INVITED ARTICLE

Nineteen Questions to Evaluate Typographic Research: Chaff and Wheat

Karel van der Waarde [a] and **Myra Thiessen** [b]

[a] Graphic Design – Research, Elewijt, Belgium; and Lucerne University of Applied Sciences and Arts, Lucerne, Switzerland; [b] Art, Design, and Architecture, Monash University, Melbourne, Australia

Corresponding author: Karel van der Waarde (waarde[at]glo.be)

Abstract: The number of experiments that investigate the "readability" or "legibility" of texts is very substantial. Literature reviews of these studies appear regularly, and many publications refer to these experiments to suggest evidence for claims. Some of these claims have led to usable recommendations. However, most of these recommendations are often hard to apply and unhelpful. When we are teaching typography, we struggled to explain why the recommendations are difficult to use, why many reviews are uncritical, and why experiments rarely provide reliable evidence to support design decisions. A literature review, guided by experience in both commercial practice and university level education, lead to a list of themes and issues. There are at least 19 reasons why the results of many typographic experiments need to be questioned. This article provides 19 guidelines that could be used to evaluate experimental research into the ways in which texts are read. This list of reasons can be used as a checklist to assess and guide new typographic experiments. We hope to make sure experiments are worthwhile, future reviews are based on reliable sources, and recommendations are effective.

Implications for practice: There are three practical applications of the findings of this review. Firstly, the 19 guidelines might help to critically review experimental findings and assess if they are relevant for practice — Table 1 is a handy checklist for this assessment. Secondly, the review shows that a typographic practice must be reader-focused. It is essential to involve readers throughout design processes, especially when the intention of information is to enable people to act. Performance criteria, evaluation methods and performance levels need to be relevant for readers. The result of this involvement is qualitative: a single remark from a single person can change the frame of a design project. And thirdly, the review shows that it is beneficial to look more intentionally at differences across readers and across reading activities. Involving people with

@: ISSUE〉 ARTICLE〉

Cite this article:

van der Waarde, K., & Thiessen, M. (2025). Nineteen questions to evaluate typographic research: Chaff and wheat. *Visible Language, 59*(1), 77–99. https://www.visible-language.org/59-1/nineteen-questions-to-evaluate-typographic-research-chaff-and-wheat

First published online April 27, 2025.

https://visible-language.org/journal/

Visible Language Consortium:
University of Leeds (UK)
University of Cincinnati (USA)
North Carolina State University (USA)

different backgrounds and experiences will provide new insights into the ways visual information can be interpreted and applied. Listening and observing are fundamental design skills that need to be trained and honed. Even after decades of practice, it remains a humbling experience to find out how people really look at visual design.

Keywords: legibility; readability; research methods; typographic research; typography

1. Legibility and Readability Research: Some Starting Questions

Typographic research investigates the ways in which the visual format of a text affects how people read and understand a text. These experiments are usually classified as "legibility research" or "readability research." Although several attempts have been made to clearly separate "legibility research" from "readability research," these terms remain opaque (Luna, 2018; Gonzales Crisp, 2012). Reading is a complex activity that starts from recognizing individual letterforms and word shapes and ends with the conscious interpretation of continuous text. In this article, we group all experiments that investigate relations between visual texts and the reading activities under a more general "typographic research." We focus on the Latin script because most of the research has used test materials with this alphabet.

This article only includes those experiments where people were asked to read a specific text. Experiments without actual readers, such as the research for "readability formulae" (Dubay, 2004) or other "expert evaluations" (Schriver, 1997) are not included. Furthermore, the focus is on "continuous text," and not on the letterforms of individual characters (Bigelow, 2016).

In our practices, we noticed the following:

1. Teaching typography and providing typographic advice that is based on evidence is difficult. The published recommendations somehow do not seem to be applicable (Gonzales Crisp, 2012; Schriver, 1997). This can hamper education and the status of the design profession, which might seem to some to prioritize craft skill and aesthetics over purpose and social outcome (Thiessen & Kelly, 2019; Frascara, 2022).
2. Reviews of the publications about typographic experiments, which is an integral part of any experimental study, keep referring to the same, sometimes outdated, experiments. These reviews rarely look critically at the relevance or applicability of the experiments they cite and take the outcomes for granted. This hampers new research.

3. Carefully applying the recommendations does not make texts more readable or more legible. Recommendations are often in conflict with each other, are not prioritized, ignore genres and readers, and do not specify a context or a language. This hampers practice and reduces the confidence in typographic research.

We look first at each of these three experiences. After that, we present 19 questions that might prevent some of the problematic repetitions in research, education, and practice.

1.1. Evidence for Typographic Decisions to Be Used in Education and Practice

In practice and in education, it is very difficult to apply any of the typographic recommendations outlined in research experiments. They require informed and specialized interpretation and often are not literal in a practical sense in that they cannot typically be applied universally.

Application in education. Students cannot apply the guidelines in their work and in observation they can find it difficult to understand why the information is important.

For example, co-author Thiessen draws regularly on scientific studies in her own typography studio teaching and thinks that studies such as Dyson and Haselgrove (2001) and Dyson (2004, 2013) are important examples to illustrate that reading actions and behaviors are different on screens compared to print media. Principles cannot be directly transposed from print to screens. Dyson and Haselgrove show that when reading from screens, readers can be quite proficient with line lengths up to 100 characters, which is substantially longer than the accepted 60–75 long believed to be ideal for printed materials. There are a few reasons that this may be the case, including issues surrounding reading distance, screen size and resolution, or paging/scrolling functions.*

However, in response to this discussion, Thiessen regularly sees progress work from students that contain line lengths in literal translation of Dyson and Haselgrove's results. In these cases, the application of "experimental results" in practical guidelines does not lead to the required results. It does not result in a text-design that is comfortable to read and that relates to the design of a genre. This observation points to an important gap in the teaching of typography. It suggests that students are lacking exposure to, first, rigorous typographic research publications and, second, the practice of doing scientific methods themselves to investigate typographic problems. These are both important skills for enabling capacities to question and test the function and accessibility of reading materials and developing a basic understanding of how to translate

* Technology has obviously progressed substantially since Dyson and Haselgrove published their study; however, we think this only makes our point here stronger.

an experiment's result for application. Typographic designers working in any range of professional practice would find value with these skills.

Application in practice by designers. It is common to hear phrases like "designers always have to 'break the rules' to make effective typographic designs" (Keedy, 1993). This can be misleading — or further, false — if the "rules" are not clearly defined, esoteric, or if there is no clear evidence supporting the so called "rule." It is worth looking at "best practice" to see which practical guidelines are recommended in particular contexts, but it is important to keep in mind or to question the reading condition the "rule" has been designed for. As mentioned above, reading on screens is proving to be very different from reading print based materials. So too are different environments, different reading actions, and different readers.

Functional reading (Thiessen et al., 2020) examines reading contexts including environmental factors such as lighting, stresses, distractions, awareness of the reader, and reader differences, to determine how to best support a particular reading action. If the guidelines are worth applying, then they should consider aspects of functional reading and be flexible and clear enough to respond to specific contexts. Similar arguments are made by Waller (2012), Moys (2017), Noël et al. (2019), and Larson and Picard (2005). They all attempt to make evidence-based design decisions and apply the recommendations, but the effects on readers do not seem to be directly related to the suggested generalizable outcomes.

For example, the recommendations that *line length, line space, and type size are directly related; if the line gets longer, more line space is needed; if the type size gets larger, more line space is needed* are repeated (e.g., Luna, 2018, p. 109; Spencer, 1968, p. 55), without mentioning in which context and in which kinds of genres for which kinds of reading, and for which kinds of readers the resulting text would be readable or legible. Van der Waarde (1999) shows that designers do divert from published rules, but clearly adhere to established visual patterns in novels, academic journals, and brochures.

Application in practice by non-designers. We think considering the non-designer is also important because some publications about typographic guidelines are read, cited, and used by people who are not trained to consider typography or scientific research — sometimes in ways that are worrying. For example, whether or not difficult to read, or disfluent, typefaces are valuable for learning has attracted attention in both typographic and cognitive psychology circles (Thiessen et al., 2020). A series of experiments were published assertively concluding that school age learners tested better when they were learning curriculum using reading materials set with difficult to read typefaces. What is worrying about this paper is that it suggested to teachers that presenting content in a way that is hard to read is a cost effective way to help their students learn more (Diemand-Yauman et al., 2011). Other examples have suggested that difficult to read

typefaces improve memory (RMIT, 2018). However, the problem of disfluency and its relationship to learning is far more complex. At best, there will be no effect, but if reading materials are hard to read, it is more likely that teachers reduce learners' motivation to engage with content or even impede learning outcomes for children who might struggle with reading (Astley et al., 2023).

Typographic solutions are rarely "one-size-fits-all," and it is essential to consider the consequences of poor guidance for non-typographers and to provide limits where the advice does not apply. It is therefore necessary that there is a clear link between the experimental data, the conclusions, and the recommendations. A "strength of evidence-scale," as it is used in the medical world (Jerkert, 2021), might be worth considering.

1.2. Uncritical Reviews

Some reviews and summaries that attempt to translate scientific research seem to uncritically quote experimental results and copy its advice and guidelines rather than contextualizing it for practical or educational application. These reviews tend to follow a tradition that draws on the same questionable experiments and repetitively extrapolate the same doubtful guidelines.

There seem to be three main reasons.

- *Reason 1: People ignore different research approaches.* It is important to note that there is a key difference in the way psychology and typography approach legibility research. This difference impacts the sort of research questions that are asked and how experiments are subsequently designed. Dyson (2013) helps by separating the aims of psychologists and typographers. Psychologists are primarily concerned with the mechanisms by which we read and the differences across people. These are "how questions." Typographers are more concerned with the materials used for reading, the environments in which reading actions take place, and the goal the reader aims to achieve. These are "what questions." Answers to the "how questions" explain why things happen. These do not suggest what to do to make things happen.

- *Reason 2: People do not question generalizations.* Reading activities and behaviors are contextual (Britt et al., 2022). This makes the replication of specific results more difficult across different kinds of readers, reading materials, and reading environments. It means that individual studies may only be able to tell us about very specific reading conditions. The outcomes of specific experiments cannot easily be directly translated into generalizable findings.

- *Reason 3: People do not question the reliability and validity.* In contrast with other academic disciplines, such as medical, pharmaceutical or educational

psychology, experimental results are not always scrutinized thoroughly by typographers, and the link between experimental findings and generalized guidelines are not always carefully considered.

Reliable recommendations are based on experiments. These experiments need to be replicated. If the same test materials are used again in a similar context and study design, the same result must come out if they are to be trusted. For example, typographic practitioners have held onto the belief that the word-shape reading model is favored by fluent readers. The suggestion is that the activity of reading consists of recognizing the shape of whole words. This idea has a long history (Cattell, 1886). The most often cited article (Bouma, 1973) seemed to provide evidence, but a closer reading reveals that this is based on a misunderstanding (Larson, 2004). Experimental evidence does not support the word-shape reading model.

However, striving for reliability is still important to further our understanding of the performance of certain variables that might be important for a wider range of readers or conditions. Examples of these variables are motivations of readers (attention and interest), situations in which a text is read (stress, lighting), and availability of alternative information sources (practical validity). Understanding this distinction between "significant research results" and "reliable recommendations" makes critically evaluating legibility studies for typographic application even more important. A result in any given experiment might be significant, but if it cannot be replicated under similar conditions, the result is only meaningful to the circumstance in which it was created. This is not in itself very useful for furthering knowledge or understanding an effect but could point to something unique about the experiment itself. There might be an influencing factor that was not controlled for in the original experiment related to environment, material, context, or participant.

1.3. Generalization of Research Findings into Recommendations

One of the stark characteristics of typographic recommendations that are based on experimental studies in readability and legibility is that they have hardly changed since the first articles appeared about 140 years ago (Javal, 1878).

The reason why this is problematic is because reading is cultural, and the cultures of reading have changed substantially in this time. Readers and reading materials and behaviors are very different from those seen 140 years ago, with observed differences in how children develop reading skills. In some cases, children show a marked reduction in motivation and underdeveloped skills associated with deep reading (Wolf, 2018). Where once, reading was primarily a private and solitary act, it is now very public and permeates all parts of modern life. For some, communication by email or by text is preferred over face-to-face, telephone, or videophone conversation. These shifts in

reading behaviors have been enabled and shaped by changes to reading objects and environments. Where readers once relied on books and paper-pamphlets, they are now very likely using screens far more than printed materials. Furthermore, where reading may have been primarily an act undertaken for long periods of time, today readers may find they are expected to act or take action in response to a sign or label with only a few words or symbols. The introduction of the internet has also had a considerable effect on how readers engage with texts and their expectations related to both reading and content. As Carr (2010) explains, no longer is reading a "slow drip" of content for contemplation and integration with previous knowledge and experience. Rather, the internet has provided the means to access a flood of information, changing reading action from a substantiated task to something more akin to "skimming": seek and search. The result is readers who are less able to concentrate for long periods of time, and are now very practiced at darting around a text and picking up small snippets of information (Carr, 2010).

The aims of this article are:

- To prevent the repetition of uncritical citations and reviews by pointing out some of the issues. These comments might form a basis for further research.
- To reduce the attempts to apply unsupported recommendations ("rules") in education and in practice. The advice about typographic specifications that educators and researchers provide needs to be reliable and evidence based.
- To support practitioners with effective instructions and reliable advice about the visual design of texts.

Our ultimate aim is to examine norms in typographic research for their continued value. If research is conducted correctly, it is more likely that the resulting recommendations will lead to improved text-designs because they consider aspects related to functional reading and thus improve reading experiences and outcomes.

2. Nineteen Questions

The 19 questions we suggest below explore reasons why much legibility research that has been undertaken and reported needs to be reconsidered. Some are more serious than others, but they are all fairly damaging, in our view. We would like to stress that it is not our aim to list publications or shame authors who do this and have in other ways done very good work. Typographic research often seems to be based on questionable assumptions. These only come to light when many studies are compared, and patterns start to form. We did therefore not include the references to authors of reviews or recommendations who expected that the assumptions of previous researchers were correct. We may all be guilty of this at some stage in our research careers.

This list is not comprehensive. We raise these particular questions based on our observations and recognize our own interests and biases in them. However, it is our goal not to present definitive conclusions but to draw attention to what we see as a problem and, hopefully, incite some discussion and action that aims to improve the situation. In our opinion any of these 19 questions, in any combination, should place serious doubt on the validity of original data experiments in typography research.

The questions are grouped under five headings:

1. Participants,
2. Test materials,
3. Measurements and criteria,
4. The experiments,
5. Sources: authors and sponsors.

These themes were selected because each of these needs to be considered in the evaluation of an experimental study. They determine the accuracy, validity, and reliability of the results. Replication of an experiment is not really possible without an exact description of these five themes.

2.1. Participants

Question 1: Reading processes — Does the study describe the readers, conditions, and models? Reading is a complex cognitive activity. Models for how a reader can identify and comprehend letters and words were debated in early psychology literature (Cattell, 1886; Javal, 1878). However, it is relatively understood that readers mostly likely use a method of template matching where one uses specific letter features to identify letters (Grainger et al., 2008).

Experienced readers also likely rely on several strategies simultaneously during extended reading tasks including phoneme mapping to sound out unfamiliar word and semantic context to anticipate upcoming words (Dehaene, 2009). However, readers may differ dramatically and any reader who experiences any kind of difficulty or impairment that interferes or obstructs the reading process may not behave in expected ways.

Experiments that do not describe readers or reading conditions and any factors that deviate from what might be considered typical could be naively assuming that “all reading is equal” and that “all people read in an identical way.” For example, most experimental environments are relatively sterile set ups with good reading conditions. The experiment is conducted in well-lit rooms with little to no distraction unless it is part of the task and readers are fluent and are unlikely to have experienced a reading or language related difficulty such as dyslexia or aphasia.

If an experiment does not recognize reader differences and reading strategies that may be at play, be cautious.

Question 2: The starting point — What do we know about the participants? Was there a test beforehand, or some detailed description, that outlines the reading ability, existing knowledge, language skills, and motivation of participants? If not, discard the results. Without first understanding the reading level and capabilities of the readers that comprise participant groups it is impossible to know the extent that reading capabilities may differ. Reading is a learned skill that requires practice and maintenance, an individual who reads extended texts for three or more hours a day will likely read faster than someone who reads a lot less, reads infrequently, or reads primarily messages and e-mails (Suk, 2016). Readers will read new and more complex information slower than content they are familiar with (Schriver, 1997).

Readers who are marginalized for any reason do not typically form part of participant groups. Marginalized readers may have a different reading behavior due to dyslexia or compromised linguistic ability because they are reading in a second or third language (i.e., not their native language). Or they might have low vision, memory issues, or learning difficulties. These are not the groups of readers that typographic research is often most concerned about.

However, these capabilities are fundamental for functional typography. Typographers are designing texts for specific readers to support a specific reading task (entertainment, learning, searching, instruction) in a specific environment (library, home, café, public transport, driving). Readers may have any range of ability or disability, and these must be considered before results from experiments can be drawn.

Question 3: End points and aims — Is there a clear purpose for the person who reads? People read texts for different purposes: reading to do, reading to learn to do, reading to enjoy, reading to assess (Schriver, 1997). In these actions readers adopt different strategies such as skimming, scanning, slow careful reading, or searching (Muijselaar & de Jong, 2015). Ignoring these differences and not discussing the specific action, content, and context an experiment aims to support will likely invalidate the conclusions and make it very difficult to generalize the results and apply them in real world situations. No one reads a text for no reason at all; there is always a goal.

Question 4: Different people — Are the participants university students? Many experimental studies have been conducted on students in a university environment, most often psychology students. The underlying assumption is that this is a homogenous group. However, anyone who has ever taught a classroom of students knows that there are many differences within a group caused by motivation, experience, reading skills, and so on.

Psychology and related fields are desired programs of study, and they are therefore able to accept the brightest and most capable students from the range and applicants. This means that they are more typically skilled and experienced readers. And they often come from a certain position of privilege in society that has enabled them to enroll in a university program. Students may also be more practiced in experiments than more diversely represented participants since they are regularly recruited for experiments due to the nature of their program of study. Although not likely intended, this practice and lack of diversity may be perpetuating cultures of exclusivity and exclusion through the reading materials that are developed using these study results. It is furthermore unlikely that student populations can be compared over time or geographical locations. The reading behavior of students in the 1950s in the USA might not be comparable with the reading behavior of students in the 2020s in Australia.

These first four questions indicate that reading processes, starting points, aims, and participants vary. These questions need to be taken into account when experimental studies into the effects of the typography of reading materials is compared.

2.2. Test Materials

Question 5: Repeatability — Are the test materials available or is there an accurate visual representation? A verbal description of typography test materials is not sufficient, and it is impossible to evaluate the research if no accurate visual representation of what was tested is provided. It is also important to know how the material was finished. If it is meant to examine print materials, then dimensions, paper stock, and details of the print output are important if there is any hope of replication. For experiments that test onscreen reading a record of the type of screen, the resolution, colors, and sharpness should be documented. It is essential to record what participants looked at exactly during an experiment. Some test materials might have been archived, but many have been lost, deleted, or can only be shown on obsolete technology.

Question 6: Design of materials — If the test materials are shown, are they appropriate for the research question? Evaluate the test material to determine whether it is appropriately designed to address the research question. Many test materials have been criticized for poor design or because they use a typographic specification that would not occur in practice. However, it is important to understand that the material may be very effective for isolating a specific variable and in some cases, it is necessary to push the boundaries to investigate a hypothesis. When evaluating the appearance of test material, it is essential to consider if these materials address the research question. The design of reading material plays a very large role in how a reader interacts with it, as well as their capacity to perform any range of cognitive tasks (Walker, 2001).

In some cases, disciplinary differences play a role in how materials are reported and the value placed on their development. For example, a psychology researcher may be less concerned about the visual appearance and output criteria of test materials because their research questions tend to be framed around mechanisms that influence perception, behavior, or cognition. However, aspects of these mechanisms that are related to reader impression, motivation, and belief (Song & Schwartz, 2008, 2010), along with certain reading processes (Thiessen et al., 2022) are closely tied to the visual appearance of a text. This makes accurate reproductions of the materials important, especially in order to translate results for typographic practice.

Question 7: Boundaries — Are the recommendations generalized without any genre-limits? Research is often undertaken with type stimulus presented onscreen, but the results are often suggested for application across a range of print media with little critical evaluation of the likely differences in reading behavior media causes. Reading acts are different across reading materials because the reader's goals are different. Reading a newspaper differs from reading an online instruction because readers come to reading tasks with different expectations. Newspapers are designed to support a nonlinear reading strategy where the reader can scan headlines quickly and dip in and out of an article. Online instructions are designed for step-by-step reading and thus support a specific sequence of consecutive activities. Supporting different reading goals has led to different genres that are based on different typographic configurations (Moys, 2013). For example, the line length in a paper newspaper differs from the line length in an online instruction because the reading strategy employed by readers engaging with these materials is different. To suggest that there are "optimal factors" that can be applied across reading objects and tasks disregards the differences between genres and is not supportive of a variation of reading strategies.

Question 8: Language characteristics — Is the language, alphabet, and/or writing system defined and its unique characteristics described? It is common to assume experiments have been undertaken in English, or that what is appropriate for English will also work for other languages. This is not the case. Word length can vary dramatically across languages, and this may influence optimal line length and necessary line space. Avoiding hyphenation is a common recommendation for typesetting English but this could be far more problematic for Dutch, Danish, or Finnish where words, on average, have more characters. Plus, the use of diacritical marks and accents can also influence reading. Because of the discernability of diacritical marks, it is necessary to design texts in French with more vertical space and a slightly larger x-height than texts in Dutch. Data collected using Latin script is not likely to extrapolate to languages using other scripts such as Greek, Hebrew, or Cyrillic, or to scripts like Arabic, and languages that use characters like Korean, Japanese, or Chinese languages. There is no one-size-fits-all solution to typographic design across languages or scripts.

Question 9: External validity — Does the study deal with realistic looking texts and/or realistic reading scenarios? Test materials that are used in typographic experiments must be directly related to practical uses. The less realistic it is, the less it can speak to functional reading scenarios. Materials that are designed to more closely replicate realistic reading material can speak more broadly about reading behaviors and performance based on typographic variables. For example, texts need to be read under pressure because the reader is on a motorway and searching for specific instructions, they may be reading under low lighting conditions, or they may be reading from a backlit device. We still lack sufficient knowledge about the impact of environmental factors more broadly and the various distractions a reader engaged in continuous reading actions may encounter at any one time on their reading and cognition.

This is not to say that studies that evaluate how readers are able to identify letters or words in isolation are not valuable. These studies can tell us a great deal about legibility (Beier, 2012), how letters are identified (Pelli et al., 2006), how word reading is impacted by environmental visual noise (Sawyer et al., 2020), and how cognitively demanding the process of reading is (Thiessen et al., 2015). However, this is only one piece in the puzzle and these studies are limited if the results have not been tested for reliability under more realistic reading conditions.

Question 10: Variables — Is the combination of typographic factors described? It is usually acknowledged that it is "a combination of typographic factors" that makes a text legible. These combinations include factors such as typeface, type size, line-space, line length, color, and type-weight. Usually, these individual factors are described without reference to each other. Just investigating a single variable, without acknowledging the interactions between these variables, invalidates many typographic studies.

Very few publications bother to describe the non-tested typographic factors, such as the paper quality, the dimensions of the margins, the dimensions of the paper, or the characteristics of screens, all of which contribute to functional readability.

Question 11: Date of studies — Does the study rely either moderately or heavily on old or outdated science? As mentioned above, cultures of reading change over time and this means that scientific studies undertaken more than 20 years ago* will not be able to inform typography today. Unfortunately, it is common to see work from before 1950 turn up in reference lists. These studies were limited for reasons related to technology, such as a limited capacity to modify type sizes without very time consuming and costly type-setting compositions. This along with the shifts in reading cultures seen since these early works make it difficult to see how knowledge about reading behavior and reading materials produced using letterpress will be able to satisfactorily inform

* Twenty years is even a generous timeframe in our opinion.

typographic design output with modern offset or digital presses. Even more worrying is to think those historical studies can inform screen reading.

Technological advances have also been instrumental in how reading and reading materials have evolved. Consider how mobile technology has changed the way readers engage in reading acts and how their expectations for how information is delivered has changed — and this can be observed over a very short period of time, relatively speaking. Historically, reading acts, typographic preferences, and what readers will tolerate has changed very slowly, but it has changed. As the mechanics for developing texts changes so do the materials that are developed, and readers change in response.

This means that experiments that examine a reader's response using test stimuli that is no longer relevant are not useful for typographic design today, nor will they say anything about contemporary readers.

In conclusion, these seven questions about typographic experimental materials indicate that it is essential to look at the original test materials, genres, scripts, practical validity, combinations of typographic variables, and reproduction technologies. Again, the variation of these factors in the literature is substantial, and it is unlikely that generalizable conclusions across these factors can be drawn.

2.3. Measurements and Criteria

Question 12: Measurement units — Is the type size specified in points? If the answer is yes and different typefaces are being compared, the results and recommendations are not usable. "Point sizes" are not directly related to the vertical dimension of printed letters, not in print nor on screens.

Every character in the Latin script is contained in a rectangle. A designer of a typeface can choose the vertical dimension of each character within this rectangle, as long as all characters are positioned on the same baseline. A second complicating factor is that the actual dimensions of "a single point" have changed several times (Boag, 1996). The points that we use at the moment are PostScript points. Although the difference might seem small, it is significant because there is no guarantee that the dimensions are kept identical in this conversion. Even as a plain description as "12 point Times New Roman" does not give an exact dimension. Which points? Which "Times New Roman"? Which technology?

The combination of these variations in typefaces and point sizes makes it impossible to estimate the size of the type used in experiments or recommendations. Type size is an influential variable in typographic research, and comparing different fonts at the same point size will not result in useful data.

Question 13: Appreciation — Is aesthetic preference considered? A reader's preference for particular presentations of information do not always correlate with their capacity to use them (Wright, 1979), and the impact of preference for typography is seen from the first stages of reading development (Thiessen & Dyson, 2009; Walker & Reynolds, 2003). Although aesthetic preference may not factor into reading performance, it has shown to be influential in capturing attention, influencing motivation, and affecting the mood of the reader (Larson, 2007). One cannot lose sight of how influential motivation is on reading acts and that motivation is closely tied to aesthetic preference. If a reader does not want to read something because they do not like it, they will not read it.

Question 14: Reading measure — Is "the speed of reading" really an appropriate measure? Unless the differences are substantial, readers are unlikely to care about, or even notice, differences in their reading rate. However, in a typographic experiment, changes in reading rate can suggest particular typefaces or typographic arrangements are easier or harder to read. For example, disfluent typefaces are likely to slow down reading due to their visual complexity (Thiessen et al., 2020). This means that readers may need to exert more cognitive energy to perform simple tasks related to letter and word identification. Since the working memory is limited in capacity this means that the reader could struggle to perform higher-order tasks necessary to interpret and assimilate the content they are reading (Thiessen et al., 2015), which is an undesirable outcome for the reader. So, while a reader may be unconcerned with small changes in their reading speed, this can provide a way to identify and improve factors affecting reading efficiency.

However, reading speed as an isolated measure is not likely to provide a very clear picture about performance and is not sufficient to evaluate typographic design. If this is the only measure and it is uncontextualized, approach with caution.

In conclusion, selecting appropriate dimensions and criteria in typographic experimental research has proven to be challenging. Traditional point sizes are not reliable, aesthetic preferences are hard to determine, and criteria such as reading speed are hardly relevant for readers.

2.4. The Experiments

Question 15: Context — Is the experiment related to reading in real life situations? When drawing on experimental research to inform typographic practice it is important to understand that each study can only speak to one narrow set of variables, and that the study must be tightly controlled to eliminate any distractions or unintentional effects. In order to achieve this, many studies take place in laboratory settings and bear very little resemblance to environments in which texts are normally read. Laboratory settings are important to ensure that measurements are accurate and able to address

the intended variable manipulation. It is clear that scientific methods have been able to tell us a great deal about people and reading behaviors. However, these methods are limited in their capacity to examine reading in real life situations. The environments in which people normally read are riddled with distractions that affect attention and comprehension.

A successful typographic experiment is one that considers functional readability (Thiessen et al., 2020) and aims to create more ideal reading scenarios through the combination and evaluation of the science, craft skill, reading environment, and reader goal. For example, will there likely be noise, poor lighting conditions (you do not want to keep your partner awake), painful reading positions (not enough pillows, poorly designed sofas), or environmental distractions (children, television, colleagues)? Data generated in laboratory settings can only speak to isolated factors and only speculate about specific aspects related to functional legibility. The best way to understand how typography functions in a broader sense is through combined laboratory and real-world evaluation.

Question 16: Global standard — Does the experiment aim for a gold standard? Is the aim of a series of experiments to find a "gold standard"? The assumption that it is possible to find an ideal combination of typographic variables that is effective across different contexts is incorrect. It is a fallacy. Reading depends on the combination of a text (visual material), a reader (eyesight, background knowledge, aims, intentions, linguistic knowledge), and a situation (late at night, on a beach, in a train, in an office). Examining only text variables ignores functional reading principles and will not lead to more usable typographic outcomes.

Question 17: Impact — Do the authors overclaim impact? When undertaking research, one typically has vision for how the work might expand or in what contexts it may be translated and applied. However, it is important to represent results accurately and avoid inflating the importance of individual or isolated experiments.* As we have seen in the points above, typography is a complex and dynamic system of interconnected variables and readers are individual cases with unique challenges and gifts. Claims about easy fixes or that altering single variables, like a typeface, are effective are likely to be untrue.

In conclusion, the context in which typographic experiments are conducted affects its results. Conclusions drawn from data in specific contexts are unlikely to be generalizable or have a similar impact in another context.

* We recognize that bodies of work that have been developed over long and focused careers have impacted the fields of legibility, perception, and typography in profound ways (for example, Denis Pelli or Gordon Legge). This is not the work we allude to here.

2.5. Sources: Authors and Sponsors

Question 18: Credibility — Are the sources credible? Uncritically accepting everything that has appeared in academic journals without checking might lead to the wrong conclusions. A prime example are the publications by Prof. Dr. Sir Cyril Burt (e.g., Burt, 1959). Although he published in the most prestigious journals and with respected publishers, his typographic experiments were discredited. There is simply very little truth in Burt's work in typography (Hartley & Rooum, 1983). Since this work proved to be fraudulent and of no value, articles that quote it as valid typographic guidance are called into question for their rigor, accuracy, and validity. A more recent example can be seen in the debate surrounding the typeface Sans Forgetica, which saw claims suggesting that using the font would improve memory by drawing on the principle of desired difficulty (RMIT, 2018). Details of the experiments undertaken by the font researchers and developers are not published to our knowledge and subsequent studies have not been able to replicate the same effect (Geller et al., 2020; Huff et al., 2022).

Referring to disproven claims shows limited engagement with the academic debate in typography and psychology and is simply bad academic practice.

Question 19: Sponsor — Who paid for the experiments? One of the important influential factors is the sponsor of the research. It is always worth asking who paid the researchers, who paid the participants, and who will benefit from the publication of the results. Legibility research paid for by Microsoft in 1996 and undertaken by Carnegie Mellon University concluded that screen fonts Georgia and Verdana — which were specifically designed for Microsoft — were more legible on screens than Times New Roman (Boyarski et al., 1996). Is it likely that any other result would have been published? Although it is not very common in legibility/typographic research, it is an important question to ask, nonetheless.

In conclusion, a check of the authors and their sponsors might reveal reasons to critically reconsider the outcomes of typographic experiments.

3. Discussion: What Can We Learn?

Based on the previous 19 questions (reiterated in Table 1), we propose two recommendations for typographic practice and research and see these as opportunities to move forward and strive to achieve conditions grounded in principles of functional readability (Thiessen et al., 2020).

First, in opportunities that allow it, adopting a typographic practice that is reader focused and works to integrate the processes of writing, designing, and testing is more likely to result in texts that are suitable for specific readers and support them through

Table 1. Nineteen questions for evaluating typographic research, with two recommendations.

Participants	1	Reading processes	Does the study describe the readers, conditions, and models?
	2	The starting point	What do we know about the participants?
	3	End points and aims	Is there a clear purpose for the person who reads?
	4	Different people	Are the participants university students?
Test materials	5	Repeatability	Are the test materials available or is there an accurate visual representation?
	6	Design of materials	If the test materials are shown, are they appropriate for the research question?
	7	Boundaries	Are the recommendations generalized without any genre-limits?
	8	Language characteristics	Is the language, alphabet, and/or writing system defined and its unique characteristics described?
	9	External validity	Does the study deal with realistic looking texts and/or realistic reading scenarios?
	10	Variables	Is the combination of typographic factors described?
	11	Date of studies	Does the study rely either moderately or heavily on old or outdated science?
Measurements and criteria	12	Measurement units	Is the type size specified in points?
	13	Appreciation	Is aesthetic preference considered?
	14	Reading measure	Is “the speed of reading” really an appropriate measure?
The experiments	15	Context	Is the experiment related to reading in real life situations?
	16	Global standard	Does the experiment aim for a gold standard?
	17	Impact	Do the authors overclaim impact?
Sources: authors and sponsors	18	Credibility	Are the sources credible?
	19	Sponsor	Who paid for the experiments?
Recommendations	1	Adopt a typographic practice that is reader focused and works to integrate the processes of writing, designing, and testing.	
	2	Look more intentionally at differences across readers and across reading activities in typography studies — rather than focusing on detailed manipulations of layout.	

the reading act. More often than not this requires an approach that draws on multi-disciplinary expertise and knowledge so one can best understand the specific reader, the environment in which the reading act will take place, the content of the text, and what is the desired action or outcome. A text, inclusive of all its component parts, such as illustrations, captions, and headings, is written for a specific group of readers in a specific situation. Based on integrated knowledge of existing frameworks and empirical research results, a prototype is developed that shows what the result might look like. Best practice suggests that diagnostic and user tests* are conducted to confirm or disprove assumptions that were made during writing and designing. Based on performance results and preference data collected during conversations with readers, the prototype is modified and can be tested again.

This means that the participants, test materials, measurements/criteria, and contexts are completely integrated into a larger project. This approach allows for different models of reading, genres, languages, and characteristics of different readers. In usability tests, or reader-interviews, or participatory design, they all provide reactions and feedback about both the contents as well as the typographic design of specific information in a specific context. An approach such as this allows typographers to tailor texts to specific kinds of readers using specific kinds of texts, but it should not be assumed that these results can be extrapolated to texts or readers more broadly. This approach is ideal when the need and opportunity to create bespoke documents is present.

Second, it may be more productive for typography studies to look more intentionally at differences across readers and across reading activities rather than focusing on detailed manipulations of layout. Reading contexts differ dramatically as well as the intent of the reader and the reasons they engage with texts at all. Consider the commotion and high stakes of a hospital emergency room and the importance of administering the right medicine at the right dose. What if that reader was tired or distracted during any of this action? How might that impact reading? What if they are dyslexic?

The consequences of these two recommendations are likely to affect four areas: typography research, the design of materials for experiments, education, and practice.

Typography research that investigates hypotheses that are focused on material design rather than on the reader may be limited in capacity to further the progress of the field more generally. Typographic design is contextual, and every reading action is different, but typography must find a way to generalize from the knowledge generated

* Dyson (2017) provides a thorough examination of research methods relevant to design for reading and offers a discussion of how and when this range of methods may be useful, including historical accounts, applying frameworks, drawing on heuristics or expert opinion, diagnostic testing, user research, and empirical research experiments.

scientifically and from the rich history of craft knowledge to create reading objects and scenarios that are usable and useful. Typography is likely to find more value in examining how different readers interact with similar materials under similar reading conditions. This is likely to say more about design for functional readability. This suggests that typography research that is collaborative and multidisciplinary will be most productive for the field of knowledge today.

With this said, however, it is also important for typographic researchers to better communicate their value within a collaborative model and to be clearer about the importance of typography at problem framing and study design stages of research — it cannot be an afterthought. Discussions in this realm more readily understand the value of psychology to typography but the value of typography to psychology is less well understood. When designers are included in study framing and design, the question and hypothesis change. More considerations can be given to how the reading materials might be impacting reading processes and reader behaviors so to create a better understanding of readers as well as reading material.

One key concern in education is the limited exposure students in undergraduate and postgraduate design programs have to research methods informed by the social sciences. This leaves many practitioners with a limited knowledge of the contribution psychology has made to reading research as well as minimal knowledge in how those studies might be interpreted. This has resulted in a history of perpetuating ideas. This article aims to point to some of the issues caused by the limited exposure by providing questions that could help to start assessing an experimental study.

Lastly, it is important for practitioners to recognize that psychology is asking different questions than typography researchers and this dictates the data that is collected and how it can be applied. This does not mean that the data is not informative, but that it can only speak to a narrow set of variables that are often not practical for typographic application without interpretation and a compilation of the knowledge across the body of research and broader reading contexts. This implies that typographic practitioners need to be widely read and actively generating a knowledge of good typographic practice by stitching together the results into a cohesive story. This also requires active critical reflection of their own work and a thorough knowledge of experimental methodology so to test the knowledge generated in labs in actual reading environments. Just stating that “science is not forthcoming with a seamless web of rules” (Lupton, 2004) is an indicator that practice has unreal expectations of experimental research.

4. Conclusion

Of course, this is not the first article that claims to question the quality and validity of typographic research. Wheeler (1928), Spencer (1969), Lupton (2004), and Ole Lund's 1999 thesis clearly indicate the severe shortcomings of studies investigating the difference between the legibility of serif and sans-serif typefaces. Rob Waller's discussion about "single typographic variables" (1991) and Karen Schriver's hesitation (1997) should be used as pointers that the pre-1985 typographic research needs to be discarded.

The lack of application of the results of typographic research in practice is probably the most damaging critique. If the results were clear, helpful, and effective then they would be used immediately. We see this as a problem of the narrow focus of disciplinary approaches to research across both psychology and typography.

Since typography is typically concerned with the reading objects, research questions rarely examine individual reader differences. On the other hand, psychology is very concerned with individual reader differences but does not focus on an understanding about how the reading object and environment affects the reading act regardless of who the reader might be.

By listing the issues, we suggest four considerations that avoid the pitfalls of the readability/legibility research:

1. Typographic research must be interdisciplinary and collaborative;
2. Test materials must be based on best practice and have a high practical validity;
3. Design education must include a critical approach and scientific methods;
4. Researchers must focus on a clear user-action and establish the differences between people.

It is clear that there is never a single way of reading, and that a single typographic design of a text cannot suit all readers. We need to find out what kinds of reading people apply to different kinds of texts.

5. References

Astley, J., Keage, H., Kelson, E., Callahan, R., Hofmann, J., Thiessen, M., Kohler, M., & Coussens, S. (2023). Font disfluency and reading performance in children: An event-related potential study. *Brain and Cognition, 169,* 105986. https://doi.org/10.1016/j.bandc.2023.105986

Beier, S. (2012). *Reading letters: Designing for legibility.* BIS publishers.

Bigelow, C. (2016). Typeface features and legibility research. *Vision Research, 165,* 162–172. https://doi.org/10.1016/j.visres.2019.05.003

Boag, A. (1996). Typographic measurement: A chronology. *Typography Papers, I,* 105–121. https://typography.network/wp-content/uploads/2024/04/Boag_TypPp_1_Typographic_measurement_a_chronology.pdf

Bouma, H. (1973). Visual interference in the parafoveal recognition of initial and final letters of words. *Vision Research, 13,* 762–782.

Boyarski, D., Neuwirth, C., Forlizzi, J., & Regli, S. H. (1998). A study of fonts designed for screen display. In *Proceedings of CHI '98* (pp. 87–94). ACM Press. https://dl.acm.org/doi/pdf/10.1145/274644.274658

Britt, M. A., Durik, A., & Rouet, J.-F. (2022). Reading contexts, goals, and decisions: Text comprehension as a situated activity. *Discourse Processes, 59*(5–6), 361–378. https://doi.org/10.1080/0163853X.2022.2068345

Burt, C. (1959). *A psychological study of typography.* Cambridge University Press. https://archive.org/details/psychologicalstu0000sirc/page/n5/mode/2up

Carr, N. (2010). *The shallows: How the internet is changing the way we think, read and remember.* Atlantic Books Ltd.

Cattell, J. (1886). The time taken up by cerebral operations. *Mind, 11,* 524–538.

Dehaene, S. (2009). *Reading in the brain: The new science of how we read.* Penguin Books.

Diemand-Yauman, C., Oppenheimer, D. M., & Vaughan, E. B. (2011). Fortune favors the bold (and the italicized): Effects of disfluency on educational outcomes. *Cognition, 118*(1), 111–115.

Dubay, W. (2004). *The principles of readability.* Impact Information.

Dyson, M. C. (2004). How physical text layout affects reading from screen. *Behaviour & Information Technology, 23*(6), 377–393.

Dyson, M. C. (2013). Where theory meets practice: A critical comparison of research into identifying letters and craft knowledge of type design. *The Design Journal, 16*(3), 271–294.

Dyson, M. C. (2017). Information design research methods. In A. Black, P. Luna, O. Lund, & S. Walker (Eds.), *Information design: Research and practice* (pp. 451–466). Routledge.

Dyson, M. C., & Haselgrove, M. (2001). The influence of reading speed and line length on the effectiveness of reading from screen. *International Journal of Human-Computer Studies, 54*(4), 585–612.

Frascara, J. (2022). Revisiting "Graphic Design: Fine Art or Social Science?"—The question of quality in communication design. *She Ji: The Journal of Design, Economics, and Innovation, 8*(2), 270–288. https://doi.org/10.1016/j.sheji.2022.05.002

Geller, J., Davis, S. D., & Peterson, D. J. (2020). Sans Forgetica is not desirable for learning. *Memory, 28*(8), 957–967. https://doi.org/10.1080/09658211.2020.1797096

Gonzales Crisp, D. (2012). *Typography.* Thames and Hudson.

Grainger, J., Rey, A., & Dufau, S. (2008). Letter perception: From pixels to pandemonium. *Trends in Cognitive Sciences, 12*(10), 381–387.

Hartley, J., & Rooum, D. (1983). Sir Cyril Burt and typography: A re-evaluation. *British Journal of Psychology, 7,* 203–212. https://doi.org/10.1111/j.2044-8295.1983.tb01856.x

Huff, M. J., Maxwell, N. P., & Mitchell, A. (2022). Distinctive Sans Forgetica font does not benefit memory accuracy in the DRM paradigm. *Cognitive Research: Principles and Implications, 7*(1), 102. https://doi.org/10.1186/s41235-022-00448-9

Javal, E. (1878). Hygiène de la lecture. *Bulletin de la Société de Médecine Publique,* 569–575.

Jerkert, J. (2021). On the meaning of medical evidence hierarchies. *Philosophy of Medicine, 2*(1), 1–21. https://doi.org/10.5195/pom.2021.31

Keedy, J. (1993). The rules of typography according to crackpots/experts. *Eye, Winter,* 48–55.

Larson, K. (2004). The science of word recognition. *Advanced Reading Technology, Microsoft Corporation.* https://learn.microsoft.com/en-gb/typography/develop/word-recognition

Larson, K. (2007). The technology of text. *IEEE Spectrum, 44*(5), 26–31.

Larson, K., & Picard, R. (2005). *The aesthetics of reading* [Paper presentation]. Human-Computer Interaction Consortium, Colorado, United States. https://www.media.mit.edu/publications/the-aesthetics-of-reading-2/

Luna, P. (2018). *Typography: A very short introduction.* Oxford University Press.

Lund, O. (1999). *Knowledge construction in typography: The case of legibility research and the legibility of sans serif typefaces* [PhD thesis]. Reading University.

Lupton, E. (2004). The science of typography. *Typotheque*. https://www.typotheque.com/articles/the_science_of_typography

Moys, J.-L. (2017). Visual rhetoric in information design. In A. Black, P. Luna, O. Lund, & S. Walker (Eds.), *Information design: Research and practice* (pp. 205–220). Routledge.

Moys, J.-L. (2013). Investigating readers' impressions of typographic differentiation using repertory grids. *Visible Language, 47*(3), 96–121.

Muijselaar, M. M. L., & de Jong, P. F. (2015). The effects of updating ability and knowledge of reading strategies on reading comprehension. *Learning and Individual Differences, 43,* 111–117. https://doi.org/10.1016/j.lindif.2015.08.011

Noël, G., Frascara, J., & Wong, C. (2019). Designing bowel preparation patient instructions to improve colon cancer detection: Evidence-based design criteria for patients' documents. *Information Design Journal, 25*(1), 110–121. https://doi.org/10.1075/idj.25.1.09noe

Pelli, D. G., Burns, C. W., Farell, B., & Moore-Page, D. C. (2006). Feature detection and letter identification. *Vision Research, 46*(28), 4646–4674. https://doi.org/10.1016/j.visres.2006.04.023

RMIT (2018). Sans Forgetica: New typeface designed to help students study. *RMIT University.* https://www.rmit.edu.au/news/all-news/2018/oct/sans-forgetica-news-story

Sawyer, B. D., Wolfe, B., Dobres, J., Chahine, N., Mehler, B., & Reimer, B. (2020). Glanceable, legible typography over complex backgrounds. *Ergonomics, 63*(7), 864–883. https://doi.org/10.1080/00140139.2020.1758348

Schriver, K. A. (1997). *Dynamics in document design.* John Wiley & Sons.

Song, H., & Schwarz, N. (2008). If it's hard to read, it's hard to do: Processing fluency affects effort prediction and motivation. *Psychological Science, 19*(10), 986–988.

Song, H., & Schwarz, N. (2010). If it's easy to read, it's easy to do, pretty, good, and true. *The Psychologist, 23*(2), 108–111.

Spencer, H. (1969). *The visible word.* Lund Humphries.

Suk, N. (2016). The effects of extensive reading on reading comprehension, reading rate, and vocabulary acquisition. *Reading Research Quarterly, 52*(1), 73–89. https://doi.org/10.1002/rrq.152

Thiessen, M., Beier, S., & Keage, H. (2020). A review of the cognitive effects of disfluent typography on functional reading. *The Design Journal, 23*(5), 797–815. https://doi.org/10.1080/14606925.2020.1810434

Thiessen, M., & Dyson, M. C. (2009). Typography for children with reading difficulties: Preferences for type in reading books. *International Journal of the Book, 6*(2), 115–122.

Thiessen, M., Keage, H., Hwang, I., Astley, J., & Beier, S. (2022). Effect of typeface complexity on automatic whole word reading processes. *Visible Language, 56*(3), 8–31.

Thiessen, M., & Kelly, V. (2019). But, it won an award: A look at communication design 'excellence.' In *The Routledge companion to criticality in art, architecture, and design* (pp. 369–385). Routledge.

Thiessen, M., Kohler, M., Churches, O., Coussens, S., & Keage, H. (2015). Brainy type: A look at how the brain processes typographic information. *Visible Language, 49*(1/2), 175–189.

van der Waarde, K. (1999). Typographic dimensions and conventional wisdom: A discrepancy? *Technical Communication, First Quarter,* 67–74.

Walker, S., & Reynolds, L. (2002/2003). Serifs, sans serifs and infant characters in children's reading books. *Information Design Journal, 11*(2/3), 106–122.

Walker, S. (2001). *Typography & language in everyday life: Prescriptions and practices.* Routledge.

Waller, R. (1991). Typography and discourse. In R. Barr, N. L. Kamil, P. B. Mosenthal & P. D. Pearson (Eds.), *Handbook of reading research, volume II* (pp. 341–380). Longman.

Waller, R. (2012). Graphic literacies for a digital age: The survival of layout. *The Information Society: An International Journal, 28*(4), 236–252.

Wheeler, H. E. (1928). Suggestions for research on the typography of school textbooks. *The Elementary School Journal, 29*(1), 27–31. https://doi.org/10.1086/456172

Wolf, M. (2018). *Reader, come home: The reading brain in a digital world.* HarperCollins.

Wright, P. (1979). The quality control of document design. *Information Design Journal, 1,* 33–42. https://doi.org/10.1075/idj.1.1.05wri

Authors

Karel van der Waarde studied information design in the Netherlands (The Design Academy, Eindhoven) and in the UK (De Montfort University, Leicester, and the University of Reading). In 1995, he started a design–research consultancy in Belgium specializing in testing of information design. Most of the projects are related to information about medicines for patients, doctors and pharmacists. His research focuses on the effects of visual information. Karel van der Waarde frequently publishes and lectures about information design. Van der Waarde is a board member of International Institute for Information Design (IIID, Vienna, Austria) and the International Plain Language Federation (IPLF), and editorial board member of *Information Design Journal, Journal of Visual Communication, The Journal of Visual Political Communication,* and *Visible Language.*

Dr. Myra Thiessen is a researcher in the Design Health Collab at Monash University and is the program coordinator of the Communication Design program in the Faculty of Art, Design, and Architecture. She has rounded expertise as a design practitioner, educator, and researcher with expertise in design for legibility, readability, and usability. Her work focuses on information translation and accessibility, developing communication systems that enable people, including those who may be marginalized due to cognitive differences, to share and use information in healthcare settings and other complex environments. Dr. Thiessen is especially interested in how motivation, context, and environment affect comprehension and decision making, and she specializes in evidence-based design drawing on empirical research methods to test both the preference for and performance of visual materials. As part of the Design Health Collab this interest is applied across a range of user experience design contexts.