–304. https://doi.org/10.1145/2470654.2470696

[40] Lauren Turpin, Diane Kelly, and Jaime Arguello. 2016. To Blend or Not to Blend? Perceptual Speed, Visual Memory and Aggregated Search. *Sigir 2016* (2016), 1021–1024. https://doi.org/10.1145/2911451.2914809

[41] Maria C. Velez, Deborah Silver, and Marilyn Tremaine. 2005. Understanding visualization through spatial ability differences. *Proceedings of the IEEE Visualization Conference* (2005), 65. https://doi.org/10.1109/VIS.2005.108

[42] Ryen W. White, Susan T. Dumais, and Jaime Teevan. 2009. Characterizing the Influence of Domain Expertise on Web Search Behavior. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining* (Barcelona, Spain) *(WSDM '09)*. ACM, New York, NY, USA, 132–141. https://doi.org/10.1145/1498759.1498819

[43] G. J.S. Wilde. 2013. Immediate and delayed social interaction in road user behaviour. *Applied Psychology* 29, 4 (2013), 439–460. https://doi.org/10.1111/j.1464-0597.1980.tb01105.x

[44] Daniel Zimprich and Tanja Kurtz. 2013. Individual differences and predictors of forgetting in old age: The role of processing speed and working memory. *Aging, Neuropsychology, and Cognition* 20, 2 (2013), 195–219. https://doi.org/10.1080/13825585.2012.690364