Web Table Formatting Affects Readability on Mobile Devices

Christopher Tensmeyer
tensmeye@adobe.com
Adobe Research
College Park, Maryland, USA

Zoya Bylinskii
bylinski@adobe.com
Adobe Research
Cambridge, Massachusetts, USA

Tianyuan Cai
tcai@adobe.com
Adobe Research
San Francisco, California, USA

Dave Miller
dave.miller@tufts.edu
Tufts University
Somerville, Massachusetts, USA

Ani Nenkova
nenkova@adobe.com
Adobe Research
College Park, Maryland, USA

Aleena Gertrudes Niklaus
aleena@adobe.com
Adobe
San Jose, California, USA

Shaun Wallace
shaun_wallace@brown.edu
Brown University
Providence, Rhode Island, USA

ABSTRACT

Reading large tables on small mobile screens presents serious usability challenges that can be addressed, in part, by better table formatting. However, there are few evidenced-based guidelines for formatting mobile tables to improve readability. For this work, we first conducted a survey to investigate how people interact with tables on mobile devices and conducted a study with designers to identify which design considerations are most critical. Based on these findings, we designed and conducted three large scale studies with remote crowdworker participants. Across the studies, we analyze over 14,000 trials from 590 participants who each viewed and answered questions about 28 diverse tables rendered in different formats. We find that smaller cell padding and frozen headers lead to faster task completion, and that while zebra striping and row borders do not speed up tasks, they are still subjectively preferred by participants.

KEYWORDS

mobile device, reading, web tables, quantitative study

ACM Reference Format:
Christopher Tensmeyer, Zoya Bylinskii, Tianyuan Cai, Dave Miller, Ani Nenkova, Aleena Gertrudes Niklaus, and Shaun Wallace. 2023. Web Table Formatting Affects Readability on Mobile Devices. In *Proceedings of the ACM Web Conference 2023 (WWW '23), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3543507.3583506

1 INTRODUCTION

Whether comparing product specs to make a purchasing decision, examining historical data to make financial decisions, or inspecting a menu or nutrition guide to make dietary decisions - in all these cases, we rely on tables to retrieve the information that we need. Tables can display both numerical and textual data, and can be more efficient in communicating numerical data than bar and line charts [22]. However, in looking up tables on small smartphone screens, large tables can overflow, leading to difficulties reading, wasted time [18], additional effort [31], or worse - wrong interpretations of the underlying information [27]. Coupled with the fact that smartphones account for the majority of web traffic [16], this raises the need to examine how formatting affects web data table readability on smartphones so that readers can quickly find the information they need with minimal effort.

Often, the process of manually adapting a large desktop table for viewing on mobile devices requires a skilled designer and domain expertise of the data displayed. For example, some best practices for this conversion include removing less important columns, vertically stacking data from multiple columns, or replacing text with icons [13]. Such manipulations are difficult to automate and manual redesign is cost prohibitive for web sites that have many tables that display data from diverse schema. On the other hand, automatic table detection and conversion to HTML from non-HTML documents such as scanned documents, digital PDFs, and scene images are possible [10], allowing such tables to be reformatted for smartphone display. In such automated workflows, table data schemas are not known a priori, so manual table redesign is not feasible. In all these cases, it is helpful to have guidelines for improving mobile table readability.

Most prior studies on table readability have either used large screens [2, 9, 22] or an older generation of mobile devices [18, 26, 31, 32, 34, 35, 37]. In large screen studies, typically the entire table is displayed and comparative studies between large and small screens find significant differences in task accuracy and speed based on screen size [31]. Table navigation on older mobile devices, such as the Palm Pilots used in previous studies, was discrete and performed by tapping hardware buttons or screen buttons (e.g., via stylus) to scroll to the next table row/col or page of entries in large paginated tables. In contrast, modern touch screen smartphone navigation allows for faster continuous scrolling with intuitive gestures. Thus,

WWW '23, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04…$15.00
https://doi.org/10.1145/3543507.3583506

it is unclear to what extent conclusions from prior studies apply to smartphones. Generalization may also be limited due to most prior studies having few participants, tables, and tasks, and running experiments in controlled lab settings on a single device.

To address these shortcomings, we first conduct a survey (Section 3) to investigate common interactions with tables on mobile devices which also confirms that tables are indeed consumed frequently on mobile devices, and come with significant usability challenges. Based on the survey responses, we designed 28 diverse tables and associated tasks for running large scale quantitative studies. We asked seven designers to design mobile-friendly formats for 14 of these tables (Section 4), and then performed a small preference study to understand which design considerations matter most. As a result of these investigations, we converged on systematically evaluating the effects of text density, cell separators (including rule lines and zebra striping), and frozen headers on table readability. For our main evaluation, we ran three large-scale quantitative studies to investigate how these format factors impact table readability and collected over 14K valid responses from 590 participants. Among our study population of US-based, English-speaking adult crowdworkers, most without reported reading or learning difficulties, we found that smaller spacing between cells consistently lead to faster task completion. This suggests that, for this representative population of mobile readers, the benefit from reducing the amount of scrolling needed to view the table outweighs the increased reading difficulty for denser text observed in previous readability studies [12, 23]. The methods, results, and discussion of our quantitative studies are presented in Sections 5-7 respectively, and we offer our study materials, including all 28 tables with associated tasks in the supplemental material.

2 RELATED WORKS

In this section, we review relevant literature on general text readability on mobile devices and then discuss in more detail the smaller set of studies that focus on table readability on mobile.

Text Readability on Mobile Devices: There is a global societal shift towards reading on digital devices, particularly on mobile devices [25]. However, searching for information on long web pages is more efficient on a large screen compared to a small mobile device screen [3, 17]. Several studies have established that font size, font family, and line spacing can impact reading speed and comprehension on mobile devices. Darroch et. al. [8] found that objective readability did not improve by increasing font size beyond 6pt, though participants subjectively preferred larger fonts. Yamabe and Takahashi evaluated mobile interfaces for walking users and found mixed user preferences for font sizes — those preferring smaller font sizes cited their preference for greater information density, and others preferred larger text for added visual clarity [36]. Smaller fonts have been found to increase reading speed on small screens in Korean, while the reverse is true for larger screens [19]. Moderate line spacing tends to improve readability on mobile devices [15], while line spacing seems to have a smaller effect on larger displays as long as it is not extreme [23]. Recent research builds on Rello et al.'s work to study readability in non-English texts on mobile devices [30] and tablets [5], measuring reading speed, comprehension, and subjective preference. Wang et al. [30] conducted in-lab studies using Chinese texts on mobile devices to study font size, line spacing, paragraph spacing, and page margin. Chatrangsan and Petrie [5] conducted in-lab studies using Thai and English texts on tablets among young and old participants while manipulating font family and font size. Notably, their results show that reading in size 18 point on tablets improved reading speed and comprehension. Other studies have shown the effect of typography on readability is individualized on both large [28, 29] and small screens [14].

Table Readability: Prior work highlights the important of readability with tables, notably how columns improve data organization and cell coloring might improve comprehension [4]. Table 1 gives an overview of prior mobile table readability studies. One study concluded that a larger mobile screen reduced the number of navigation actions for a simple task, but not for a complex task that involved two value lookups with a larger table [31]. Another found that task time slowed substantially for tables wider than the screen when horizontal scrolling was required to find the answer [18].

Motivated by similar findings, other work explored alternative presentation modes for tables. Collapsible rows/cols, Record Mode (view single row/col), and Cell Mode (view local cells with headers) were proposed [26] but not verified with usability studies. Zooming and allowing users to select subtables for display was shown to speed up task time, but displaying tables as text with one cell per line was problematic for complex comparison tasks [32]. Follow up work demonstrated that Cascade Mode [37] – which initially shows a short preview of cell content and allows users to select cells to view the full cell and its row/col – was much faster than having to expand rows/cols to view data. Interactive sorting by column value, allowing users to hide/show individual rows/cols, and frozen headers was verified in a small study to speed up comparison tasks [34]. While conducted on larger screens, Ender's study found that row zebra striping was preferred by participants, but only produced quantitative gains for one of six tasks evaluated [9].

While these are foundational works performed in a lab setting, the generalization of the results may be limited due to small numbers of participants, tables, and tasks. Our study builds on this work by greatly increasing the scale and breath of the tables, tasks, and study population. While we investigate some similar aspects of table formatting as prior work (e.g. frozen headers [34] and zebra striping [9]), we also examine text density and cell borders. Our studies are conducted remotely on participants' own smartphone devices to avoid any effects due to artificial lab conditions.

3 MOBILE TABLE READING HABITS SURVEY

Before conducting our quantitative study on how table formatting impacts reading performance on mobile devices (Sections 5-7), we first performed the preliminary survey described in this section to better understand table reading behaviors and ensure we selected a representative and naturalistic sample of tables for our main study.

3.1 Survey Methods

We created a survey with 12 multiple-choice questions (Table 2), to understand general mobile table reading preferences, ask participants about recently-viewed tables, and collect demographic information about our proposed study population. By asking participants to recall a specific table they recently viewed, we hoped to

Study	Year	Device/size(s)	N	Population	Tables	Type	Factors Investigated
Watters [31]	2003	3x5, 4x6 in	84	University	2	Numeric	Search, Context, Screen Size
Kim [18]	2003	3.1x2.3	28	University	3	Numeric	Table Width, Answer Location
Watters [32]	2005	PDA (Simulated)	18	Unspecified	1	Numeric	Linear View, Overview Mode
Zhang [37]	2006	2.1x2.8 in	10	University	1	Numeric	Row/Col Collapse, Lookahead Cascade
Xu [35]	2009	Desktop, 240x320px	7	University	2	Numeric, Text	Sort, Hide, Screen Size
Xu [34]	2011	Mobile Phone	6	University	2	Numeric, Text	Sort, Hide, Save
Our work	2022	Mobile Phones	590	Crowdworkers	28	Various	Text Density, Cell Borders, Frozen Headers

Table 1: Comparison of our work with previously published user studies on mobile table readability. Our work stands out in number of participants, type of participants, and number/diversity of tables used (see Section 3.3).

#	Paraphrased Question
	Demographics
1	Gender
2	Age
3	Education Level
	Think about a recent table you read on your phone
4	What was the table about?
5	Within how long ago did you read it?
6	How many times had you looked at it before?
7	What did you do with it?†
	Habit Questions
8	What kinds of tables do you usually look at on your phone?†
9	Is reading tables on phones easy or difficult?
10	What do you do if a table is hard to read on your phone?
11	What is difficult about reading tables on your phone?†
12	What would make it easier to look at tables on phones? (Free response)

Table 2: Survey questions to understand table reading habits. † indicates multiple answers allowed.

elicit more precise answers about table viewing behaviors, including viewing frequencies and common tasks.

We recruited 304 participants on Amazon's Mechanical Turk (MTurk) with an approval rate >97% and >1000 previous tasks completed. Participants were compensated $1 USD based on the average completion time of <5 minutes. To ensure quality data, we discarded 150 responses with incorrect answers for our attention check question. Then, we examined the free-response answers and discarded 14 responses that appeared to be automatically generated. For additional validation, we distributed the survey amongst the authors' professional networks and found the collected 41 responses to be consistent with the 140 remaining MTurk responses.

3.2 Survey Results

Q5: Reading tables on phones is a frequent activity, with the majority (87%) of participants recalling a table they read in the last day (34%) or week (53%). **Q6:** Participants often looked at the same table more than once (51%). **Q7:** Overwhelmingly, the most frequent table task is to locate specific information (83%). **Q8:** The four categories *Finance*, *Menu*, *Product Specs*, and *Nutrition* were each frequently viewed (≥ 50%), though *Schedule* was most frequent for non-MTurk participants.

Q9: On average, participants rated the difficulty of reading tables on phones as 2.6 using a 5-Point Likert scale, where 1=*Very Easy* and 5=*Very Hard*. **Q10:** However, when a table is difficult to read on mobile, most participants prefer to switch to a larger screen device (61%) rather than continue with the phone (27%). **Q11:** We acknowledge that the provided answer choices for *What is hard about reading tables on phones?* reflect the authors' own experiences with mobile table readability and may introduce some bias into the responses. However, it is interesting that the top complaints are *Can't see all needed information at the same time* (64%), *Too Small Text Size* (64%), and *Table Too Big for Screen* (54%) involve a tradeoff since smaller text would allow for more content to fit the screen. Full survey results are presented in Table 7 in the Appendix.

Based on these responses, we conclude that reading tables on mobile is a frequent behavior that presents readability challenges. Though overall participants rated the difficulty between *Easy* and *Neutral*, 3 common complaints were selected by over half of respondents. For example, since *Text Too Dense* (27%) was only identified as a challenge by a minority of respondents, reformatting tables with denser text could address the other complaints by allowing more content to fit the screen while allowing a larger font size.

While mobile design tradeoffs have been discussed by prior work, they have not been evaluated in a systematic way, leading to some design guidelines without quantitative evidence. For example, our survey indicates that text is often too small for users, despite some guidelines suggesting that "[u]sing smaller text should actually make tables more readable because it provides enough room for the content to fit well" [13].

3.3 Study Table Selection

One aim of this survey was to identify a diverse and representative set of tables and tasks for our quantitative study, to address the limitations of prior studies on table format readability that typically used one or two tables [9, 31, 32, 35] and/or task types [9, 35].

Based on the reading challenges identified in the survey, we created a simple taxonomy of table size and content type and ensured coverage of these attributes. Possible table heights and widths are *small* (<1 screen), *medium* (1-2 screens), and *large* (>2 screens), where 1 screen is equal to device viewport height and width respectively. To determine the height/width category of a table, we rendered the table on a representative medium sized phone, i.e.

iPhone 11 (5.9 x 2.9in), with no cell margins/padding and size 6vw Helvetica font. For reference, *vw* is a relative CSS unit equal to 1% of the device width, so 6vw on an iPhone 11 is approximately 4.4mm or 12pt. A *small* height table requires no vertical scrolling and a small width table no horizontal scrolling. *Medium* height/width tables are scrollable, but the table center can be a reference point to the reader as it is always in view. Tables bigger than 2 screens lack this reference point and are categorized as *large*.

Table *content type* is based on body cell text and is categorized as: *Numbers; Short Text; Mixed; Long Text.* In *Numbers* tables, all body cells contain either a number, possibly with a unit prefix or suffix, or a date/time composed primarily of digits. *Short Text* tables have body cells with single words or short phrases (≤ 6 words). Tables with longer sentences qualify as *Long Text.* Short and long text tables can include occasional numbers within the text, but a table with a dedicated numeric column would be classified as *Mixed.*

Based on this taxonomy, we selected 14 real tables to cover all individual attributes at least twice and all 9 combinations of height/width. Table topics included finance, statistical, schedule, product specs, definitions lists, menu, movies and tv. In some cases, we extracted a subtable to obtain the desired height/width attribute and for cases where usage rights were unclear, we replaced cell text with content of the same type and length. Then, for each of the 14 tables, we found or created a *paired* table with the same table structure, content type, and similar cell content length. Paired tables allow us to perform A/B testing on formats, while keeping content and task controlled, without having to show a participant the same table more than once (thereby avoiding learning effects and similar biases).

4 TABLE DESIGN AND PREFERENCE STUDIES

The design space for tables is impractically large to study exhaustively, so we first engaged designers to prioritize which aspects of table formatting to focus our quantitative studies on.

4.1 Design Task

We created a simple web GUI tool for editing the following CSS formatting attributes for a given HTML table:

- Font face, size and text alignment for row headers, col headers, and body cells.
- Bold, italic, underline for header text.
- Freezing row and col headers during scrolling (*position:sticky*).
- Row and col header background.
- Cell borders and alternate row shading (zebra stripes).
- Cell padding and minimum height/width.

The tool was designed to be used on a large display with a mobile-sized preview of the formatted table. See Appendix A for additional details on the tool.

We recruited two professional designers from the authors' professional networks and five participants with design experience from usertesting.com to design 14 tables (one per pair) using the tool to be "as easily readable as possible on a mobile phone screen."

4.2 Preference Task

The seven designs produced by all the designers per table differed along many formatting attributes. To understand which attributes were considered most important by readers, we conducted a small blind ranking survey with participants recruited from the author's professional networks. Respondents viewed the seven designs side-by-side for each table (randomized ordering) and selected up to two designs they considered most readable on mobile. An average of 10-15 respondents reviewed the designs for each of the 14 tables. We collected 406 total votes across all the designs, resulting in the following observations:

(1) Row striping/borders were included in nearly every table's top three formats.
(2) Col and row header backgrounds were present in nearly every table's top two formats (those with headers).
(3) Near zero row or col spacing formats almost never top three, even when cells were bordered. However, larger tables tended to have smaller spacings preferred.
(4) Frozen header preference appeared random. Likely respondents infrequently scrolled the tables as instructed.
(5) Non-centered text alignment and non-default font size/style were rarely present in formats or preferred by respondents.

Based on these observations and prior work on table readability, we opted to focus on text density via row/col spacing, cell borders/striping, and freezing headers in our quantitative studies.

5 LARGE SCALE READABILITY STUDIES

This section describes our quantitative studies for measuring the impact of mobile table formatting on readability. Study results are presented in Section 6 and discussed in Section 7.

5.1 Table Task Types and Formats

Based on prior literature (Section 2) and our survey (Section 3), we design four information seeking task types of varying complexity (Table 3) that we distribute across the 26 tables in the main study.

Based our analysis in Section 4, we chose to investigate three table format attributes - text density, cell separators, and freezing header regions when scrolling. All combinations of attributes would be infeasible to test exhaustively, so we adopted a sequential approach. We initialized our experiments with a readable default format from the highly ranked formats in Section 4 (see Appendix B.1) and varied one formatting attribute at a time in a sequence of three consecutive studies. After determining the best value for a given attribute, we fix it, and continue with the next study.

Study 1: Text Density refers to the amount of space between rows and columns achieved by setting both the horizontal and vertical CSS *padding* attributes to be one of four values: 0.75vw, 1.75vw, 3.75vw, 5.5vw (Figure 1). Since we are not controlling for device size, we used the relative CSS unit vw where 100vw = screen width. For reference, at 100dpi and a device width of 2.8in, these spacing values would be approximately 2, 5, 10, 15 pixels.

Non-extreme line spacing has been shown to improve readability of paragraphs [23], and in Section 4 we also found moderate table spacing to be preferred. Our hypothesis was that the moderate 1.75vw would be fastest because larger spacings reduce the amount of visible content. Surprisingly, 0.75vw produced the fastest task

Type	N	Definition	Example
Cell Lookup	6	Given the necessary header cell, find the body cell.	When does train X leave stop Y?
Reverse Cell Lookup	5	Given partial or no header cell information and a body cell value, find the corresponding header cell.	Which train leaves stop Y at time T?
Intra-row/col Comparison	10	Given an identified row or column, compare values within that row or column to find the body cell (or its header) that meets some criteria.	For train X, which 2 consecutive stops have the longest time gap between their departures?
Inter-row/col Query	5	Examine or compare values across multiple rows or columns.	Which is the latest departing train from stop Y1 that arrives at stop Y2 before time T?

Table 3: The four task types used in our large scale table readability studies. In prior work, Cell Lookup and Reverse Cell Lookup are considered simple tasks, while the remainder would be complex [31, 37]. We distribute these task types across the 26 tables (excluding practice tables), with the number of instances of each task type listed under N. The example provided is for a train schedule table where columns correspond to trains, rows to stops, body cells list departure times.

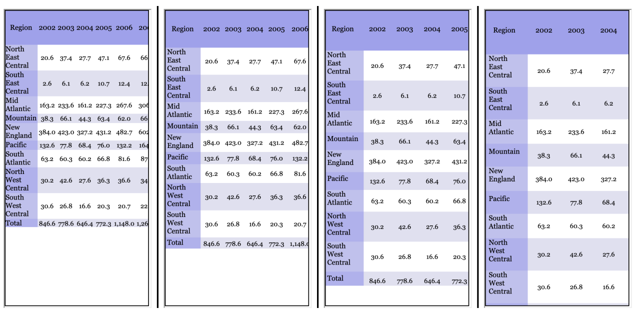

Figure 1: Example rendering of a table in the 4 cell padding values we used in our study. From left to right - 0.75vw, 1.75vw, 3.75vw, 5.5vw.

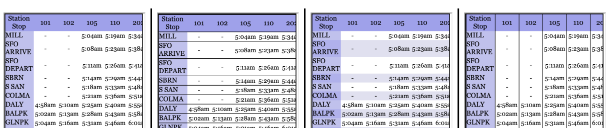

Figure 2: Example rendering of a portion of a table in the 4 separator formats we used in our study. From left to right - plain, row borders, zebra striping, col borders.

times despite being quite tight. We fixed spacing to 0.75vw for the studies that came next.

Study 2: Cell Separators can be implemented as thin rule lines between rows/cols and as alternating row background colors (i.e., zebra stripes). We tested four separator variations: no separators (plain), row zebra striping, row rule lines, and col rule lines (Figure 2). Prior work has found zebra striping to be preferred by readers and to marginally speed up complex table tasks [9], but also to decrease simple search task accuracy [21]. Row or column rule lines are simple alternatives that also visually indicate cell borders. With a spacing of 0.75vw, having no separators produced the fastest task times, so both settings were carried over to the third study.

Study 3: Frozen Headers are when the col and/or row header cells have a fixed position on the screen when the table is scrolled. A common complaint identified in Section 3 is not being able to see all the required information at the same time. Often, headers are needed to interpret body cells, and freezing headers ensures they are always visible. For this study, we compare no freezing, freeze only row headers, freeze only col headers, and freeze both row and col headers. In our implementation, for hierarchical headers, only the inner most header row/col are frozen. If a table does not have row or col headers or is too short or narrow for scrolling, freezing headers is not meaningful. Of the 28 tables, only 14 have col headers and can be vertically scrolled. A (different) set of 14 tables with row headers scroll horizontally. Only 8 tables have both row and col headers and scroll in both directions.

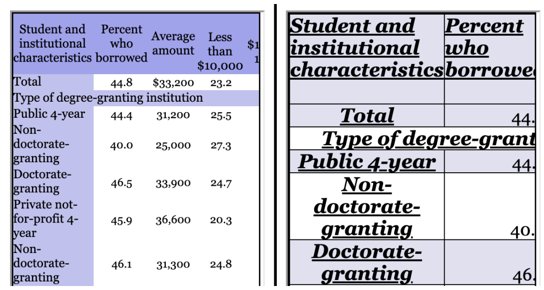

Figure 3: Example rendering of a table in the visual format used by the 4 scrolling variations (left) and the poorly designed format (right) in the frozen header segment.

Since some tables would only have the *no freezing* format, we introduced another table format. So far, tested formats have followed good design principles, so we include a poorly designed format in order to measure the impact of bad formatting. From our preference study, 0-spacing, larger text, header text styles, and non-centered text alignment were rarely part of top ranked formats. In addition, this *bad* format uses all border styles to increase the visual clutter and does not freeze the headers (Figure 3).

5.2 Study Methods

Overview: First, participants completed a short survey with questions on demographics, reading habits/preferences, eyesight, and learning disabilities. Then, we instructed participants to open the study URL on their own mobile devices in a web browser using a QR code. On the landing page, participants are shown the study instructions. Then there are 2 practice trials and 26 main study

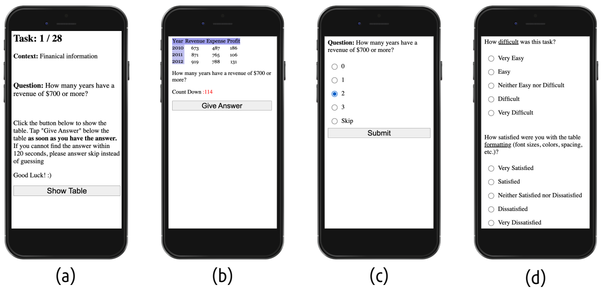

Figure 4: An example table task. (a) task question, task context, and repeated instructions (b) formatted table displayed with countdown timer and question (c) task question shown with answer choices (d) follow up questions.

trials. In each trial, a table is shown in a randomly-selected format out of the options in the given study (Section 5.1), with the exception that paired tables are never assigned the same format. Each table is presented with its corresponding study task (Figure 4). The sequence of tables is shuffled for every participant.

Each study trial is composed of 4 parts (Fig. 4). First, participants see only a context sentence and a task question about a table (4a). When participants tap the *Next* button, the formatted table appears with the question repeated below it along with a 120 second count down timer (4b). The timer encourages participants to find the answer quickly, and when they do they tap the *Give Answer* button to be shown the question with 4 answer choices ordered alphabetically (4c). A fifth *Skip* answer choice is also given to discourage random guessing. Participants received 1 point for a correct answer, 0 points for skipped questions, and -0.33 for wrong answers to discourage guessing. After answering the question, participants are shown 2 follow up questions (4d), which are the same for all trials:

(1) How difficult was the task?
(2) How satisfied were you with the table formatting (font sizes, colors, spacing, etc.)?

Participants pick among the 5-point Likert scale answers for both questions. Once participants answer these 2 questions, we briefly display whether their provided task answer was correct and then advance them to the next trial. After completing all the trials, we ask the following free response questions and show participants their total score.

(1) What made a task easy or difficult?
(2) What made you satisfied/unsatisfied with a table format?
(3) Any other comments about the tasks? Any Issues?

Participants: MTurk was used to distribute our study over a period of 6 weeks with HITs released at various hours throughout the day to help ensure coverage of the eligible MTurk population. We required participants to be located in the US, have 99% HIT approval, and have completed at least 5000 HITs. Participants were aged 19-78 (μ = 39.3, σ=11.2). The vast majority of participants completed the study in 20-30 minutes, so we set the compensation for study completion to be \$8 USD with a \$1 bonus for participants that scored in the top 25% and a \$5 bonus for perfect scores. We cleaned the data by discarding responses with wrong answers or outlier timings according the interquartile range method. Additionally, we excluded all responses from participants who had at least half (13) task responses excluded. See Appendix B.2 for more details.

5.3 Readability Metrics

Readability can be measured by considering many metrics, including glanceability of structure/content, induced cognitive load, legibility, vocabulary, clarity of language, etc. We adopt a task-based interpretation and measure table readability in terms of task completion time, task accuracy, and self-reported task difficulty and format satisfaction.

Comparing or aggregating absolute timings between tables of various size and task complexity is not meaningful, so absolute timings for each table are converted to *z-scores*. With $T_n = \{t_{n1}, \dots, t_{nM}\}$ as the task timings in seconds for M participants on table n, $(1 \le n \le N = 26)$, we can compute the absolute mean and standard deviation timings, μ_n, σ_n, for each T_n. Per-table scaled z-scores are then computed from the absolute timings as

$$Z_n = \{100\frac{t_{n1} - \mu_n}{\sigma_n}, \dots, 100\frac{t_{nM} - \mu_n}{\sigma_n}\} \tag{1}$$

with z_{ni} as the i^{th} entry of Z_n. Positive z_{ni} indicate that participant i took more time (i.e. slower) than average on table n, while negative indicates less time was taken (i.e. faster). To simplify result presentation, we scaled these z-scores by 100 so that $z_{ni} = 100$ indicates that participant i took one standard deviation longer than the mean time on table n.

Similarly, we define D_n, S_n, and F_n respectively as the M integer task difficulty ratings (1 = Very Easy; 5 = Very Difficult), M format satisfaction ratings (1 = Very Satisfied; 5 = Very Unsatisfied), and the M table formats seen by participants.[1] For clarity, $f_{ni} = k$ means that participant i saw table n in the k^{th} format.

We then can compute the average z-score for all tasks seen in format k as

$$mZ(k) = \frac{\sum_{i=1}^{M} \sum_{n \in F(i,k)} z_{ni}}{\sum_{i=1}^{M} |F(i,k)|} \tag{2}$$

where $F(i,k) = \{n | f_{ni} = k\}$ is the set of table indices where participant i saw format k.

While mZ measures average format speed, it is also desirable to directly measure within-participant time differences of format pairs. The average z (i.e. speed) of participants varies widely, and there can be interactions between individual's relative speed and table type or size. For example, a participant may be faster than average on small tables, but slower on large tables. So to measure within-participant differences in z-scores of format pairs we use

$$\Delta Z(k,j) = \frac{\sum_{i=1}^{M} \sum_{n,n' \in F_p(i,k,j)} (z_{ni} - z_{n'i})}{\sum_{i=1}^{M} |F_p(i,k,j)|} \tag{3}$$

where $F_p(i,k,j) = \{(n,n') | f_{ni} = k, f_{n'i} = j, (n,n') \text{ are paired}\}$ is the subset of paired table indices where participant i saw table n in format k and the corresponding paired table n' in format j. Since the numerator of Eq. 3 only includes time differences from the same participant on identically structured tables, it avoids variance due to differing participant speeds. $\Delta Z(k,j) < 0$ indicates that format k

[1] For simplicity, our notation ignores that some participants have incomplete task data.

Format	N Tasks	mZ	Task Accuracy	ES	SS
0.75vw	1620	-10±6	78.8±2.0	58	51
1.75vw	1539	-3±6	80.4±2.0	59	48
3.75vw	1629	-1±6	78.0±2.0	56	47
5.5vw	1570	14±6	78.5±2.0	52	40
plain	1070	1±6	85.1±2.1	52	34
zebra	1029	-2±6	83.0±2.3	55	47
row rules	1097	0±6	84.8±2.1	56	45
col rules	1079	1±6	83.4±2.1	57	40

Table 4: Results for studies 1 (text density) and 2 (cell separators). 95% confidence intervals are given for mZ=mean Z-score time and Task Accuracy. See Section 5.3 for definitions of ES=Easiness Score and SS=Satisfaction Score. Lower mZ is better (faster); higher is better for the other metrics.

	mZ	N	ES	SS	mZ	N	ES	SS
Format	All Tables				Col Freeze Possible			
bad	22±6	1176	49	8	37±10	478	43	4
none	-6±6	1208	63	40	-2±9	508	60	40
col	-	-	-	-	-21±9	486	58	43
	Row Freeze Possible				Both Freeze Possible			
bad	21±11	476	22	-14	35±18	201	-7	-32
none	3±11	500	50	33	13±16	226	27	8
col	-	-	-	-	-8±15	235	23	14
row	-8±9	534	42	33	-8±15	232	26	21
both	-	-	-	-	-26±14	228	44	28

Table 5: Study 3 (frozen headers) results. Row/col headers can only be frozen in some tables, so we compare average results over four (overlapping) sets of tables.

is faster than format j. Appendix B.3 shows an example calculation of this metric for clarity.

To compute the Easiness Score (ES) for a format, we compute the percentage of responses for that format that answered *Very Easy* or *Easy* and subtract the percentage that answered *Very Difficult* or *Difficult*, ignoring the neutral responses. Higher ES is better. We similarly compute the Satisfaction (SS), where a higher score indicates that participants were more satisfied with the table format. Task accuracy is the number of correct responses as a fraction of the total number of responses (with non-outlier timings).

6 RESULTS

We collected 18,927 responses from 733 unique participants completing 26 table tasks (excluding 2 practice tasks). After filtering, we had 14,265 responses from 590 participants, of which 12,255 (81%) responses were correct. Accuracy for individual tasks ranged from 35-99% with a median task accuracy of 88%.

Table 4 shows the metrics from Section 5.3 for studies 1 (text density) and 2 (cell separators) which had 261 and 174 participants respectively. For study 3 (frozen headers) with 155 participants, results are in Table 5, aggregated over the 4 sets of tables where each tested format applies (some tables lack headers or cannot be scrolled). Table 6 contains the pair-wise format metric ΔZ for all pairs of metrics within each study. Note that since data filtering and z-scores normalization is specific to each study, metrics are not directly comparable across studies. Furthermore, after study 1, we edited the context and question text for clarity and replaced a question with extremely low accuracy scores.

Density	0.75vw	1.75vw	3.75vw	5.5vw
0.75vw	-	-14±12	-10±11	-24±12
1.75vw	14±12	-	-6±12	-19±13
3.75vw	10±11	6±12	-	-8±12
5.5vw	24±12	19±13	8±12	-

Separators	plain	zebra	row rules	col rules
plain	-	-1±15	1±15	6±15
zebra	1±15	-	-1±15	3±15
row rules	-1±15	1±15	-	-6±14
col rules	-6±15	-3±15	6±14	-

Freeze	bad	none	col	row	both
bad	-	29±11	54±20	30±19	30±62
none	-29±11	-	13±20	4±18	57±28
col	-54±20	-13±20	-	-34±40	-1±35
row	-30±19	-4±18	34±39	-	5±36
both	-30±62	-57±28	1±35	-5±36	-

Table 6: Pair Results for the 3 studies. ΔZ computed based on eq. 2 to compare the formats in the column to those in the row, where negative table values indicate that the format in the column leads to faster task timings.

Text Density: Table 4 shows a clear trend that smaller spacings are faster and more preferred on average. With the exception of a marginally higher ES for 1.75vw than 0.75vw, every spacing always ranks better than all larger spacings in mZ, SS, ES. Based on the pair results in Table 6, a smaller spacing is faster that all larger spacings. The smallest spacing, 0.75vw, performed fastest with $mZ = -10$, compared to the next fastest $mZ = -3$ for 1.75vw. While these unpaired mZ are not statistically different according to an independent sample t-test at p=0.05, ΔZ for 0.75vw and 1.75vw in Table 6 is significantly different according to a paired sample t-test at p=0.05. We found no significant differences in task accuracy.

That 0.75vw outperforms 1.75vw in speed and SS is surprising given the results of the format preference study, where formats with near-0 spacing were never ranked in the top 3 formats for any of the 14 tables. Possibly participants who have to perform a task with a table in a given format use different subjective criteria than those who only evaluate the aesthetics of the formatting. That larger spacings (3.75vw and 5.5vw vs. 1.75vw) are slower and less preferred in this study is consistent with the results of the preference study.

Separators: While Tables 4 and 6 show no clear benefit in speed or accuracy for any format, the ES/SS scores for the plain format are lower than the other tested formats. This concurs with Enders' [9] findings that while zebra striping is often preferred over a plain table, striping does not consistently speed up tasks on large tables. The difference in zebra striping rows and separating rows with rule lines is minimal w.r.t. ES/SS. However, both ways of visually distinguishing rows are more preferred (SS) than col rule lines, even though all 3 formats have similar task difficulty perception (ES).

Frozen Headers: The mZ scores in Table 5 support our hypothesis that freezing header rows and columns on tables with scrolling allow participants to accomplish their tasks faster. We see a repeating trend that freezing all applicable headers results in faster task completion times than not freezing. Unfortunately, the number of tables where freezing applies is small, and these larger tables tend to be more difficult. Thus, more tasks are excluded due to skipped or incorrect answers, and the confidence intervals for the estimated mZ are very wide with few statistically significant results. Notably, the mZ of freezing col or both headers is significantly lower than freezing no headers on their respective table subsets. This makes sense since many of our tasks have more informative column headers, and our tables tend to be taller than they are wide. This is reflective of tables in the wild, where column header-based tables are 3x more frequent than row header-based tables [7].

7 DISCUSSION

Our results indicate that on average, the smallest cell padding leads to the fastest task completion times. Previous work found that reducing table width speeds up table tasks by reducing the need for horizontal scrolling [18]. However, previous work also found that very tight spacing may reduce readability [23] and reduce within-cell proximity grouping cues [33] important for distinguishing individual cells. If the small spacing actually negatively impacted the visual segmentation of cells, we would have expected this effect to be negated after adding row or column separators to provide additional visual cues. However, when we added row/column separators to the small spacing format, we observed no performance gains beyond increased reported satisfaction and decreased perceived difficulty. Overall, we found that the reduction in scrolling from denser text leads to a net positive effect.

Freezing headers in scrollable tables was shown to generally improve readability. This effect was most pronounced for the large tables that scrolled in both directions but was also important in tables that scrolled in only one direction. Zooming and scrolling can cause users of mobile devices to lose focus and reset their information search [36]. One possible explanation for the effectiveness of freezing header is its ability to help readers maintain context and focus amidst interface changes.

Design guidelines for web tables abound [6, 11, 13, 24], but perhaps due to a lack of quantitative research in this area, such guidelines are often based on subjective aesthetics rather than evidence. Studies such as this one are needed to clarify which design elements are merely stylistic — like cell borders that only increase reader satisfaction, — and which ones objectively improve task completion speed and/or accuracy.

The Web Content Accessibility Guidelines (WCAG) v2.1 [20] do not dictate table formatting and only require that table structure be programmatically available for assistive technologies and that table content is preserved under format changes (criteria 1.4.4; 1.4.12). However, WCAG criterion 1.4.10 permits 2D scrolling in large tables, and based on our findings, we would add to WCAG a recommendation to freeze headers for such tables.

Prior work has made different readability recommendations for younger and older adults [1, 5, 29]. While we did not explicitly recruit participants uniformly across ages, when we compared participants older than 40 to those younger (a similar age split as prior work) [1], we didn't find significant differences in the effects of format attributes on readability. Instead, we found that older participants were slower across all conditions and exhibited greater variability in their task completion times. We also did not find any significant differences among gender groups.

Limitations: We did not exhaustively explore table format attributes or format attribute values, so we do not claim any one format is optimal, only that some formats we studied are better on average than others. In order to control for the same content and content density being shown despite the actual width of the user device, we scaled font size with the user device width. However, we did not investigate possible differences due to device size or variation in viewing distances across participants. We also used a single font face throughout the study similar to prior work [23], though recent work suggests that font face is an important and individualized factor that affects paragraph reading speed and comprehension [28, 29]. To reduce variance due to user familiarity with features, our study design excluded common interactive features that prior work shows to be beneficial: column sorting, filtering, show/hide columns/rows, text search, and viewport zoom in/out. The participant responses have high enough variation that some differences are not statistically significant, particularly if we examine subsets of the tables to draw conclusions about the interactivity between table content/size and format. Since we were not observing participants during the study (e.g. via eye tracking), we cannot determine if smaller spacing affected participants' ability to visually segment rows, columns, or cells. Despite these limitations, our study materials provide opportunities to continue to systematically study any additional formatting attributes or attribute values.

We did not specifically recruit for and received < 10 participants with uncorrected vision impairment, reading or learning disabilities, or English disfluency. and only received a few responses from these populations. Future work is needed to confirm that our recommended formats for the general population do not make reading tables harder for such populations.

8 CONCLUSION

This work presented the first large scale studies of how web table formatting affects readability on modern mobile devices. We found that smaller cell spacing leads to faster task completion on average and was preferred by participants, but that row and column separators increased only participant satisfaction. For large tables that require scrolling to view, frozen headers greatly speed up task completion and increase participant satisfaction. We hope these results can begin to provide evidence-based guidelines for formatting tables for viewing on mobile devices.

Additional work on table readability is needed to examine additional formatting aspects like font size or family and to investigate the impact on specific populations, such as those with reading disabilities like dyslexia. While our study aimed to find the best formatting on average, it is very likely that individuals, both with and without disabilities, differ in which formats are most readable. Future studies would need to collect more responses per individual and format pair to properly test this hypothesis.

REFERENCES

[1] Tianyuan Cai, Shaun Wallace, Tina Rezvanian, Jonathan Dobres, Bernard Kerr, Samuel Berlow, Jeff Huang, Ben D Sawyer, and Zoya Bylinskii. 2022. Personalized Font Recommendations: Combining ML and Typographic Guidelines to Optimize Readability. In *Designing Interactive Systems Conference*. 1–25.

[2] Sophia Caspár. 2018. *Visualization of tabular data on mobile devices*. Ph. D. Dissertation. Luleå University of Technology.

[3] Minhee Chae and Jinwoo Kim. 2004. Do size and structure matter to mobile users? An empirical study of the effects of screen size, information structure, and task complexity on user activities with standard web phones. *Behaviour & information technology* 23, 3 (2004), 165–181.

[4] George Chalhoub and Advait Sarkar. 2022. "It's Freedom to Put Things Where My Mind Wants": Understanding and Improving the User Experience of Structuring Data in Spreadsheets. In *CHI Conference on Human Factors in Computing Systems*. 1–24.

[5] Maneerut Chatrangsan and Helen Petrie. 2019. The effect of typeface and font size on reading text on a tablet computer for older and younger people. In *Proceedings of the 16th Web For All 2019 Personalization-Personalizing the Web*. ACM, ACM, New York, NY, USA, 5.

[6] Andrew Coyle. 2017. Design better data tables. https://medium.com/nextux/design-better-data-tables-4ecc99d23356

[7] Eric Crestan and Patrick Pantel. 2010. A fine-grained taxonomy of tables on the Web. In *Proceedings of the 19th ACM International Conference on Information and Knowledge management*. 1405–1408.

[8] Iain Darroch, Joy Goodman, Stephen Brewster, and Phil Gray. 2005. The effect of age and font size on reading text on handheld computers. In *IFIP conference on human-computer interaction*. Springer, 253–266.

[9] Jessica Enders. 2007. Zebra striping: does it really help?. In *Proceedings of the 19th Australasian conference on Computer-Human Interaction: Entertaining User Interfaces*. 319–322.

[10] Khurram Azeem Hashmi, Marcus Liwicki, Didier Stricker, Muhammad Adnan Afzal, Muhammad Ahtsham Afzal, and Muhammad Zeshan Afzal. 2021. Current status and performance analysis of table recognition in document images with deep neural networks. *IEEE Access* 9 (2021), 87663–87685.

[11] Molly Hellmuth. 2019. The Ultimate Guide to Designing Data Tables. https://medium.com/design-with-figma/the-ultimate-guide-to-designing-data-tables-7db29713a85a

[12] Nafiseh Hojjati and Balakrishnan Muniandy. 2014. The effects of font type and spacing of text for online readability and performance. *Contemporary Educational Technology* 5, 2 (2014), 161–174.

[13] Steven Hoober. 2020. Designing Mobile Tables. https://www.uxmatters.com/mt/archives/2020/07/designing-mobile-tables.php

[14] Ding-Long Huang, Pei-Luen Patrick Rau, and Ying Liu. 2009. Effects of font size, display resolution and task type on reading Chinese fonts from mobile devices. *International Journal of Industrial Ergonomics* 39, 1 (2009), 81–89.

[15] Shih-Miao Huang. 2020. Effects of the Chinese Character Size, Page Spacing and Scrolling Frequency on Reading Time of Smartphones. In *International Conference on Human-Computer Interaction*. Springer, 257–269.

[16] Simon Kemp. 2022. DIGITAL 2022: JULY GLOBAL STATSHOT REPORT. https://datareportal.com/reports/digital-2022-july-global-statshot

[17] Loel Kim and Michael J Albers. 2001. Web design issues when searching for information in a small screen display. In *Proceedings of the 19th annual international conference on Computer documentation*. 193–200.

[18] Loel Kim and Michael J Albers. 2003. Presenting information on the small-screen interface: effects of table formatting. *IEEE Transactions on Professional Communication* 46, 2 (2003), 94–104.

[19] Minjee Kim, Sung Hee Ahn, Youlbeen Kang, Soo-chan Jee, Sondo Kim, Ji Hwan Park, Hunsik Shin, and Myung-Hwan Yun. 2016. The effect of font and display sizes on the readability for mobile devices. In *Proceedings of HCI Korea*. 468–475.

[20] Andrew Kirkpatrick, Joshue O Connor, Alastair Campbell, and Michael Cooper. 2018. Web Content Accessibility Guidelines (WCAG) 2.1. https://www.w3.org/TR/WCAG21/ EdTechHub.ItemAlsoKnownAs: 2405685:8TX6IQZM.

[21] Michael Lee, Travis Kent, C Melody Carswell, Will Seidelman, and Michelle Sublette. 2014. Zebra-striping: visual flow in grid-based graphic design. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 58. SAGE Publications Sage CA: Los Angeles, CA, 1318–1322.

[22] Joachim Meyer, David Shinar, and David Leiser. 1997. Multiple factors that determine performance with tables and graphs. *Human factors* 39, 2 (1997), 268–286.

[23] Luz Rello, Martin Pielot, and Mari-Carmen Marcos. 2016. Make it big! The effect of font size and line spacing on online readability. In *Proceedings of the 2016 CHI conference on Human Factors in Computing Systems*. 3637–3648.

[24] Richard Rutter. 2017. Web Typography: Designing Tables to be Read, Not Looked At. https://alistapart.com/article/web-typography-tables/

[25] Somipam R Shimray, Chennupati Keerti, Chennupati K Ramaiah, et al. 2015. An overview of mobile reading habits. *DESIDOC Journal of Library and Information Technology* 35, 5 (2015), 343–354.

[26] Keishi Tajima and Kaori Ohnishi. 2008. Browsing large HTML tables on small screens. In *Proceedings of the 21st annual ACM symposium on User Interface Software and Technology*. 259–268.

[27] Howard Wainer. 1997. Improving tabular displays, with NAEP tables as examples and inspirations. *Journal of Educational and Behavioral Statistics* 22, 1 (1997), 1–30.

[28] Shaun Wallace, Zoya Bylinskii, Jonathan Dobres, Bernard Kerr, Sam Berlow, Rick Treitman, Nirmal Kumawat, Kathleen Arpin, Dave B Miller, Jeff Huang, et al. 2022. Towards Individuated Reading Experiences: Different Fonts Increase Reading Speed for Different Individuals. *ACM Transactions on Computer-Human Interaction (TOCHI)* 29, 4 (2022), 1–56.

[29] Shaun Wallace, Rick Treitman, Nirmal Kumawat, Kathleen Arpin, Jeff Huang, Ben Sawyer, and Zoya Bylinskii. 2020. Individual differences in font preference & effectiveness as applied to interlude reading in the digital age. *Journal of Vision* 20, 11 (2020), 412–412.

[30] Junxiang Wang, Jianwei Yin, Shuiguang Deng, Ying Li, Calton Pu, Yan Tang, and Zhiling Luo. 2018. Evaluating User Satisfaction with Typography Designs via Mining Touch Interaction Data in Mobile Reading. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, ACM, New York, NY, USA, 113.

[31] Carolyn Watters, Jack Duffy, and Kathryn Duffy. 2003. Using large tables on small display devices. *International Journal of Human-Computer Studies* 58, 1 (2003), 21–37.

[32] Carolyn Watters, Rui Zhang, and Jack Duffy. 2005. Comparing table views for small devices. In *Proceedings of the 2005 ACM symposium on Applied computing*. 975–980.

[33] Max Wertheimer. 1923. Investigations into the theory of gestalt. II. *Psychological research* 4, 1 (1923), 301–350.

[34] Wenchang Xu and Yuanchun Shi. 2011. Enabling efficient browsing and manipulation of web tables on smartphone. In *International Conference on Human-Computer Interaction*. Springer, 117–126.

[35] Wenchang Xu, Xin Yang, and Yuanchun Shi. 2009. A new mode of browsing web tables on small screens. *Proceedings of UCS* (2009), 75–79.

[36] Tetsuo Yamabe and Kiyotaka Takahashi. 2007. Experiments in Mobile User Interface Adaptation for Walking Users. In *The 2007 International Conference on Intelligent Pervasive Computing (IPC 2007)*. IEEE, IEEE, 280–284. https://doi.org/10.1109/ipc.2007.94

[37] Rui Zhang, Carolyn Watters, and Jack Duffy. 2006. Lookahead Cascade for Table Access on Small Devices. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*, Vol. 1. IEEE, 23c–23c.

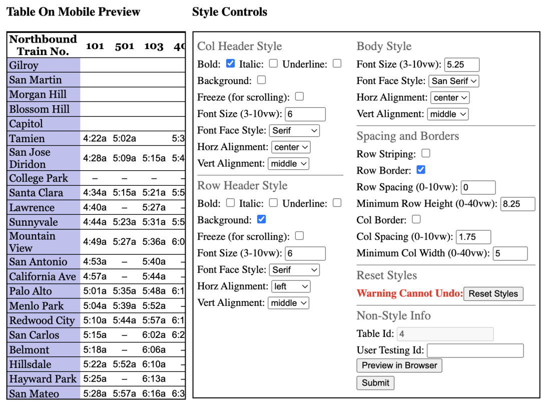

Figure 5: Our design tool for editing table mobile formats. The styled table is shown on the left at a fixed size to mimic a mobile screen size. The checkboxes and drop down boxes in the center allow the designer to edit the table format.

A TABLE FORMATTING TOOL

Figure 5 shows the web GUI tool for formatting mobile HTML tables that was used by the designers in Section 4. The tool was designed to be used on a large display, but a mobile-sized preview of the formatted table is shown on the left. As with a mobile device, if the table is larger than the display area, it is scrollable. A QR code is also provided (off screen) for the user to scan to preview the formatted table on their own mobile device, though this was rarely used by our designers. The controls in the middle allow the designer to modify the table formatting.

B ADDITIONAL METHODS DETAILS

In this appendix, we provide additional details about the methodology of our main quantitative study.

B.1 Default Table Format

The default table format was chosen based on the findings of Section 4. Row header text is left aligned, while col header and body cell text is center aligned. All cell text is vertically middle-aligned and uses the serif font *Georgia*. Row and col header font size is 5vw with body text slightly smaller at 4.5vw. Light purple background colors #C0C0F0 and #A0A0F0 are used to emphasis row and col headers respectively. The colors used for zebra striping were chosen to be lightly contrasting – #FFFFFF (white) and #E0E0F0 – and in our implementation, row header backgrounds on the dark stripes are changed to #B0B0F0. Cell CSS margins are fixed at 0, and the table element is set to *border-collapse*, so the spacing between cells is controlled by the *padding* attribute, which is varied in our first set of experimental formats and then set based on the results of that experiment.

B.2 Data Cleaning

Initial examination of the collected data revealed many outlier times for all tables (both slow and fast). Large outlier timings, including tasks that triggered the 120 second time out, may indicate that the participant was interrupted or unfocused during the task. Small outlier timings could indicate that the participant did not examine the table to answer the question, so either the participant randomly guessed or already knew the answer. Only a few questions could possibly be answered with certainty without examining the table (by bot or human), so we likely had some repeat participants with multiple MTurk accounts. This was also evidenced by a slight shift in participant scores higher as the study progressed, especially in the number of perfect scores.

We followed an iterative process to exclude both individual task responses and all responses by some participants. This filtering of the 26 non-practice tasks is performed independently for the 3 study segments:

(1) Exclude tasks that were skipped or were answered incorrectly.
(2) Use the currently included tasks to compute table-specific first and third quartiles, q_1, q_3, and interquartile range, $r = q_3 - q_1$, task timings.
(3) Exclude tasks with time less than $(q_1 - 1.5r)$ or more than $(q_3 + 1.5r)$.
(4) Exclude all tasks belonging to participants who have at least 13 (50%) excluded tasks
(5) Repeat until no more tasks are excluded.

One reason we iterative on this filtering is that for some tables, there is a large cluster of task times near our enforced minimum time of 3 seconds (since the button to advance past the table viewing screen is only enabled after 3 seconds) that skews $q1$ to be small enough that $q_1 - 1.5r < 3$. For some tasks, 3 seconds may be valid if the task is simple and the answer initially visible. However, for other tables, many of those 3 second timings get excluded by removing entire participants due to those participants having more than 13 outliers or incorrect answers for other tasks. After identifying bad participants, we recompute q_1, q_3, and r since the data distribution has shifted.

B.3 Readability Metric Example

As an example of how to compute our readability metrics Z, mZ, ΔZ, let us consider the example dataset of $M = 10$ participant responses for $N = 4$ tables (2 pairs: $(1, 2)$ and $(3, 4)$) with $K = 3$ formats presented in Table 8.

First, we convert each T_n in Table 8 to scaled z-scores Z_n using μ_n and σ_n according to Eq. 1 to yield the values in Table 9 (rounded to the nearest integer). Note that the mean is 0 and the standard deviation is 100 for each set of table z-scores. This allows us to compare the impact of format on task time across tables that vary in size and task complexity. Then to compute mZ (Eq. 2) of format k, we collect all z-scores from every column in Table 9 that were seen in that format and compute the average z-score. E.g., for $k = 1$, we take all 14 of the bolded values, i.e., -100, -50, -70, 70, -100, -147, 33, 83, -157, -130, 25, 75, 50, 150. Averaging these values together yields mZ(1)=-19. Similarly, mZ(2)=7 and mZ(3)=14 leading to the empirical conclusion that averaged across tasks and participants, it took the least time to complete tasks when viewing format 1.

Participants do not always see each format the same number of times and random sampling procedure could give faster or slower participants a disproportionate allocation of a particular format. In this example, participant 5 is much faster than other participants, but saw format 2 twice and formats 1 and 3 only once. We can

#	Paraphrased Question	Answers
		Demographics
1	Gender	Male (63%); Female (37%); Non-binary; Prefer not to say; Other
2	Age	Free Response Number
3	Education Level	Bachelor Degree (55%); High School/GED (19%); Graduate Degree (18%); Undergrad Student (6%); Grad Student (1%); Other (1%); Prefer not to say;
		Think about a recent table you read on your phone
4	What was the table about?	Nutrition (27%); Finance (26%); Product Specs (18%); Menu (17%); Schedule (6%); Other (6%)
5	Within how long ago did you read it?	1 week (53%); 1 day (34%); 1 month (9%); longer (4%)
6	How many times had you looked at it before?	More than Once (51%); Never (27%); Once (21%);
7	What did you do with it?†	Locate specific info (83%); Compare rows/cols (53%); Copy/paste text (43%); Read entirely (26%); Edit (4%); Other
		Habit Questions
8	What kinds of tables do you usually look at on your phone?†	Finance (60%); Menu (56%); Product Specs (56%); Nutrition (50%); Schedule (31%); Other (4%)
9	Reading tables on phones is generally	Very Easy (11%); Easy (42%); Neither (18%); Hard (27%); Very Hard (2%)
10	What do you do if a table is hard to read on your phone?	View on a larger device (61%); Use phone anyway (27%); Give up (6%); Print the table (4%); Other (2%)
11	What do you find hard about reading tables on your phone?†	Can't see all needed info at the same time (64%); Text too small (64%); Table too big for screen (54%); Difficulty seeing row alignment (31%); Text too close together (27%);Difficulty seeing col alignment (20%); Other (2%)
12	What would make it easier to look at tables on phones?	Free Response

Table 7: Survey questions and responses to understand table reading habits. Multiple choice answers are separated by ";" and listed in order of response frequency. Answer choice "Other" allows free response. † - Select all answers that apply.

Participant i	T_1	T_2	T_3	T_4
1	40	70	53	75
2	**20**	50	**105**	50
3	**25**	37	85	**75**
4	35	**25**	116	**85**
5	15	35	**33**	40
6	45	**18**	96	**80**
7	23	**45**	102	80
8	37	53	63	75
9	**23**	32	106	**100**
10	**37**	35	**41**	40
avg μ_n	30	40	80	70
std σ_n	10	15	30	20

Table 8: Example dataset of participant task times in seconds and formats seen. Format 1 is indicated by bold. Format 2 by underline. Format 3 with default font.

remove these effects by using the metric ΔZ (Eq. 3) to compare formats by examining within-participant and within-table-pair time differences.

For $\Delta Z(1,2)$, we first identify which participants saw table pairs (1,2) or (3,4) in formats 1 and 2. For table pair (1,2), these are participants 2, 6, 7 and for table pair (3,4) these are participants 2, 4, 5, 6, 9. Then, for these participants and table pairs, we subtract the z-scores for the table seen in format 1 from the z-scores of the table seen in format 2 and take the average difference:

$$\begin{aligned}\Delta Z(1,2) = \frac{1}{8}\Big[&(-100-67)+(-147-150)+(33-(-70))\\ &+(83-(-100))+(75-120)+(-157-(-150))\\ &+(50-53)+(150-87)\Big]\\ =\ &-21\end{aligned} \tag{4}$$

Note that $\Delta Z(a,b) = -\Delta Z(b,a)$, so $\Delta Z(2,1) = 21$. Similarly we have $\Delta Z(1,3) = 17$ and $\Delta Z(2,3) = -70$.

Participant i	T_1	T_2	T_3	T_4
1	100	200	-90	25
2	**-100**	67	**83**	-100
3	**-50**	-20	17	**25**
4	50	**-100**	120	**75**
5	-150	-33	**-157**	-150
6	150	**-147**	53	**50**
7	-70	**33**	73	50
8	70	87	-57	25
9	**-70**	-53	87	**150**
10	**70**	-33	**-130**	-150
avg μ_n	0	0	0	0
std σ_n	100	100	100	100

Table 9: Conversion of raw task times from Table 8 to Z-scores using Eq. 1. Format 1 is indicated by bold. Format 2 by underline. Format 3 with default font.