



Discussion paper

Experiment Smartphone-First Questionnaire Layout

Deirdre Giesen, Maaïke
Kompier & Jan van den
Brakel

oktober 2025

Summary

This paper presents the purpose, design, and results of a large-scale field experiment investigating the effects of a smartphone-first redesign of the login process and questionnaire layout for CBS household surveys. This study investigates the effects of six experimental conditions:

1. A smartphone-first developed login screen versus the regular login screen.
2. A smartphone-first developed questionnaire layout versus the regular questionnaire layout.
3. Four designs for grids: three smartphone-first developed options for grid design (stem fix, carousel and accordion) versus the regular design (classic table for large screen and stem fix for small screen).
4. The use of smileys and icons in answer options: text only, text and smileys + icons and smileys only + icons.
5. Showing an instruction on smartphones to encourage the use of speech to text for open questions versus not showing such an instruction.
6. A shorter (17 minutes) versus a longer version (25 minutes) of the questionnaire.

Across a broad set of response behavior indicators (e.g., logging in, breaking off, choosing 'no answer', straightlining, mean scores on various scales, and consistency with register data for several items), as well as indicators of respondent satisfaction, the experiment revealed remarkably few significant effects. This is reassuring, as it suggests that respondents provide answers largely independent of the tested design alternatives, reducing concerns about disruptions in time series or systematic measurement bias when new designs are introduced. At the same time, the experiment did not succeed in alleviating login difficulties or reducing the relatively high break-off rates on smartphones.

In a separate project, the login process was simplified drawing on a literature review and user feedback. Changes such as lowercase-only passwords, caps lock notifications, and clearer error messages were implemented in January 2025. Since then, the number of login errors has shown a significant decline.

The smartphone-first design will be gradually implemented in CBS questionnaires, as it aligns with the corporate style guide, improves accessibility, and ensures a consistent experience across devices.

Keywords

Login design; Questionnaire design; Grid design; Blaise 5; Break-off; Response behavior; Respondent satisfaction

Content

1. Introduction	4
2. Research questions	7
2.1 Experimental conditions	7
2.2 Hypotheses	18
3. Methodology	20
3.1 Sample design and experimental design	20
3.2 Fieldwork	26
3.3 Independent variables and control variables	26
3.4 Dependent variables	27
3.5 Analytical approach for main analysis	29
3.6 Additional analysis login screen	30
3.7 Additional analysis speech-to-text encouragement	31
4. Results	33
4.1 Response	33
4.2 Device use and circumstances of survey completion	34
4.3 Dependent variables	35
4.4 Overview results analyses	36
4.5 The login screen	38
4.6 Questionnaire layout	41
4.7 Grid design	42
4.8 Smileys and icons	44
4.9 Speech-to-text encouragement	46
4.10 Short and longer version of the questionnaire	48
4.11 Device effects	50
5. Discussion	52
Acknowledgments	54
References	55

1. Introduction

Online household surveys have increasingly become mixed-device surveys. Currently, the proportion of questionnaire logins on smartphones for CBS (Statistics Netherlands) household surveys is slightly higher than the proportion of logins on PCs (see Figure 1).

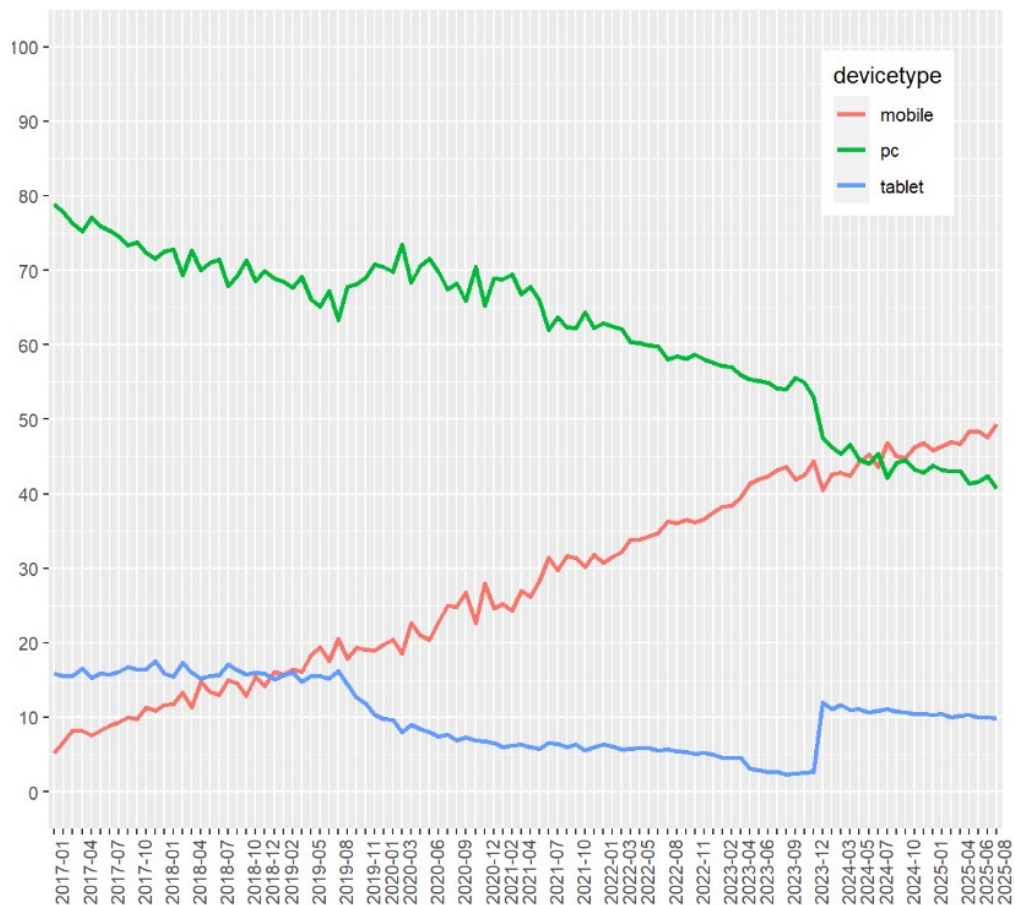


Figure 1 Device use on first login for CBS continuous household surveys January 2017-August 2025. Note: as of January 2024, a better methodology is used to detect tablets that were before misclassified as PCs

As shown in Figure 2, the proportion of break-offs on smartphones has dropped sharply—from about 37% to roughly 8%—between 2017 and 2021. This decline coincides with the gradual transition at CBS from Blaise 4 to Blaise 5, the software used for questionnaire development. With Blaise 5, a responsive design and a new, more smartphone-friendly layout were introduced, including the use of buttons and the elimination of horizontal scrolling. However, the break-off rate on smartphones remains higher than on PCs. See Roberts & Bakker (2018) for a more detailed description and analysis of device specific logins and break-offs in CBS surveys.

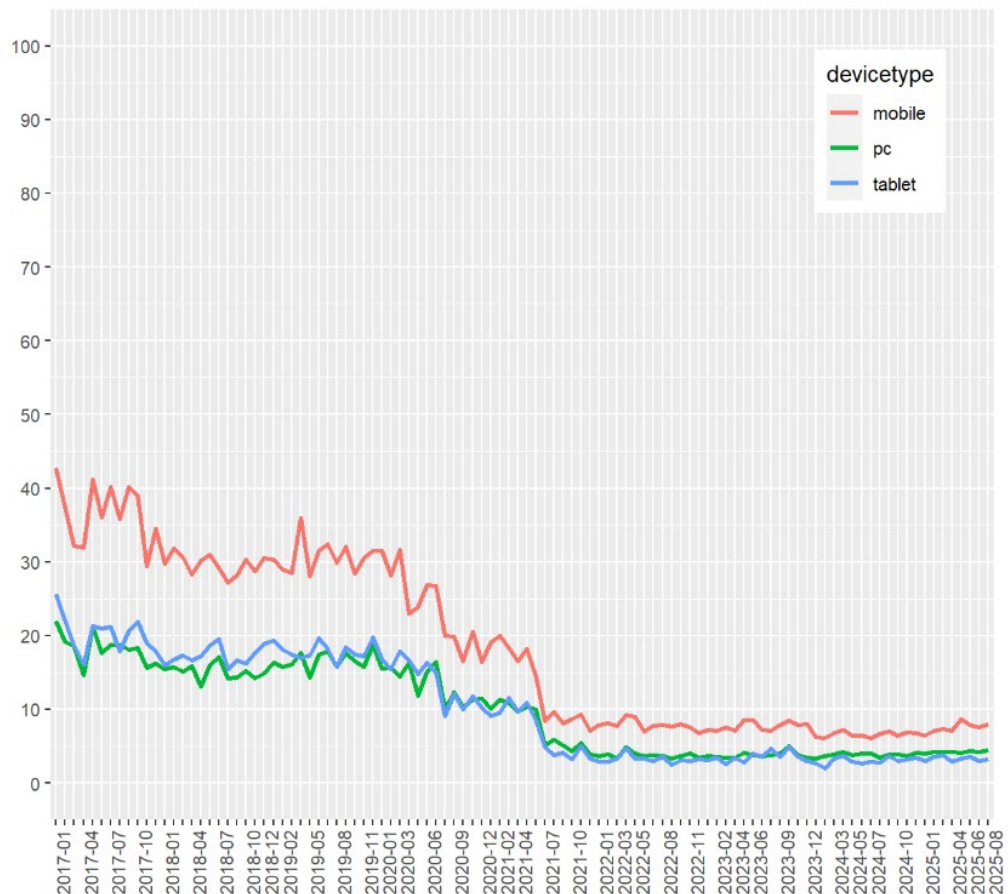


Figure 2 Break-off by device on last login for CBS continuous household surveys January 2017-August 2025. Note: as of January 2024, a better methodology is used to detect tablets that were before misclassified as PCs

Userlab tests and a literature review (Giesen, Bakker, Meertens, & Luiten, 2019; Schouten, Blanke, Gravem, Luiten, Meertens and Paulus, 2018) indicated that the improved, smartphone-friendly, layout (hereafter: the regular layout) still had some usability issues. Issues identified included the small space between buttons that made it difficult to select the correct button on a touch screen, the readability of instruction texts in blue italic and difficulty to select the correct answer in classic grid tables on PCs. In Userlab tests it was also observed that respondents often struggle with correctly typing in the login information, especially on a phone. Concerns about usability and device-related effects prompted research into a new questionnaire design.

For this new design we used a smartphone-first approach, in which a questionnaire and the login screen were (re)designed with the smartphone as the starting point. This design is then applied across all devices and modes. It also seemed worthwhile to explore how questionnaire innovations—such as voice input, the use of emojis or smileys, and alternative grid formats—can enhance both the user experience and data quality.

In addition to the developments mentioned above, there were two other key reasons for developing a new layout: aligning with the CBS corporate style guide and meeting accessibility requirements.

To investigate the effects of a smartphone-first approach on the quality of survey response, a large-scale field experiment is conducted. This paper describes the purpose, design and results of this experiment. In Section 2 the research questions of this experiment are explained. Section 3 describes the experimental design and the analysis methods used for this study. Section 4 describes the results. The paper concludes with a discussion in Section 5.

2. Research questions

The main research questions for this study are:

1. What is the effect of a smartphone-first redesign of the login screen on response rates and break-off?
2. How does a smartphone-first questionnaire layout affect data quality and respondent perception? Does this vary between design alternatives?
3. Do device specific measurement effects occur? Does this vary between design alternatives?

2.1 Experimental conditions

In this study, the following conditions were compared:

1. Design login screen:
 - a. Regular login screen
 - b. Experimental login screen
2. Design questionnaire layout
 - a. Regular questionnaire layout
 - b. Experimental questionnaire layout
3. Grid design
 - a. Regular design, i.e. classic table for large screen and stem fix for small screen
 - b. Stem fix for large screen
 - c. Carousel
 - d. Accordion
4. Graphical symbols
 - a. No smileys, no icons for question about social media (regular approach)
 - b. Smileys with text and icons
 - c. Smileys without text and icons
5. Encouraging the use of speech to text for open questions
 - a. No encouragement (regular approach)
 - b. Encouragement
6. Length questionnaire
 - a. Short questionnaire (17 minutes)
 - b. Long questionnaire (25 minutes)

2.1.1 The login screen

CBS approaches sample units for online surveys with a letter with a URL, a username and a password. See Figure 3 for an example of how the login information is presented in a regular letter and Figure 4 for the regular login screen as it was used at the time of the experiment.

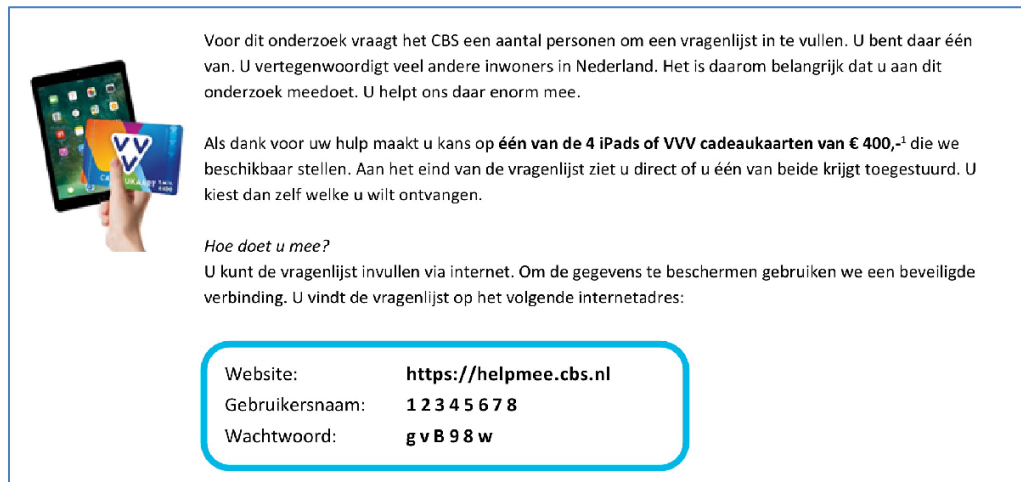



Figure 3 Excerpt from example letter for the regular login screen.

Inloggen

Gebruikersnaam

Wachtwoord 


Inloggen

Figure 4 Example regular login screen. Note: the "eye icon" to show the password is not shown on smartphones and only shown on PCs if characters are typed in the password field.

An experimental version of the login letter and screen were developed with the aim of making the login process easier (see Figure 5 and Figure 6).

The main changes were:

- Adding a short welcome text explaining that they need to enter the username and password from the letter and a link to a website with help on logging in.
- Segmenting the username and password (both in the login screen and in the letter) to make it easier to read and copy this information.
- Adding an (always visible) "show password" button to make the typed in password visible.
- Using the [CBS design system](#), this is a set of standardized guidelines and components—such as fonts, colors, and button styles—that ensure a consistent visual identity and user experience across all products and communications.
- Allowing a maximum of eight characters in the username (whereas in the regular version any number of characters could be entered), clarifying the expected input and providing immediate feedback on an entry of an invalid number of characters.



Voor dit onderzoek vraagt het CBS een aantal personen om een vragenlijst in te vullen. U bent daar één van. U vertegenwoordigt veel andere inwoners in Nederland. Het is daarom belangrijk dat u aan dit onderzoek meedoet. U helpt ons daar enorm mee.

Als dank voor uw hulp maakt u kans op **één van de 4 iPads of VVV cadeaukaarten van € 400,-¹** die we beschikbaar stellen. Aan het eind van de vragenlijst ziet u direct of u één van beide krijgt toegestuurd. U kiest dan zelf welke u wilt ontvangen.

Hoe doet u mee?
U kunt de vragenlijst invullen via internet. Om de gegevens te beschermen gebruiken we een beveiligde verbinding. U vindt de vragenlijst op het volgende internetadres:

Website:	https://doemee.cbs.nl
Gebruikersnaam:	1 2 3 4 - 5 6 7 8
Wachtwoord:	g v - B 9 - 8 w

Figure 5 Excerpt from letter for the experimental login screen.

Welkom bij dit onderzoek van het Centraal Bureau voor de Statistiek

Voer uw gebruikersnaam en wachtwoord in om te beginnen met het onderzoek. Deze staan in de brief die u heeft ontvangen. Hulp nodig? [Neem contact met ons op.](#)

Gebruikersnaam -

Wachtwoord - -

[Toon wachtwoord](#)

Figure 6 Experimental login screen (auto forward between the dashes).

For the experiment, two URLs were used that were similarly easy to understand and type: <https://helpmee.cbs.nl> for the regular design, <https://doemee.cbs.nl> for the experimental design. 'Help mee' can be translated as 'contribute' and 'doe mee' as 'take part'. Note that the normal URL used for logging into CBS questionnaires is <https://antwoord.cbs.nl> (with the regular design) was also available during the experiment for data collection for production surveys. Respondents who googled how to login to CBS questionnaires would probably find this URL and could also login through this URL.

2.1.2 Questionnaire layout

The experimental, smartphone-first design was designed according to three main principles: a minimalistic design, ease of selection and similarity over devices (Van Hees, 2019, based on e.g. Antoun, Couper, & Conrad, 2017; Schouten, Blanke, Gravem, Luiten, Meertens and Paulus, 2018). This design was furthermore aligned with the general CBS design system.

Figure 7 and Figure 8 show some examples of the regular and experimental design, both on large and small screens. The main changes made in the experimental design compared to the regular design are:

- Minimalistic design: less color and elements, for example the questionnaire name is not displayed on top of each screen (only on the welcome screen).
- Single choice answer options are only displayed with large buttons both on the large and small screen (in the regular design radio buttons were shown on the large screen).¹
- Larger buttons and more space between buttons.
- Much more space between the navigation buttons.
- Font and colors according to CBS design system, with instruction text no longer in italics, as italicized text is harder to read.
- Extra vertical space between substantial answer options and the "no answer" option; This was not a smartphone related change but implemented to better communicate that this answer option is different and to make it easier for answer scales to determine the middle point.

Figure 7 displays four examples (A, B, C, D) of questionnaire layouts comparing regular and experimental designs on large and small screens.

A (Large screen, regular design): Shows a questionnaire titled "Nederland Vandaag". The first question is "Zijn u en uw partner getrouwd of heeft u een geregistreerd partnerschap?" with radio button options: "Ja, getrouwd", "Ja, geregistreerd partnerschap", and "Nee". The second question is "Wat is het geslacht van uw partner?" with radio button options: "Man", "Vrouw", and "Overig". The third question is "Wat is de geboortedatum van uw partner?" with a date field (DD-MM-JJJJ) and a "Geen antwoord" option. Navigation buttons "Vorige" and "Volgende" are at the bottom.

B (Small screen, regular design): Shows the same questionnaire on a small screen. The date field is formatted as "dd-mm-jjjj".

C (Large screen, experimental design): Shows the same questionnaire on a large screen. The answer options are displayed as large buttons instead of radio buttons. The date field is formatted as "dd-mm-jjjj".

D (Small screen, experimental design): Shows the same questionnaire on a small screen. The answer options are displayed as large buttons instead of radio buttons. The date field is formatted as "dd-mm-jjjj".

Figure 7 Example single choice question and date field in regular layout on large (A) and small screen (B) and in experimental layout large (C) and small screen (D)

¹ Please note that this option was later identified as not conforming to accessibility standards.

A

Nederland Vandaag

De volgende vragen gaan over nieuws, internetgebruik en sociale media.

Volgt u weleens het nieuws?
U kunt meerdere antwoorden kiezen.

- ☐ Ja, via de krant (papier of digitaal)
- ☐ Ja, via de radio
- ☐ Ja, via de TV
- ☐ Ja, via nieuwswebsites, blogs of apps
- ☐ Ja, via sociale media
- ☐ Ja, via podcasts
- ☐ Nee, ik volg geen nieuws
- ☐ Geen antwoord

Vorige Volgende

B

prepenquete2.cbs.nl

Nederland Vandaag

De volgende vragen gaan over nieuws, internetgebruik en sociale media.

Volgt u weleens het nieuws?
U kunt meerdere antwoorden kiezen.

- ☒ Ja, via de krant (papier of digitaal)
- ☒ Ja, via de radio
- ☒ Ja, via de TV
- ☐ Ja, via nieuwswebsites, blogs of apps
- ☐ Ja, via sociale media
- ☐ Ja, via podcasts
- ☐ Nee, ik volg geen nieuws
- ☐ Geen antwoord

Vorige Volgende

C

De volgende vragen gaan over nieuws, internetgebruik en sociale media.

Volgt u weleens het nieuws?
U kunt meerdere antwoorden kiezen.

- ☐ Ja, via de krant (papier of digitaal)
- ☐ Ja, via de radio
- ☐ Ja, via de TV
- ☐ Ja, via nieuwswebsites, blogs of apps
- ☐ Ja, via sociale media
- ☐ Ja, via podcasts
- ☐ Nee, ik volg geen nieuws
- ☐ Geen antwoord

Vorige Volgende

D

prepenquete2.cbs.nl

De volgende vragen gaan over nieuws, internetgebruik en sociale media.

Volgt u weleens het nieuws?
U kunt meerdere antwoorden kiezen.

- ☒ Ja, via de krant (papier of digitaal)
- ☒ Ja, via de radio
- ☐ Ja, via de TV
- ☐ Ja, via nieuwswebsites, blogs of apps
- ☐ Ja, via sociale media
- ☐ Ja, via podcasts
- ☐ Nee, ik volg geen nieuws
- ☐ Geen antwoord

Figure 8 Multiple select question in regular layout on large (A) and small screen (B) and in experimental layout large (C) and small screen (D)

2.1.3 Grid design

In the regular questionnaire design, grid questions are displayed differently on a large screen than on a small screen (Figure 9). For the large screen, grid questions are shown as a classic table. The question stem is shown in the top of the table, each row represents one of the question items and the labels of the answer options are displayed as headers of the columns.

For a small screen (smartphones), a so called “stem fix” design is used. The question stem is shown fixed on the top of the screen. This is done as the stem may include important information, such as a reference period, which should be easily accessible to respondents. The items and answer options are displayed vertically below the stem. Please note that as in all CBS questionnaires, at the bottom of the screen a “next” button must be selected to proceed to the next question. This ensures respondents always scroll down to the bottom of the screen.

Figure 9 consists of two screenshots, A and B, illustrating different grid question formats.

Screenshot A (Large screen): Shows a classic table format. The title is "Nederland Vandaag". The question stem is "Als u het nieuws volgt, in hoeverre bent u dan geïnteresseerd in:". The table has three rows of questions: "Binnenlands nieuws?", "Buitenlands nieuws?", and "Nieuws over politiek?". The columns represent levels of interest: "Heel erg geïnteresseerd", "Erg geïnteresseerd", "Redelijk geïnteresseerd", "Een beetje geïnteresseerd", "Niet geïnteresseerd", and "Geen antwoord". Each cell contains a radio button. Navigation buttons "Vorige" and "Volgende" are at the bottom.

	Heel erg geïnteresseerd	Erg geïnteresseerd	Redelijk geïnteresseerd	Een beetje geïnteresseerd	Niet geïnteresseerd	Geen antwoord
Binnenlands nieuws?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Buitenlands nieuws?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Nieuws over politiek?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Screenshot B (Small screen): Shows a stem fix design. The title is "Nederland Vandaag". The question stem is "Als u het nieuws volgt, in hoeverre bent u dan geïnteresseerd in:". The answer options are listed vertically: "Heel erg geïnteresseerd", "Erg geïnteresseerd", "Redelijk geïnteresseerd", "Een beetje geïnteresseerd", "Niet geïnteresseerd", and "Geen antwoord". Below this, the question "Nieuws over politiek?" is shown with the same set of answer options. Navigation buttons "Vorige" and "Volgende" are at the bottom.

Figure 9 Regular grid design: classic table on a large screen (A) and stem fix design on a small screen (B).

The difference in presentation of questions between small and large screens might affect how respondents perceive and answer the questions. This may generate device specific measurement error (e.g. Couper, Antoun & Mavletova, 2017; Menold & Toepoel, 2022). Additionally, we had observed in Userlab tests that for some respondents the classic table format was quite challenging. To reduce the risk of device specific measurement error and to make grid questions easier, a uniform format of questions for all screen sizes was developed. Three alternative experimental uniform grid designs were tested.

The first experimental grid design uses the stem fix smartphone layout also for large screens. The stem of the question is fixed on top of the screen. The items and answer options scroll underneath. This new stem fix design was developed both for the regular questionnaire design and for the experimental design (see Figure 10).

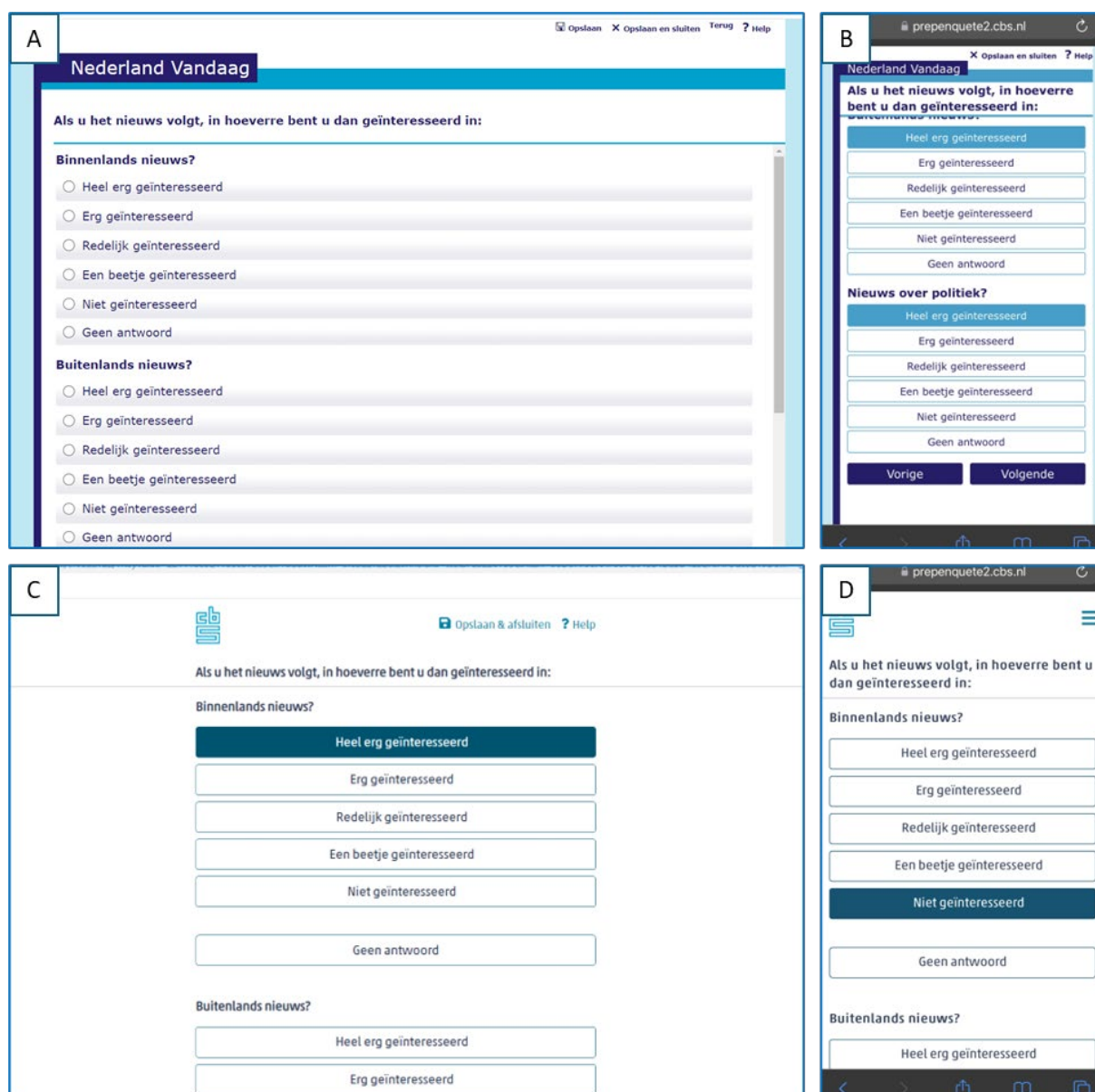


Figure 10 Stem fix for both large (A) and small screen (B) in regular layout and for both large (C) and small screen (D) in experimental layout.

The second design was the carrousel question type (see Figure 11). This design was only developed within the experimental layout design. In the carrousel, the question stem is located at the top. Autoforward was implemented: the item and answer options slide in automatically after an answer option has been selected. Respondents could also navigate manually forward or backward to answered items by clicking on the numbers at the bottom².

² Small scale user tests indicated that the difference in functionality of the numbers at the bottom and the previous button was unclear.

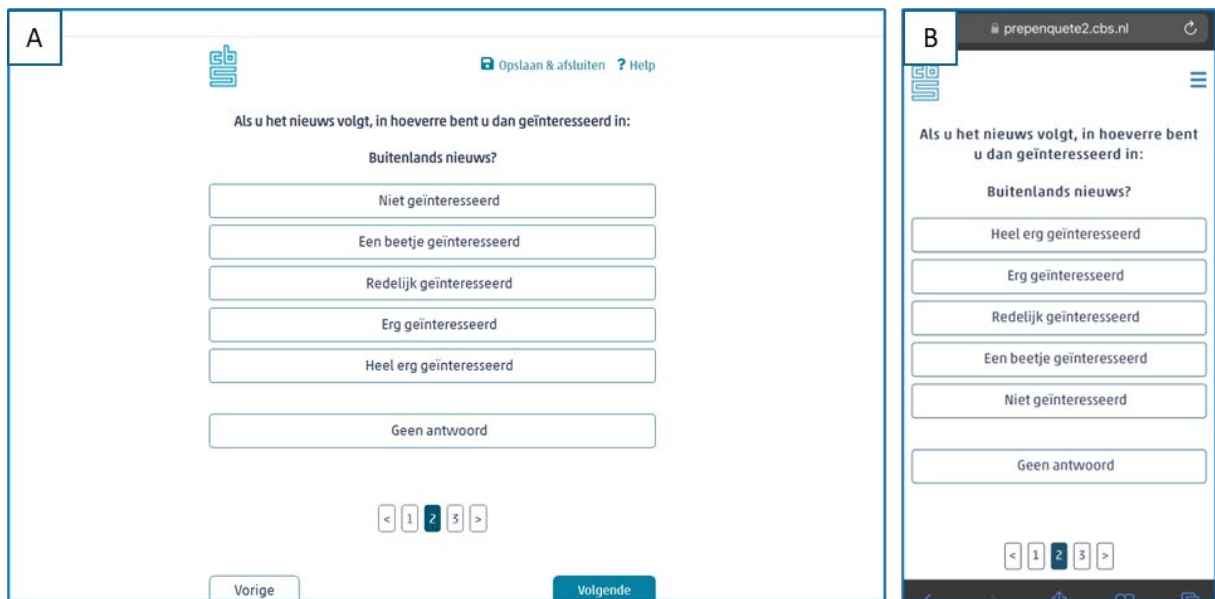


Figure 11 Carousel design (in experimental layout only) on a large screen (A) on a small screen (B).

The third experimental grid design was the accordion design. This grid was also only developed for the experimental layout. Here the question stem and items are always shown, but only the answer options of the active field are shown (see Figure 12). After selecting an answer, this answer is displayed directly below the item and the answer options of the next item are displayed. At the last item, the respondent was forwarded automatically to the next question after selecting the answer option³.

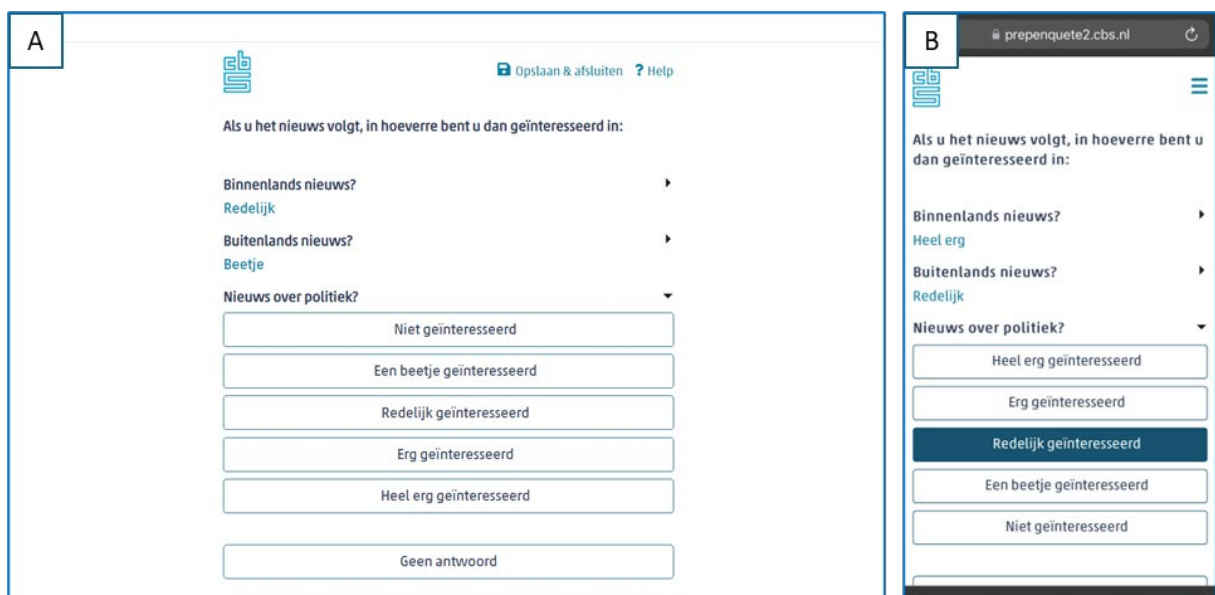


Figure 12 Accordion design (in experimental layout only) on a large screen (A) on a small screen (B).

³ Small scale user tests indicated that respondents preferred to stay on the page and press next themselves (in line with the logic of the rest of the questionnaire).

Note that for both the carousel and the accordion designs, the question stem was not fixed. Consequently, in questions with many items and/or answer options, the question stem may have become invisible. Additionally, for both the classic grid and the accordion, respondents can more easily view the answers provided to previous items compared with the stem fixe and carousel designs. For items that may be evaluated relative to one another, this visibility difference could influence the response process.

The questionnaire was designed to contain a variety of grid questions that are common in CBS questionnaires. Depending on the version of the questionnaire (short or long) and routing, respondents received at least six and a maximum of twelve grid questions. The number of items per grid varied between three and eleven. The number of substantial answer options used in the grids varied from 2 (yes/no) to 10 (rating of trust in various institutions). Several of the grids contained items that may be evaluated relative to one each other (such as the trust in various institutions). See Appendix A for details on the grid questions used.

2.1.4 Smileys and icons

Visual cues such as smileys and icons may help respondents to process information easier. Toepoel, Vermeeren & Metin (2019) found that the use of smileys in answer options was evaluated positively by respondents, while producing comparable average answer scores as traditional radio buttons.

Especially on small screens, where reading text is more burdensome, smileys and icons may be helpful. The use of smileys and icons was only tested within the experimental layout.

Respondents with the experimental layout condition received one of three alternatives for the display of the answer options for all questions that asked about satisfaction: text only, text with a smiley and smileys only (see Figure 13). Respondents who received the questionnaire in the regular layout only had a text only option.

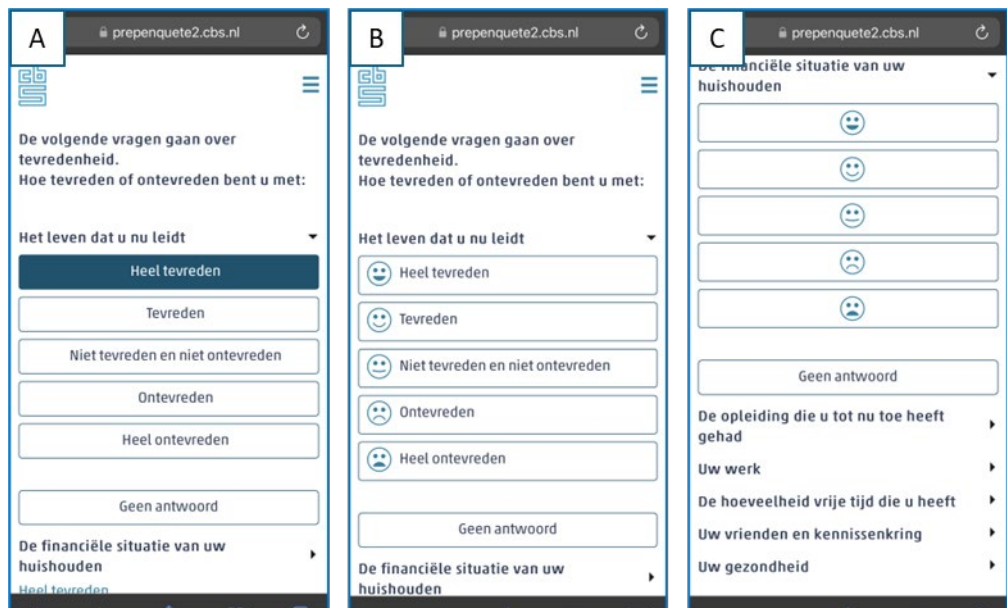


Figure 13 Smiley condition: text only (A), text and smileys (B) and smileys only (C) (experimental layout only).

Respondents in the experimental layout who were exposed to either the smileys with text or the smileys only condition, also received icons with the answer options for a question on social media use. The respondents in the text only condition received the social media questions with only text in the answer options (see Figure 14), as did the respondents in the regular layout condition.

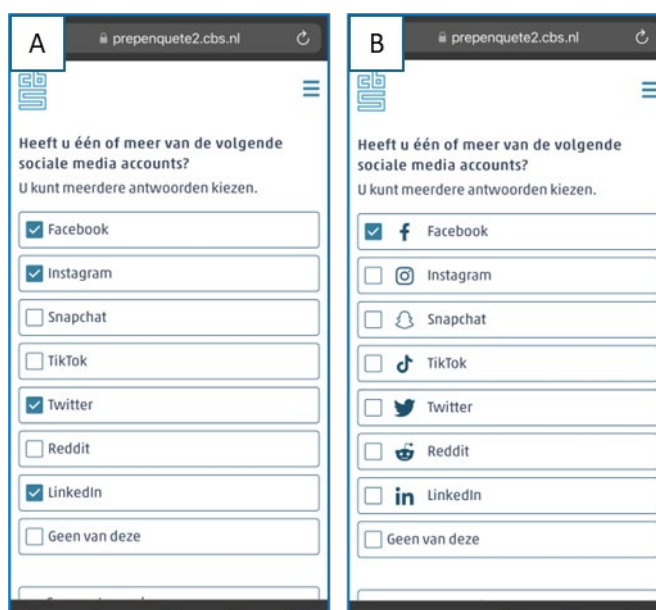


Figure 14 Question on social media accounts with text only (A) or text and icons (B).

2.1.5 Speech-to-text encouragement

Typing in answers to open questions can be burdensome, especially on smartphones. Several studies found that respondents on smartphones provide shorter answers than PC respondents (e.g. Revilla & Ochoa, 2016), whereas others found no (Toepoel & Lugtig, 2014) or contrary effects (Antoun, Couper & Conrad,

2017). In this experiment we tested if encouraging the use of speech-to-text would affect the length of the open answers. This was only possible for respondents who used a smartphone while completing the questionnaire.

In the experimental condition, open-ended questions in the survey included a short instructional prompt indicating that participants could use voice input instead of typing (Figure 15). The translated prompt was: "Tip: On many phones, it is also possible to dictate your answer. To do so, press the microphone on your keyboard. Your spoken response will be displayed as text."

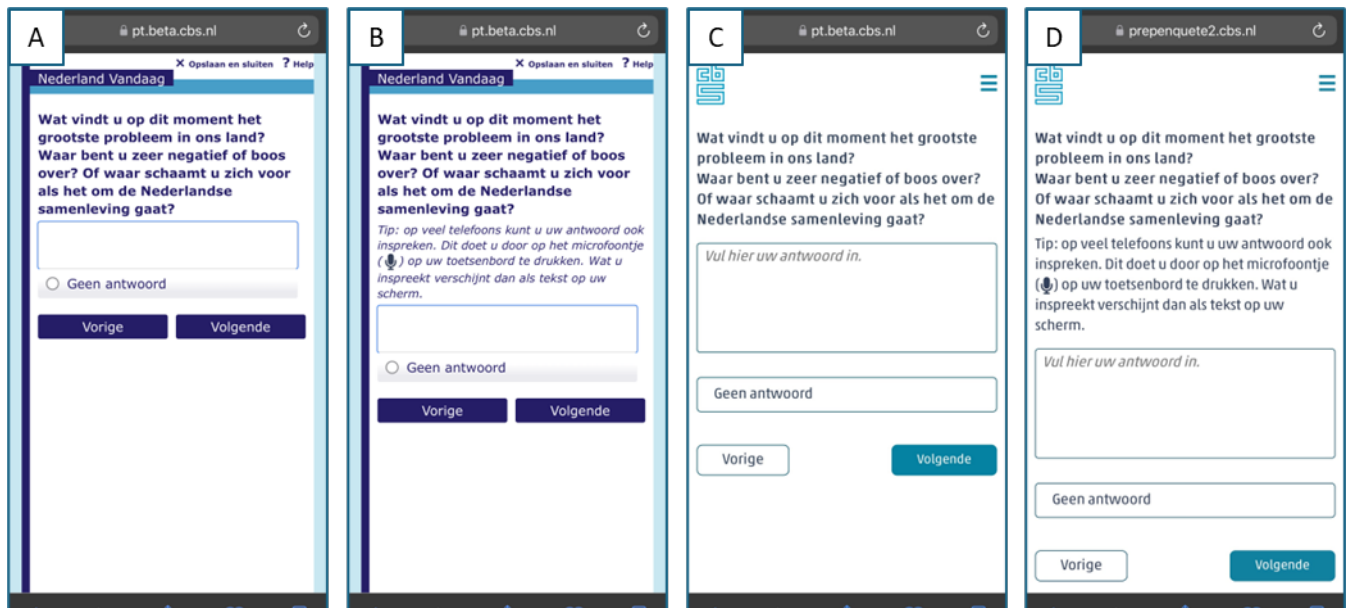


Figure 15 Open answer question without and with hint text to use speech to text in the regular design (A and B) and in the experimental design (C and D).

2.1.6 Shorter and longer version of the questionnaire

Questionnaire duration may impact the respondent's motivation and thus may impact how careful respondents answer the question and whether they break off the questionnaire. Especially on smartphones, questionnaire completion can be more burdensome as respondents have to read from a small screen and use a small touchscreen to select or type their answers. In this experiment, respondents randomly received either a shorter or a longer version of the questionnaire. The long questionnaire contained three extra question blocks compared to the short questionnaire, as can be seen in Table 1.

Table 1 Questionnaire blocks in short and long version of the questionnaire

Questionnaire Blocks	Short	Long
Household Composition	✓	✓
Type of Housing	✓	✓
Social Economic Status	✓	✓
News and Social Media	✓	✓
Satisfaction Life	-	✓
Society & Politics	✓	✓
Personality & Mental Health	-	✓
Employment	-	✓
Education	✓	✓
Allowances	✓	✓
Vehicles	✓	✓
Survey Attitudes & Evaluation	✓ ⁴	✓

The mean duration of the short version of the questionnaire was 17,3 minutes (SD = 12,5). For the long version, the mean duration was 24,7 minutes (SD=14,7).

2.2 Hypotheses

In this section the hypotheses per experimental condition are described.

The hypothesis was that the experimental login screen (and the way of presenting this in the letter) was more user friendly compared to the regular login screen. We expected more successful logins, more satisfaction about the process of logging in and more satisfaction about the letter for the group who received the experimental letter and login screen compared to the regular letter and login screen. Additionally, it was hypothesized that in the experimental condition, the respondents who completed the questionnaire on their smartphone would be more satisfied with respect to these two aspects compared to respondents who completed the questionnaire on a PC.

The experimental layout was designed to simplify completion of the questionnaire on the smartphone and the PC, and it was hypothesized that this would result in less break-off and higher quality of the response. Furthermore, the purpose was to minimize device effects between smartphone and PC and to adhere to the CBS design system. It was hypothesized that the experimental layout would result in less indications of undesirable response behavior and issues with data quality. Furthermore, the hypothesis was that the experimental layout would lead to less device-specific measurement errors and more satisfied respondents compared to the regular layout.

⁴ For the short version of the questionnaire a 3-item short version of the 9-item version of the survey attitude scale (Leeuw, Hox, Silber, Struminskaya & Vis, 2019) was used.

The grid designs were also developed with the aim to optimize usability and minimize device differences. It was expected that grid questions would be easier to complete when presented in the accordion or carrousel design compared to the regular grid designs. With respect to undesirable response behavior (e.g. straightlining), it was hypothesized that the regular grid design would perform worst, followed by the stem fix design. The hypothesis was that both the carrousel and the accordion design would be better than the stem fix and the regular design. Similarly, device effects were expected to be lower in the stem fixed, carrousel and accordion grid designs compared to the regular design, with the stem fix performing worse than the carrousel and the accordion. There were no hypotheses on whether the accordion or carrousel design would perform best on usability and device effects. However, it was important to assess any differences to guide further development of a new stylesheet. With respect to satisfaction, it was hypothesized satisfaction would be highest for the carrousel and accordion grid design, followed by the stem fix and regular design. Furthermore, it was expected that satisfaction about the carrousel and accordion design compared to the stem fix would be higher on a smartphone.

Smileys and icons were implemented because of their appealing effect and simplification of information processing. This could be beneficial especially on smartphones, on which processing text on the small screen can be more difficult compared to the larger screen. Various hypotheses could be applicable with respect to the implementation of either smileys or smileys with text. The advantage of smileys only compared to smileys with text is that the amount of information of the screen is minimized and therefore easy to process. However, the disadvantage of presenting solely smileys could be that interpretation becomes ambiguous. This could result in input errors and unintended differences in responses. For these reasons, no clear hypotheses were formed on the smileys only vs. smileys with text conditions. It was hypothesized that text only would result in more undesirable response behavior. In line with this, more device effects were expected in the text only presentation together with a lower satisfaction. The latter especially for respondent who complete the questionnaire on smartphone.

The speech-to-text encouragement was included for smartphone respondents to simplify giving a response to an open question. It was hypothesized that presenting this encouragement would result in longer answers to the open questions and a higher satisfaction about the questionnaire compared to respondents who do not receive the encouragement.

At last, the length of the questionnaire was varied to examine to which extend response behavior and satisfaction would depend on the length of the questionnaire. Additionally, it was examined to which extend this differed between smartphone and PC.

3. Methodology

3.1 Sample design and experimental design

As explained in Section 2, the effect of the following six conditions is tested in a large-scale field experiment:

1. Design login screen:
 - a. Regular login screen
 - b. Experimental login screen
2. Design questionnaire layout
 - a. Regular questionnaire layout
 - b. Experimental questionnaire layout
3. Grid design
 - a. Regular design, i.e. classic table for large screen and stem fix for small screen
 - b. Stem fix for large screen
 - c. Carousel
 - d. Accordion
4. Graphical symbols
 - a. No smileys, no icons for questions about social media (regular approach)
 - b. Smileys with text and icons (S-T-I)
 - c. Smileys without text and icons (S-NT-I)
5. Encouraging the use of speech to text for open questions
 - a. No encouragement (regular approach)
 - b. Encouragement
6. Length questionnaire
 - a. Short questionnaire
 - b. Long questionnaire

The target population for this field test comprises all persons residing in the Netherlands who are 16 years of age or older. To test these factors, a sample of 12,000 persons is drawn by means of simple random sampling without replacement from this target population. A net response of around 3,600 is expected. These six factors are combined in a field experiment as presented in Table 2. The first factor, the login screen, is not shown Table 2. Please note that the number of treatment combinations in Table 2 is assigned to both treatment combinations of the first factor, i.e. the standard login screen and the smartphone login screen. Consequently, a total of 88 treatment combinations are included in this experiment. The treatment combinations are further detailed in Table 3.

A design where the levels of all six conditions, i.e. a full factorial setup, is not applicable as accordion and carousel grid designs do not conform to the standard design questionnaire layout. The same applies to the use of graphical symbols. Consequently, the number of treatment combinations under the experimental questionnaire layout is 4.5 times greater than under the regular questionnaire

layout. As a result, the number of respondents assigned to the regular questionnaire layout is three times larger than that assigned to the experimental questionnaire layout. This ensures that the allocation of respondents between the regular and the experimental questionnaire layout remains approximately balanced.

Under both levels of the questionnaire layout, 60% of respondents are assigned to encourage the use of speech text, while 40% are not. This is because speech text can only be used on a smartphone, not a desktop or laptop. As it was expected that few respondents will choose this option, the two treatment levels of the factor encouraging the use of speech text are allocated using fractions 0.6 versus 0.4.

Based on the average observed response rates of CBS household surveys for different age categories, lower response rates are expected among young people. Therefore, compared to older people, the number of younger people is overrepresented, according to the distribution specified in Table 4. The number of people in each age category in Table 4 is allocated across the 88 treatment combinations according to the allocation in Table 2. This means that the experiment is designed using a randomized block design, with the nine age categories acting as the block variable, see e.g. Montgomery (1991), Chapter 5.

Table 5 specifies the allocation of the gross sample over the treatment combinations of login screen, questionnaire layout, encouraging the use of speech text and age categories. The treatment levels of the factors length questionnaire design, grid design and use of graphical symbols are allocated as follows in the treatment combinations of Table 5:

- Regular design style with encouraging use of speech text: 1,3,5,7
- Regular design style with no encouragement of speech text: 2,4,6,8
- New design style with encouraging use of speech text: 9, 10, 11, 15, 16, 17, 21, 22, 23, 27, 28, 29, 33, 34, 35, 39, 40, 41
- New design style with no encouragement of speech text: 12, 13, 14, 18, 19, 20, 24, 25, 26, 30, 31, 32, 36, 37, 38, 42, 43, 44

The numbers refer to the treatment combination numbers as specified in the first column of Table 2.

The allocation of the gross sample over all 88 treatment combinations, within the nine age categories (block variables) is specified in Table 6. Finally, an overview of the gross sample size in each treatment combination, aggregated over the blocks, is presented in Table 7.

Finally, the answer categories for two questions are presented in different orders to investigate if and how the experimental design affects response-order effects. Respondents are randomly assigned to one of the two versions when completing the questionnaire.

Table 2 Treatment combinations for questionnaire

Treatm. Nr.	Qnr. layout (2)	Length qnr. (6)	Factor			Sample size
			Grid design (3)	Speech Text (5)	Graph. symb. (4)	
1	Regular	Short	Regular	Yes	No	357
2	Regular	Short	Regular	No	No	246
3	Regular	Short	Stem fix	Yes	No	357
4	Regular	Short	Stem fix	No	No	246
5	Regular	Long	Regular	Yes	No	357
6	Regular	Long	Regular	No	No	246
7	Regular	Long	Stem fix	Yes	No	357
8	Regular	Long	Stem fix	No	No	246
9	New	Short	Stem fix	Yes	No	120
10	New	Short	Stem fix	Yes	S-T-I	120
11	New	Short	Stem fix	Yes	S-NT-I	120
12	New	Short	Stem fix	No	No	81
13	New	Short	Stem fix	No	S-T-I	81
14	New	Short	Stem fix	No	S-NT-I	81
15	New	Short	Accordion	Yes	No	120
16	New	Short	Accordion	Yes	S-T-I	120
17	New	Short	Accordion	Yes	S-NT-I	120
18	New	Short	Accordion	No	No	81
19	New	Short	Accordion	No	S-T-I	81
20	New	Short	Accordion	No	S-NT-I	81
21	New	Short	Carousel	Yes	No	120
22	New	Short	Carousel	Yes	S-T-I	120
23	New	Short	Carousel	Yes	S-NT-I	120
24	New	Short	Carousel	No	No	81
25	New	Short	Carousel	No	S-T-I	81
26	New	Short	Carousel	No	S-NT-I	81
27	New	Long	Stem fix	Yes	No	120
28	New	Long	Stem fix	Yes	S-T-I	120
29	New	Long	Stem fix	Yes	S-NT-I	120
30	New	Long	Stem fix	No	No	81
31	New	Long	Stem fix	No	S-T-I	81
32	New	Long	Stem fix	No	S-NT-I	81
33	New	Long	Accordion	Yes	No	120
34	New	Long	Accordion	Yes	S-T-I	120
35	New	Long	Accordion	Yes	S-NT-I	120
36	New	Long	Accordion	No	No	81
37	New	Long	Accordion	No	S-T-I	81
38	New	Long	Accordion	No	S-NT-I	81
39	New	Long	Carousel	Yes	No	120
40	New	Long	Carousel	Yes	S-T-I	120
41	New	Long	Carousel	Yes	S-NT-I	120
42	New	Long	Carousel	No	No	81
43	New	Long	Carousel	No	S-T-I	81
44	New	Long	Carousel	No	S-NT-I	81

Table 3 Overview of treatment combinations. Number refers to the treatment number combinations in the first column of Table 2. 5.a: no encouragement of use of speech text, 5.b: encouragement of use of speech text.

Length	Grid Des.	Graph. Symb.	Regular login screen				Exp. login screen			
			Regular qnr. layout		Exp. qnr. layout		Regular qnr. layout		Exp. qnr. layout	
			5.b	5.a	5.b	5.a	5.b	5.a	5.b	5.a
Short	3.a	4.a	1	2	-	-	1	2	-	-
		4.b	-	-	-	-	-	-	-	-
		4.c	-	-	-	-	-	-	-	-
	3.b	4.a	3	4	9	12	3	4	9	12
		4.b	-	-	10	13	-	-	10	13
		4.c	-	-	11	14	-	-	11	14
	3.d	4.a	-	-	15	18	-	-	15	18
		4.b	-	-	16	19	-	-	16	19
		4.c	-	-	17	20	-	-	17	20
	3.c	4.a	-	-	21	24	-	-	21	24
		4.b	-	-	22	25	-	-	22	25
		4.c	-	-	23	26	-	-	23	26
Long	3.a	4.a	5	6	-	-	5	6	-	-
		4.b	-	-	-	-	-	-	-	-
		4.c	-	-	-	-	-	-	-	-
	3.b	4.a	7	8	27	30	7	8	27	30
		4.b	-	-	28	31	-	-	28	31
		4.c	-	-	29	32	-	-	29	32
	3.d	4.a	-	-	33	36	-	-	33	36
		4.b	-	-	34	37	-	-	34	37
		4.c	-	-	35	38	-	-	35	38
	3.c	4.a	-	-	39	42	-	-	39	42
		4.b	-	-	40	43	-	-	40	43
		4.c	-	-	41	44	-	-	41	44

Table 4 Number of persons and expected respondents for 9 age categories

Age	Gross sample	Expected net sample	Expected net response
16-20	1920	0,250	480
21-25	1920	0,250	480
26-30	1920	0,263	504
31-40	1320	0,288	380
41-50	1320	0,308	406
51-60	1320	0,353	465
61-70	780	0,460	359
71-80	780	0,389	303
80 and older	780	0,342	267
Total	12060		3644

Table 5 Allocation gross sample over the factors login screen, design style, and encouraging the use of speech text and the block variable age. 5.a: no encouragement of speech to text; 5.b: encouragements of speech to text

Age	Standard login screen				Exp. login screen				Total
	Regular qnr. design		Exp. qnr. design		Regular qnr. design		Exp. qnr design		
	5.b	5.a	5.b	5.a	5.b	5.a	5.b	5.a	
16-20	228	156	342	234	228	156	342	234	1920
21-25	228	156	342	234	228	156	342	234	1920
26-30	228	156	342	234	228	156	342	234	1920
31-40	156	108	234	162	156	108	234	162	1320
41-50	156	108	234	162	156	108	234	162	1320
51-60	156	108	234	162	156	108	234	162	1320
61-70	92	64	144	90	92	64	144	90	780
71-80	92	64	144	90	92	64	144	90	780
80+	92	64	144	90	92	64	144	90	780
Total	1428	984	2160	1458	1428	984	2160	1458	12060

Table 6 Gross sample size for all 88 treatment combinations within the nine blocks defined by age category. 5.a no encouragement of speech to text; 5.b encouragement of speech to text

Age	Regular login screen				Exp. login screen			
	Regular qnr. design		Exp. qnr. design		Regular qnr. design		Exp. qnr. design	
	5.b	5.a	5.b	5.a	5.b	5.a	5.b	5.a
16-20	57	39	19	13	57	39	19	13
21-25	57	39	19	13	57	39	19	13
26-30	57	39	19	13	57	39	19	13
31-40	39	27	13	9	39	27	13	9
41-50	39	27	13	9	39	27	13	9
51-60	39	27	13	9	39	27	13	9
61-70	23	16	8	5	23	16	8	5
71-80	23	16	8	5	23	16	8	5
80+	23	16	8	5	23	16	8	5

Table 7 Gross sample size for all 88 treatment combinations. 5.a no encouragement of speech to text; 5.b encouragement speech to text.

			Regular login screen				Smartphone login screen			
			Regular design style		New design style		Regular design style		New design style	
Length	Grid Des.	Graph. Symb.	5.b	5.a	5.b	5.a	5.b	5.a	5.b	5.a
Short	3.a	4.a	357	246	-	-	357	246	-	-
		4.b	-	-	-	-	-	-	-	-
		4.c	-	-	-	-	-	-	-	-
	3.b	4.a	357	246	120	81	357	246	120	81
		4.b	-	-	120	81	-	-	120	81
		4.c	-	-	120	81	-	-	120	81
	3.d	4.a	-	-	120	81	-	-	120	81
		4.b	-	-	120	81	-	-	120	81
		4.c	-	-	120	81	-	-	120	81
	3.c	4.a	-	-	120	81	-	-	120	81
		4.b	-	-	120	81	-	-	120	81
		4.c	-	-	120	81	-	-	120	81
Long	3.a	4.a	357	246	-	-	357	246	-	-
		4.b	-	-	-	-	-	-	-	-
		4.c	-	-	-	-	-	-	-	-
	3.b	4.a	357	246	120	81	357	246	120	81
		4.b	-	-	120	81	-	-	120	81
		4.c	-	-	120	81	-	-	120	81
	3.d	4.a	-	-	120	81	-	-	120	81
		4.b	-	-	120	81	-	-	120	81
		4.c	-	-	120	81	-	-	120	81
	3.c	4.a	-	-	120	81	-	-	120	81
		4.b	-	-	120	81	-	-	120	81
		4.c	-	-	120	81	-	-	120	81

3.2 Fieldwork

For this study an experimental questionnaire was developed (Cremers, Giesen & Meertens, 2022) that covered a range of different topics and re-used questions from regular CBS surveys. The questionnaire was designed to ensure that respondents answered all questions. If a respondent attempted to move to the next page without completing all questions, an error message was displayed. Nearly all questions also included a “no answer” option. The questionnaire was programmed in Blaise 5, see Pouwels & Giesen (2025) for more details on the Blaise development for this experiment.

Sampled individuals were approached with the standard CBS approach strategy for CAWI (web) only household surveys. This entails an invitation letter and two reminder letters each three weeks apart. The invitation letter invited them to complete the survey "The Netherlands today" online. The survey was introduced as research on "a variety of topics you encounter in everyday life. Examples include: the news, social media, satisfaction with society and trust in politics." There was a lottery incentive of either an iPad or €400 in gift vouchers (by choice). Whether respondents had won was revealed at the end of the questionnaire. A standard brochure about CBS household surveys was enclosed with the letter. The letter also referred to a website with more information about the specific survey. The fieldwork lasted from January 27th, 2023 until March 12th, 2023.

3.3 Independent variables and control variables

3.3.1 Experimental conditions

All experimental conditions were treated as independent variables. Extensive descriptions and examples can be found in Section 2.1. Here we will shortly repeat the conditions. Two versions of the login screen were tested: regular vs. experimental. There were also two versions of the questionnaire layout: regular vs. experimental.

There were four alternatives for the grid design, however not all of them were available in both questionnaire layouts. The regular design condition is the regular situation in which the classic grid is shown in the regular questionnaire layout on PC, and stem fixed on smartphones. This condition could therefore only co-occur with the regular layout. The stem fixed grid was available for both layouts and devices. The newly designed carousel and accordion grid designs were only available in the experimental layout, on PC and smartphone.

There were three versions of the smiley and icon condition: text only, smileys and text, and smileys only. In the regular layout, the text alternative was always shown, but in the experimental layout respondents could also receive the smileys and text or smileys only. In case respondents received one of these smiley options, they also were exposed to icons in the question on social media use.

Speech to text encouragement was only visible for respondents who completed the questionnaire on smartphone, but it was possible in both layouts. Furthermore, two versions of the length of the questionnaire were tested: short vs. long.

3.3.2 Other independent variables

The key focus of the experiment was to examine device effects in the old and the new design, which is why also a variable indicating the type of device (PC or smartphone) was included in the analyses.

Additionally, to examine interactions between design alternatives and the order of answer options, the direction of the answer options was randomized for two questions (one question on interest in various types of news and one question on social media usage). For the analyses of these questions, a variable was included indicating the direction of the answer options.

3.3.3 Control variables

The following background variables were deemed relevant and important to control for in the analyses: age, gender, migration status, education, income and urbanity. Age was summarized in 9 categories (16-20, 21-25, 26-30, 31-40, 41-50, 51-60, 61-70, 71-80 and 81+), gender in two categories (male and female) and migration status in three categories (Dutch, 1st or 2nd generation western migration background, 1st or 2nd generation non-western migration background). Education was classified in 3 categories (low, medium, high), income in quintiles and urbanity in 5 categories (ranging from non-urban to very strongly urban).

Furthermore, the survey contained questions on survey attitude (Leeuw, Hox, Silber, Struminskaya & Vis, 2019), the presence of others and other activities done while completing the survey. These factors might influence response behavior and were therefore also used as control variables. Not all items of the survey attitude scale were included in the short version of the questionnaire; therefore, a summary variable was created using the three items that were included both in the short and the long version of the questionnaire for analyses involving all respondents, whereas a summary variable for the long version of the scale was used for the analyses involving respondents to the long version of the questionnaire only (see Appendix A, SurvAt_S and SurvAt_L). These items were rated on a fully labelled 7-point scale ranging from completely agree to completely disagree. Presence of others was a dichotomous variable (yes/no), just as doing other activities while completing the survey.

3.4 Dependent variables

To analyze the effect of the experimental conditions on response behavior and satisfaction, effects on many different dependent variables were analyzed. Table 8 provides an overview. For some indicators different operationalizations were developed. This was done for example to distinguish between items received by all respondents and by respondents in the long questionnaire only, or to focus on the

questions in which smileys were an experimental condition in the answer options. See Appendix B for a complete overview of all dependent variables, as well as their interpretation.

Table 8 Overview dependent variables

Dependent variables	Description
Login	Logged in
Break-off	Logged in & broken off
No answer*	% of items at which 'no answer' is given as response (5 dependent variables)
# of answers for multiple select questions*	Total number of selected answers in multiple select questions (2 dependent variables)
Primacy*	% items at which the first option is chosen (5 dependent variables)
Recency *	% items at which the last (substantial) option is chosen (5 dependent variables)
Effect direction treatment for interest in news / social media question	Mean interest in news /% respondents that select Facebook and % respondent that select LinkedIn (3 dependent variables)
Midpoint reporting*	% grid items at which the middle response option is chosen (4 dependent variables)
Straightlining*	Dummy indicating that within a grid the same answer is chosen for all items (11 dependent variables)
Acquiescence*	% items at which the 'agree'-answers are chosen (2 dependent variables)
Duration	Duration in minutes between time first login and time last login, only for questionnaires completed on the same day, cut off at mean + 2 x SD.
Substantial results*	Mean scores on various scales (10 dependent variables).
Mismatch register data	Dummy indicting that the information reported in the questionnaire did not match register data for type of house, health care allowance, child allowance, rent allowance and childcare allowance (5 dependent variables).
Overall satisfaction with questionnaire	Mean of 5 questions on satisfaction with various characteristics of the questionnaire
Satisfaction with logging in	Score on 5-point satisfaction scale
Satisfaction with letter	Score on 5-point satisfaction scale
Satisfaction with clarity questions	Score on 5-point satisfaction scale
Satisfaction with ease of completion	Score on 5-point satisfaction scale
Satisfaction with layout questionnaire	Score on 5-point satisfaction scale
Willingness to be contacted again for a similar survey	Dummy indicating that respondent answered yes to recruitment question if they may be contacted again for similar research.

* Different operationalizations, a.o. for short and long version questionnaire. See Appendix B for more details.

3.5 Analytical approach for main analysis

The analysis of the experimental conditions on the dependent variables, as specified in Subsection 3.4, is performed using Generalized Linear Models (GLM) in SPSS. The main effects of the six experimental factors and the block variable (age category) are included in the model as explanatory variables, as they define the experimental design. At the start of each analysis, the other covariates mentioned in Subsections 3.3.2 and 3.3.3 were also included as explanatory variables in the model. An exception are the dependent variables 'Logging in' and 'Break-off'. For these variables, the set of covariates is smaller because some control variables are unavailable for non-respondents.

To avoid data dredging or p-hacking, the number of regression models fitted to the data is kept to a minimum. This is achieved by including the main effects of all covariates in the models by default, even if they are not significant. Only significant second-order interaction effects between covariates are selected for the final analysis, which is performed using a backward variable selection procedure. Furthermore, Bonferroni multiple comparison methods are used to test the significance of treatment effects to control the overall significance level of the tests on treatment effects.

Most of the dependent variables are measured on an interval scale, which justifies the use of a linear Gaussian model. Levene's test for equality of variance is used to evaluate the homogeneity assumption of the observations. If this assumption is violated at a 5% significance level, the dependent variables are log-transformed to stabilize the variance. A few dependent variables are binary. For these, a logistic regression analysis is performed using the binomial link function in GLM. Covariate selection is similar to that described in the previous paragraph for linear Gaussian models.

Treatment effects are presented if the results of a Bonferroni multiple comparisons test are significant at the 95% confidence level. The effect of a treatment or experimental factor on the model is presented in terms of a Wald statistic, along with its corresponding p-value and the estimated marginal means (M) for the respective levels. Point estimates and standard errors are calculated for the estimated marginal means, and 95% confidence intervals are reported. For log-transformed variables, the point estimates and 95% confidence intervals are first calculated on the log scale using the standard errors of the log-transformed variable. The point estimates and their 95% confidence intervals are then presented on the original scale by applying the exponential transformation to the results obtained on the log scale. Consequently, the confidence intervals for these variables are asymmetric.

Cronbach's alpha is used to evaluate the effect of different grid design on three validated scales with reversed items. To investigate differences in reliability due to different grid designs, the statistical test for differences in Cronbach's alpha, as proposed by Diedenhofen and Musch (2016), was performed using the R package cocron.

3.6 Additional analysis login screen

In addition to the analysis of the effect of the login page condition on response, data quality, substantive indicators and respondent satisfaction, descriptive analyses were performed on additional data sources that provide information about the login behavior from a different perspective. The regular analyses are done on successful login attempts, whereas these additional data sources provide insight in the unsuccessful logins. The Piwik logging provides information on what happens on the login screen, whereas the IAW logging provides more detailed information about the actual login attempts.

3.6.1 Piwik

Piwik provides data on the sessions that occur on the login screen. Every time a user visits the login screen (doemee.cbs.nl or helpmee.cbs.nl), a log is created (see Figure 16). Piwik logs cannot be linked to individual respondents. Users who do not remove their cookies can be uniquely identified. Some overestimation occurs when users who do remove their cookies are identified as new users for every new visit, and underestimation occurs for users who use specific ad-blockers for tracking websites such as Piwik.

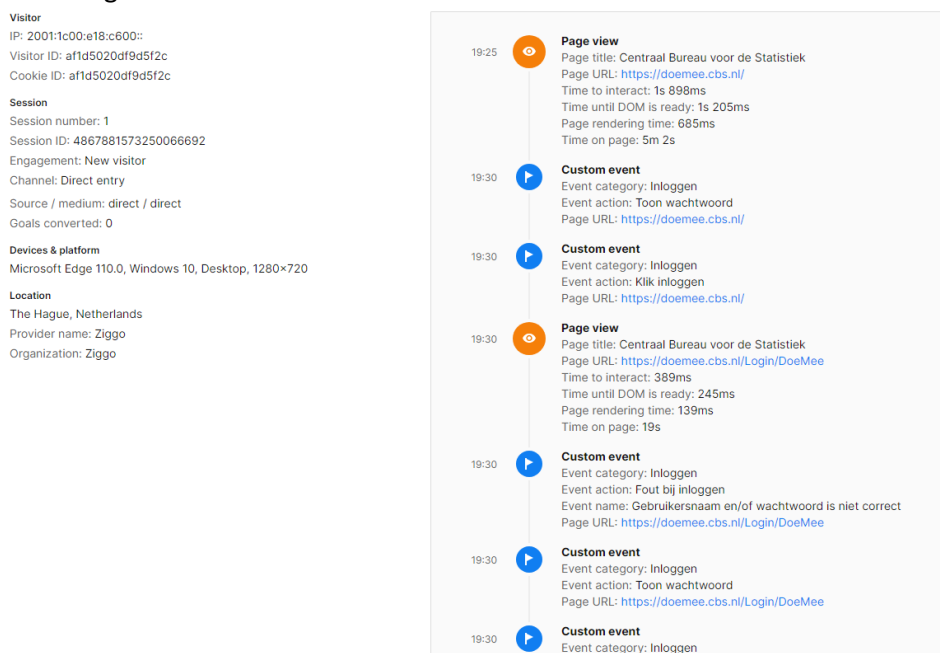


Figure 16 Example of Piwik log

Figure 16 shows that someone visited the screen directly (not via google f.e.) and tried to login. Before clicking the login button, the user looked at the password entered (event action “toon wachtwoord” means show password), but still the user provided a wrong user and password combination and ran into an error. Then the user repeated the process of looking at the password and clicking the login button. Apparently, the second login was successful, as the user left the login screen. From the Piwik logging, we cannot see what the user typed for the username and password.

For the analysis of the Piwik data, traffic between 27-1-2023 and 12-3-2023 on the doemee.cbs.nl and helpmee.cbs.nl pages were examined. Internal traffic (i.e.,

traffic from CBS employees) was excluded based on IP address. Furthermore, the data was filtered by the type of device that was used. With the Piwik data, we looked at the origin of the users on the login screens and the steps relating to logging in (f.e. use of the show password button, and the errors).

3.6.2 IAW

The IAW data provides information about the login process with information on the actual username, which (if correct) can be linked to respondent ids. Therefore, this data set provides more insight in the number of attempts that a user with one username does to login into the questionnaire. No information about the password is provided and no distinction can be made for the device that was used.

3.7 Additional analysis speech-to-text encouragement

In addition to the analysis of the effect of the speech-to-text condition on response, data quality, substantive indicators and respondent satisfaction, analyses were performed on the audit trails with the aim to investigate the actual use of the speech-to-text functionality.

Audit trails are a form of log data automatically generated in Blaise (Blaise 5.10 was used for this experiment) in which every action of each respondent is logged in an event log. For this experiment, audit trails were set to the lowest level (KeyStroke) in which every key stroke is logged. We expected that this level would provide information about the use of the microphone button on the users' keyboard, which would then identify the users who used the speech-to-text functionality. However, during data analysis we found that use of the microphone was not logged in the audit trail as the audit trail only logs what the browser receives. When using the microphone, the phone's operating system transcribes the input to text before sending it to the browser. In the input data no clear differences could be identified with which use of the microphone could be discovered, therefore respondents could not be classified as users of the speech-to-text functionality.

We tested whether showing the speech-to-text hint affected the answer length, the total number of typed characters, the difference between the answer length and the total number of typed characters and the duration of completing the question.

3.7.1 Measures

The answer length was determined by counting the number of characters in the final answer. The total number of typed characters was determined by counting the number of characters in the Keystrokes column of all edits in the question. Before doing this, special characters (f.e., 'enter' shown in the Keystrokes column as [ENTER] and 'delete' shown as [BACK]) were replaced by 1 character. The difference between the typed characters and the answer length was calculated by subtracting the number of typed characters from the answer length. At last, the question duration was calculated by subtracting the timestamp of entering the question from the timestamp of leaving the question.

For all aspects, the measure was first determined for each time a respondent worked on the question, and subsequently summarized per question. In this analysis, we only include the open questions that all respondents received: CountryProb (What do you consider to be the biggest problem in our country at the moment?), CountryPositive (And what do you think is currently going well in our country?), and EvalTips (Do you have any tips or comments about this questionnaire?) (see Table 9).

Table 9 Number of responses per open ended question

	CountryPositive	CountryProb	EvalTips
Answer	1183	1466	1729
Refusal	602	328	-

3.7.2 Analysis

All four aspects were analyzed using multilevel models in which the respondent identifier was included as a random intercept. The question name and condition were added as fixed effects, whereas the other conditions (Grid, Smiley, Order, Length and Stylesheet) were added as covariate just as Gender. The interaction between the stylesheet and the question name was tested to see whether the effect of the stylesheet differed per question, but this was not included in the final model as in none of the models this was significant.

For the significant predictors, estimated marginal means and standard error were calculated.

4. Results

4.1 Response

The overall response rate was 34%, resulting in 4088 completed questionnaires. Table 10 also shows the percentage of the sample logging in and breaking off. Additionally, the employed device to complete the questionnaire is shown. The device type is based on the screen size of the device used on the last login. "Other mobile" devices are usually new smartphone models. For the analyses smartphones and other mobile devices have been grouped together as "smartphones" and the large screen devices tablets and PCs as "PCs".

Table 10 Response and device use

Response	N	%	
Login	4586	38%	of sample
Break-off	498	11%	of logins
Completed	4088	34%	of sample
Device used for completion (completes only)			
Smartphone	1456		36%
Tablet	140		3%
Other mobile devices	199		5%
PC	2293		56%

All sample units who dropped out, did so at the beginning of the questionnaire: 10% after the introduction screen, 31% after the first question block (which verifies the sample unit's identity and assesses household composition), and 59% after the following block on dwelling type.

Table 11 provides an overview of the response by experimental condition and device. See Appendix C for an overview of the response by treatment combination. The effects of the experimental conditions on e.g. logging in and breaking off will be discussed with the main analyses in section 4.4.

Table 11 Response by experimental condition and device

		Device		Total
		PC/ tablet	Smartphone	
Login screen	Regular	1217	860	2077
	Experimental	1216	795	2011
Questionnaire layout	Regular	967	682	1649
	Experimental	1466	973	2439
Grid design	Regular	464	328	792
	Stem fixed	988	673	1661
	Carrousel	502	324	826
	Accordion	479	330	809
Smileys and icons	Text only	1447	1012	2459
	Smileys and text+icons	512	322	834
	Smileys only + icons	474	321	795
Speech-to-text encouragement	Without encouragement	2433	689	3122
	With encouragement	-	966	966
Length	Short	1234	839	2073
	Long	1199	816	2015

4.2 Device use and circumstances of survey completion

One potential cause of device effects is that the circumstances of the survey completion may differ over devices. Therefore, in the questionnaire these circumstances were assessed. As can be seen in Figure 17, most respondents completed the questionnaire at home regardless of device. This is probably because respondents are invited to participate in the survey by a letter sent to their address. Smartphone users were more often not alone while completing the survey than respondents on the PC (35% vs 27%). They were also more often engaged in other activities (31% vs 26%). Overall, responding on a smartphone appears to be more frequently associated with distraction during survey completion than responding on a PC.

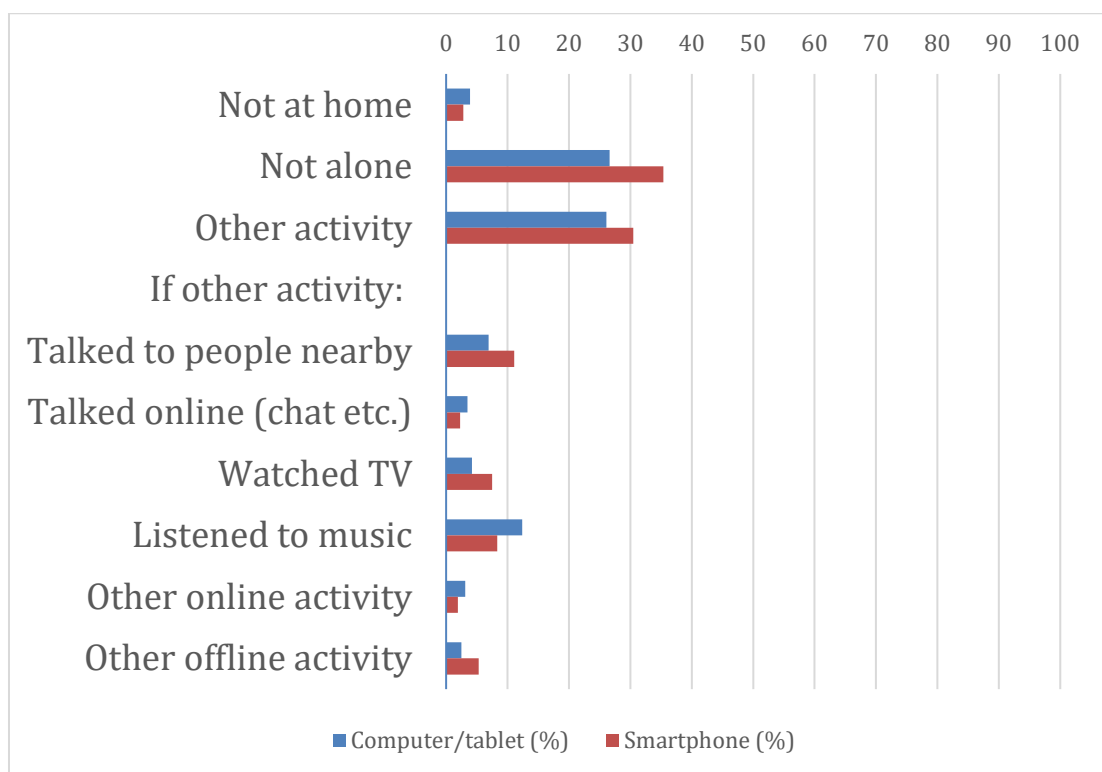


Figure 17 Survey completion conditions by device

4.3 Dependent variables

Overall, the quality of the responses was high. For example, the 'no answer'-option was not frequently used: on average in 5% of the items applicable to all and in 2% of the items in the long questionnaire only. Midpoint reporting occurred on average in 30% of the items applicable to all and in 20% of the items applicable to the long questionnaire only. Straightlining measured as choosing the same answer for all items in a grid, was assessed for 6 grid questions and varied between 2% (for a grid on trust in institution and a grid with the Ten Item Personality Inventory) to 11% for a grid with items on satisfaction with various aspects of society. For two grids with reversed items (the Brief Resilience Scale and the Mental Health Index) we also applied a stricter definition of straightlining, excluding respondents who selected 'no answer' or - if applicable - a midpoint answer for any item within the grid. Using this definition, providing the same answer for all items can be confidently regarded as an inconsistent response. Using this strict definition, straightlining was observed in 1% of respondents for both grids.

In Appendix D, all dependent variables are described in terms of minimum, maximum, mean and standard deviation.

4.4 Overview results analyses

In Table 12 and Table 13 an overview of the results is given. In the table headers it is indicated how many models of this type were tested, which depends on the number of operationalizations. See Appendix B for more details about the operationalizations. In Tables 12 and 13 it is shown how often each independent and control variables was significant in these models. Due to the two lengths of the questionnaire some variables could only be included in a selection of the models. In these cases, the number of models is shown between brackets. A hyphen in square brackets [-] indicates that an independent variable is not relevant for a model. A simple hyphen (-) indicates that an interaction effect was tested but was not significant, and therefore not included in the model. The independent variable direction is only relevant for the 3 analyses on the response order effect. For these analyses also second order interactions between response option direction, questionnaire layout, grid design and device were tested (but not included in the model as they were not significant). Full results, including all parameter estimates, are provided in the online supplement. https://www.cbs.nl/-/media/_pdf/2025/42/online-supplement-cbs-discussion-paper-experiment-smartphone-first-questionnaire-layout.pdf

Table 12 Overview of results I: indicating how often each independent and control variable was significant in the models with different operationalizations of the dependent variable. Number of analyses per type of dependent variable in brackets.

	Dependent variables						
	Login (1)	Breakoff (1)	No answer (5)	Multiple (2)	Primacy (5)	Recency (5)	Response order effect (3)
Independent variables							
Login screen	0	0	0	0	0	0	0
Questionnaire layout	[-]	0	0	0	0	1	0
Grid design	[-]	0	1	0	0	3	0
Smileys/icons	[-]	0	1	0	0	1	0
Speech*device	[-]	0	0	0	0	0	0
Length	[-]	0	0 (3)	0	1 (3)	0 (3)	0
Device	[-]	1	0	1	0	0	1
Direction	[-]	[-]	[-]	0 (1)	[-]	[-]	1
Login*Device	-	-	-	-	-	-	-
Layout*Device	[-]	-	-	-	-	-	-
Grid*Device	[-]	-	-	-	-	-	-
Smileys*Device	[-]	-	1	-	-	1	-
Length*Device	[-]	-	1	-	-	-	-
Direction*Layout	[-]	[-]	[-]	[-]	[-]	[-]	-
Direction*Grid	[-]	[-]	[-]	[-]	[-]	[-]	-
Direction*Device	[-]	[-]	[-]	[-]	[-]	[-]	-
Control variables							
Age	1	0	3	2	4	3	3
Gender	0	0	3	2	5	2	3
Migrant	1	1	2	1	1	1	0
Education	1	1	5	2	3	4	3
Income	1	0	0	1	3	3	3
Urbanity	0	0	0	2	2	0	1
Survey attitude	[-]	[-]	4	2	5	2	3
Presence others	[-]	[-]	0	0	4	0	0
Other activities	[-]	[-]	0	2	0	1	2

Table 13 Overview of results II

	Dependent variables							
	Midpoint-reporting (4)	Straight-lining (11)	Acquiescence (2)	N characters (1)	Duration (1)	Substantial results (10)	Mismatch Register (5)	Satisfaction (7)
Independent variables								
Login screen	0	0	0	0	0	0	0	2
Questionnaire layout	0	0	0	0	1	0	0	1
Grid design	0	1	0	0	0	0	0	0
Smileys/icons	2	0	0	0	0	1	1	5
Speech*device	0	0	0	0	0	0	0	0
Length	1 (2)	0 (5)	1 (1)	0	1	1(2)	0	5
Device	1	1	0	1	1	3	1	1
Direction	[-]	[-]	[-]	[-]	[-]	[-]	[-]	[-]
Login*device	-	-	-	-	-	-	-	-
Layout*device	1	-	-	-	1	-	-	-
Grid*device	-	1	-	-	1	-	-	-
Smileys*device	-	-	-	-	-	-	-	-
Length*device	-	-	1 (1)	-	-	-	-	-
Direction*Layout	[-]	[-]	[-]	[-]	[-]	[-]	[-]	[-]
Direction*Grid	[-]	[-]	[-]	[-]	[-]	[-]	[-]	[-]
Direction*Device	[-]	[-]	[-]	[-]	[-]	[-]	[-]	[-]
Control variables								
Age	3	3	2	1	1	8	3	2
Gender	4	4	1	1	0	7	0	3
Migrant	2	0	0	1	1	5	0	1
Education	2	2	0	1	1	8	1	5
Income	2	0	0	0	0	7	2	0
Urbanity	0	0	0	0	0	0	1	0
Survey attitude	3	6	2	1	1	7	1	7
Presence others	1	0	0	1	1	2	0	4
Other activities	0	1	1	0	1	5	2	2

4.5 The login screen

4.5.1 Results main analysis

The effect of the experimental login screen was examined in a total of 63 models (see Tables 12 and 13 for an overview and Appendix B for a detailed list of all dependent variables). The experimental login screen did not significantly affect any of the 56 response quality indicators analyzed. Two of seven satisfaction

indicators were affected: satisfaction with logging in (Sat_Login: Wald χ^2 (1) = 4.82, $p = .03$) and satisfaction with the layout of the questionnaire (Sat_Layout: Wald χ^2 (1) = 4.33, $p = .04$) were both significantly and negatively affected. Respondents expressed more satisfaction with the loginpage ($M = 4.20 \pm 0.03$) and the layout ($M = 4.21 \pm 0.03$) in the regular login screen condition compared to the experimental login screen condition (resp. $M = 4.16 \pm 0.03$; $p = .03$ and $M = 4.16 \pm 0.03$; $p = .04$).

4.5.2 Analyses on additional data sources

4.5.2.1 Origin of the users

For both pages, the origin of the users showed a similar pattern. 90% of the respondents directly typed in the url. 10% found the page via a search machine (Google, Bing, Yahoo, Duckduckgo and Ecosia) and about 0.25% found the page via www.cbs.nl.

4.5.2.2 Piwik

Table 14 shows that on both the regular (helpmee.cbs.nl) and the experimental (doemee.cbs.nl) screen many errors were made (20.7% vs. 49.2%), but the experimental screen performed clearly worse than the regular screen. For the old page, the difference between devices was small (19.7% on smartphone vs. 21.1% on the desktop), whereas on the experimental page more errors were made on the smartphone (52.9%) compared to the desktop (45.9%). With the experimental design, especially on the desktop the users used the show password button more frequently than in the regular design.

Table 14 Piwik data describing activity on login screens

	Sessions	Pressed the login button	Login in one go	Errors	Use show password
<i>Regular screen</i>					
Total	2364	1986	1366	412	286
smartphone	1067	866	665	171	73
desktop	1167	1020	634	215	203
<i>Experimental screen</i>					
Total	2504	2109	889	1036	580
smartphone	1062	871	407	461	106
desktop	1317	1152	444	529	453

4.5.2.3 IAW

Table 15 clearly shows more failed attempts by unique users on the experimental login screen.

Table 15 IAW data on login attempts

	Total attempts		Unique usernames	
	Successful login	Failed attempt	Successful login	Failed attempt
Regular screen	2810	1681	2411	702
Experimental screen	2564	1353	2237	1005

When examining the failed attempts more closely (Table 16), it can be noted that at the new login screen the vast majority of the failed attempts (96.5%) are with usernames that were in the sample, indicating a mistake in the password. On the old page, more errors were likely made in the username itself. To some extent this can be explained by respondents in the new screen condition that used the old screen and typed the hyphen (that was only needed in the new screen), but this does not fully explain the difference.

Table 16 IAW data on failed login attempts

	Failed attempt (unique)	Username in sample	Hyphen (-/_) in username	Username in sample if hyphen removed
Regular screen	702	433	269	216
Experimental screen	1005	979	14	0 ⁵

In total, there were 65 respondents in the sample that tried to login but never succeeded. About twenty respondents found both the old and the new screen and tried to login on both: ten of them were successful on both, ten failed (at least once) at both screens.

4.5.3 Conclusion login page

These results show that the experimental login screen did not improve survey response quality (including the likelihood of logging in) and negatively influenced respondent satisfaction: respondents reported lower satisfaction with both the login process and the questionnaire layout in the experimental condition.

Behavioral data from Piwik and IAW further underline these results: The experimental screen produced substantially more login errors (49.2% vs. 20.7% on the regular screen), particularly among smartphone users. Although the “show password” button was more frequently used on the experimental screen, this feature did not reduce errors. IAW data confirm a higher rate of failed login

⁵ As only 8 characters could be entered, it cannot be deducted if removing the hyphen would result in a correct username with the seven remaining characters.

attempts with the new screen, mostly due to incorrect password entries. In contrast, errors on the regular screen were more often attributable to mistakes in entering the username.

Note that the Piwik and IAW data show that many errors are made in the login process for both the experimental and the regular login page. The two sources do not agree on the number of login attempts because of the different methods of measuring. Furthermore, it should be noted that both sources have their drawbacks. The Piwik data does not provide insight in the data at a level that can be linked to individual respondents. The IAW data provides us more insight, but if the username is not in the sample, we still cannot deduct which user this was. Additionally, there is no information about what goes wrong in typing the password or what device was used.

4.6 Questionnaire layout

4.6.1 Results main analysis

The effect of the experimental questionnaire layout was examined in a total of 62 models (see Tables 12 and 13 for an overview and Appendix B for a detailed list of all dependent variables). For the 55 quality indicators analyzed, 3 significant effects were found, for the 7 satisfaction indicators 1 significant effect was found.

The design had a significant effect on one of the five operationalizations of recency (RE_all: Wald χ^2 (1) = 5.31, p = .02). Recency when summarized over all items applicable to all, was significantly lower in the regular questionnaire layout (M = 8.74, 95% CI [8.24-9.28]) compared to the experimental questionnaire layout (M = 9.41, 95% CI [9.02-9.82]; p = .02). Note that this effect was not found in the other four operationalizations of recency.

Furthermore, there was a significant interaction effect of the questionnaire layout and device for midpoint reporting in the question on satisfaction with society (MP_SatSoc: Wald χ^2 (1) = 15.53, p < .001). In the regular design, midpoint reporting occurred more on this question on the PC (M = 21.11, 95% CI [17.49-25.49]) compared to the smartphone (M = 14.98, 95% CI [12.42-18.07]; p < .01), whereas there were no differences between the two devices in the experimental design (PC: M = 18.60, 95% CI [16.09-21.49] and smartphone: M = 19.60, 95% CI [17.03-22.56]; p = 1.00). Note that this effect was not found in the other three operationalizations of midpoint reporting.

Duration was affected by the questionnaire layout. There was a significant main effect of questionnaire layout (Duration: Wald χ^2 (1) = 38.07, p < .001) and a significant interaction effect of questionnaire layout and device on the completion duration (Duration: Wald χ^2 (1) = 16.80, p < .001). In the regular design the duration was significantly shorter on the smartphone (M = 16.5, 95% CI [15.54-17.67]) than on the computer (M = 20.68, 95% CI [19.44-21.99]; p < .001), whereas in the experimental design there was no device effect. However, in the experimental design the duration on smartphone (M = 21.70, 95% CI [20.70-

22.75]) and PC ($M = 22.24$, 95% CI [21.24 ± 23.28]) was significantly longer than in the regular design on smartphone (resp. $p < .001$ and $p < .001$).

Finally, the experimental layout negatively affected satisfaction with the invitation letter (Sat_Letter: Wald χ^2 (1) = 4.12, $p = .04$). Satisfaction with the letter was lower in the experimental stylesheet ($M = 3.94 \pm 0.02$) compared to the regular stylesheet ($M = 4.01 \pm 0.03$; $p = .04$).

4.6.2 Conclusions questionnaire layout

Overall, the experimental questionnaire layout produced very few effects on response quality and satisfaction. Out of 62 models, only four significant effects were detected, and most of these were small and, in the case of recency and midpoint reporting, not consistent across different operationalizations. The experimental layout led to longer durations on both PC and smartphone, reducing the device differences found in duration in the regular design. Finally, a small negative effect was found for satisfaction with the invitation letter, but no other satisfaction indicators were affected.

These findings suggest that the experimental layout did not introduce systematic changes in response behavior nor satisfaction. Notably, there was no effect on the reporting of "no answer", despite its different visual presentation in the experimental design.

4.7 Grid design

4.7.1 Results main analysis

The effect of grid design was examined in a total of 62 models (see Tables 13 and 14 for an overview and Appendix B for a detailed list of all dependent variables). For the 55 quality indicators analyzed, 6 significant effects of grid were found. For the 7 satisfaction indicators no significant effects of grid were found.

The grid design had a significant effect on the frequency of no answer in the overall percentage of no answer in the 4 grids applicable to all (NAGR_all: Wald χ^2 (3) = 10.79, $p = .01$). Post-hoc analyses showed significantly less no answer in the accordion grid condition ($M = 3.26 \pm 0.49$) compared to the regular ($M = 5.19 \pm 0.50$; $p = .02$) and stem fixed conditions ($M = 4.47 \pm 0.38$; $p = .04$). In the carrousel condition, the percentage of no answer did not differ significantly from any of the other conditions ($M = 4.46 \pm 0.50$). Note that this effect was not found in the other three operationalizations of no answer in grid questions.

The grid design affected recency in three of the five operationalizations of recency (RE_all, RE_lo and REGR_all). Recency (RE_all: Wald χ^2 (3) = 10.46, $p = .02$) was significantly lower in the carrousel grid design ($M = 8.52$, 95% CI [8.03 - 9.05]) compared to both the stem fixed ($M = 9.28$, 95% CI [8.87 - 9.71]; $p = 0.01$) and the

regular grids ($M = 9.47$, 95% CI [8.92 - 10.05]; $p = .04$). When looking at the items of the long questionnaire only (all grid questions) (RE_lo: Wald χ^2 (3) = 10.35, $p = .02$), recency occurred significantly less in the group who saw the accordion grid design ($M = 5.24 \pm 0.51$) compared to the regular grid design ($M = 7.13 \pm 0.52$; $p = .03$). When zooming in on solely the grid questions applicable to all (REGR_all: Wald χ^2 (3) = 11.92, $p = .01$), the carrousel design ($M = 8.91$, 95% CI [8.19-9.70]) performed significantly better (less recency) than the stem fixed grid design ($M = 10.16$, 95% CI [9.52-10.84]; $p < .01$). The accordion ($M = 9.81$, 95% CI [9.02-10.68]) and the regular design ($M = 10.19$, 95% CI [9.35-11.10]) did not differ significantly from the other designs.

There were significant main effects of grid on straightlining on three (of eleven) operationalizations (SL_SatSoc: Wald χ^2 (3) = 11.97, $p < .01$, SL_DevInt: Wald χ^2 (3) = 18.78, $p < .001$ and SL_Trust: Wald χ^2 (3) = 9.68, $p = .02$). Post-hoc analysis showed no differences between the grids on the satisfaction with society grid. On the Device Use grid, there was significantly more straightlining amongst users who received the carrousel design ($M = 0.13 \pm 0.02$) compared to the accordion ($M = 0.09 \pm 0.02$; $p = .04$) and stem fixed design ($M = 0.08 \pm 0.01$; $p = .02$). The regular design did not differ significantly from any the others ($M = 0.11 \pm 0.02$). For the trust item, post-hoc analysis did not demonstrate any significant differences between the grids.

There was also a significant interaction effect of grid and device for straightlining (SL_MobUse: Wald χ^2 (3) = 10.51, $p = .02$). For both the regular grid design and the accordion there was more straightlining on smartphones (rep. $M = 0.30 \pm 0.03$ and $M = 0.29 \pm 0.03$) compared to computer users ($M = 0.19 \pm 0.03$; $p = .04$ and $M = 0.17 \pm 0.02$; $p < .01$).

Furthermore, duration was significantly affected by the interaction effect of grid and device (Duration: Wald χ^2 (3) = 8.68, $p = .03$). In the groups who received the stem fixed, carrousel or accordion grid design, there were device differences in duration of the questionnaire. In all three designs, respondents completed the questionnaire faster on the phone (resp. $M = 19.35$, 95% CI [18.43-20.31], $M = 18.81$, 95% CI [17.51-20.20] and $M = 17.44$, 95% CI [16.26-18.70]) compared to the computer (resp. $M = 21.57$, 95% CI [20.56-22.63]; $p < .01$, $M = 21.90$, 95% CI [20.54-23.35]; $p = .01$ and $M = 21.63$, 95% CI [20.27-23.07]; $p < .001$). This difference did not exist in the regular design (smartphone: $M = 20.37$, 95% CI [18.97-21.87] and PC: $M = 20.69$, 95% CI [19.40-22.06]).

Note that there were no interaction effects of grid design and the response order treatment. While one significant main effect of response order was found (on the mean interest in news) this did not differ over the grid designs. Also note that for two scales with reversed items (the Brief Resilience Scale and the Mental Health Index) a strict operationalization of straightlining was possible. This operationalization excluded cases with no answer and midpoint reporting (if relevant), thus ensuring that any straightlining in this strict operationalization reflects an inconsistent answer pattern. For these two analyses no differences in grids were found.

4.7.2 Results comparison Cronbach's alphas

Finally, for three validated scales with reversed items we also compared the Cronbach's alphas over the four grid designs. These involved the Brief Resilience Scale, the Mental Health Index and the Survey Attitude Scale. As shown in Table 17, the Cronbach's alphas hardly varied over the grid designs and no significant differences were found. Note that in the main analyses of strict straightlining, only the Brief Resilience Scale and the Mental Health Index were used as dependent variables, with the Survey Attitude Scale included as a control variable.

Table 17 Reliability by grid design

	Cronbach's alpha		
	Brief Resilience Scale	Mental Health Index	Survey Attitude
Grid design			
Regular	0,789	0,875	0,869
Stem fix	0,776	0,873	0,865
Carrousel	0,742	0,875	0,851
Accordion	0,735	0,875	0,864

4.7.3 Conclusion grid design

The experimental variations in grid design produced only a limited number of effects on response quality, and these were generally small and not consistent across operationalizations. Also, analyses of reliability showed no meaningful differences in Cronbach's alphas across the four grid designs, suggesting that internal consistency of the scales was unaffected by grid design. Taken together, grid design did not lead to systematic improvements or deteriorations in response quality or satisfaction. The few effects that emerged were small and not consistently replicated, indicating that different grid formats perform broadly similarly.

4.8 Smileys and icons

4.8.1 Results main analysis

The effect of smileys and icons was examined in a total of 62 models (see Tables 13 and 14 for an overview and Appendix B for a detailed list of all dependent variables). Note that the smiley and icon condition only impacted the presentation for the questions on satisfaction with society (applicable to all), satisfaction with life (applicable to the long questionnaire only) and social media use (applicable to all). Of course, it is possible that seeing smileys and icons in these questions might also affect response behavior in other questions. For the 55 quality indicators

analyzed, 6 significant effects were found. For the 7 satisfaction indicators 5 significant effects were found.

A significant effect of smileys and icons was found on the percentage of no answer in the 4 grid questions applicable to all (NAGR_all: Wald χ^2 (2) = 6.74, p = .03) and in recency in the 5 grid questions applicable to the long questionnaire only (RE_LO: Wald χ^2 (2) = 7.42, p = .02). In addition, there was a significant interaction effect with device for both these quality indicators (resp. Wald χ^2 (5) = 13.27, p = .02 and Wald χ^2 (2) = 8.55, p = .01). Amongst the group with text and smileys on the smartphone the percentage of no answer on grid questions was lowest ($M = 2.97 \pm 0.62$), and this differed significantly from the group with text only on smartphone ($M = 4.95 \pm 0.41$; p = .03). The other groups did not differ significantly. For recency, an effect was found when looking only at the group who received the long questionnaire; the group with text and smileys on the smartphone ($M = 7.65 \pm 0.63$) chose the last item more often compared to the groups with smileys only on smartphone ($M = 5.49 \pm 0.63$; p = .04) and on PC ($M = 5.39 \pm 0.59$; p = .03), and to the group with smileys and text on PC ($M = 5.29 \pm 0.59$; p = .02) Note that both the effects on no answer and on recency were not replicated in the other operationalizations of no answer and recency and also not in the questions that differed in these conditions.

Two (of the four) operationalizations of midpoint reporting were significantly affected by smileys and icons (MP_LO: Wald χ^2 (2) = 14.05, p < .01 and C5d MP_SatLife: Wald χ^2 (2) = 31.42, p < .001). Midpoint reporting occurred more often in the smileys only condition ($M = 16.65$, 95% CI [14.70-18.86]) compared to both the text only ($M = 13.70$, 95% CI [12.51-15.01]; p < .01) and the smileys and text condition ($M = 14.38$, 95% CI [12.69-16.30]; p = .04), when looking only at the group who received the long questionnaire. When zooming in on the question on Satisfaction with Life (in which smileys could be shown depending on the experimental condition), we see more midpoint reporting occurred on the item with smileys in the smileys only ($M = 25.58 \pm 1.40$) condition compared to the text and smiley ($M = 19.07 \pm 1.41$; p < .001) or text only conditions ($M = 18.89 \pm 1.02$; p < .001). Note that no effect of smileys was found in the question on satisfaction with life, the other grid question that differed with the smiley condition and was applicable to all respondents.

One operationalization of mismatch was significantly affected by the smileys in the design (Mismatch_RentAll: Wald χ^2 (2) = 7.54, p = .02). In the smileys only condition, the percentage of misreported data with respect to the rental allowance was the lowest ($M = 0.01 \pm 0.01$) and significantly lower than in the text only condition ($M = 0.03 \pm 0.01$; p = .02). Both did not differ significantly from the smileys and text condition ($M = 0.02 \pm 0.01$). Note that the presentation of this question this not differ depending on the smileys condition.

It should also be noted that icons had no effect on any of the dependent variables related to the social media use question, where icons could be displayed depending on condition: the number of social media platforms reported, the likelihood of reporting Facebook, and the likelihood of reporting LinkedIn.

Finally, the smileys significantly affected various satisfaction operationalizations: overall satisfaction (MeanSatQst: Wald χ^2 (2) = 13.30, $p < .01$), satisfaction with logging in (Sat_Login: Wald χ^2 (2) = 12.45, $p < .01$), satisfaction with the clarity of the questions (Sat_Clear: Wald χ^2 (2) = 9.34, $p = .01$), satisfaction with the ease of completion (Sat_Easy: Wald χ^2 (2) = 7.37, $p = .03$) and satisfaction with the questionnaire layout (Sat_Layout: Wald χ^2 (2) = 7.66, $p = .02$). Overall satisfaction was lowest in the text only condition ($M = 4.01 \pm 0.02$); significant differences existed with both the smileys only ($M = 4.08 \pm 0.02$; $p < .01$) and the text and smileys condition ($M = 4.08 \pm 0.02$; $p < .01$). In line with this, satisfaction with the login page was significantly lower in the text only condition ($M = 4.11 \pm 0.02$), compared to the other two conditions (smileys only: $M = 4.23 \pm 0.03$; $p < .01$ and smileys and text: $M = 4.20 \pm 0.03$; $p = .02$). Satisfaction with the clarity of the questions and the ease of completion were both significantly lower in the text only condition (resp. $M = 4.18 \pm 0.02$ and $M = 4.18 \pm 0.02$) compared to the smileys only condition (resp. $M = 4.27 \pm 0.03$; $p < .01$ and $M = 4.26 \pm 0.03$; $p = .02$). Both did not significantly differ from the smileys and text condition for these two satisfaction measures (resp. $M = 4.23 \pm 0.03$ and $M = 4.23 \pm 0.03$). Satisfaction with the layout was also lowest in the text only condition ($M = 4.14 \pm 0.02$), but this differed significantly from the text and smileys condition ($M = 4.21 \pm 0.03$; $p = .03$) and not from the smileys only condition ($M = 4.21 \pm 0.03$; $p = .05$).

4.8.2 Conclusion smileys and icons

The introduction of smileys and icons in a limited set of questions produced some significant effects, but these were scattered and not consistent across operationalizations. Effects were observed for no answer, recency, midpoint reporting, and mismatch, but they did not replicate across comparable measures and therefore do not indicate systematic changes in response behavior. Icons had no measurable impact on reporting of social media use.

By contrast, smileys did influence respondents' satisfaction ratings: the text-only condition consistently scored lowest, while both smiley conditions yielded slightly higher levels of satisfaction with the questionnaire, its clarity, and the login process.

4.9 Speech-to-text encouragement

4.9.1 Results main analysis

The effect of the speech-to-text encouragement was examined in a total of 62 models (see Tables 13 and 14 for an overview and Appendix B for a detailed list of all dependent variables). This experimental condition did not affect any of the quality indicators nor satisfaction.

4.9.2 Analysis on additional data source

In total, there were 1794 respondents who used a smartphone to complete the questionnaire in at least one of the sessions. They gave 5308 answers to the open questions and 930 refusals. In Table 18 the model results are shown.

4.9.2.1 Answer length

Answer length was not significantly affected by the speech-to-text hint. The length was significantly dependent on the question ($p < .001$) and the gender of the respondent ($p < .01$). Most characters were typed for the CountryProb question (104 ± 4), followed by CountryPositive (31 ± 4) and subsequently EvalTips (14 ± 4). Furthermore, women typed more characters (55 ± 3) than men (45 ± 4).

4.9.2.2 Typed characters

The total number of typed characters was also not significantly affected by the speech-to-text hint, but only by the question ($p < .001$) and the gender ($p = .02$). Most characters were typed at CountryProb (147 ± 9), followed by CountryPositive (35 ± 9) and EvalTips (10 ± 9). Again, women typed more characters (76 ± 8) than men (53 ± 9).

4.9.2.3 Difference in length and characters

The differences between the answer length and the number of typed characters was also not affected by showing the speech-to-text hint. The difference was only significantly dependent on the question ($p < .001$). For CountryProb (-36 ± 6) and CountryPositive (-1 ± 6) the difference was negative, meaning users types more characters than there were in the final answer. For EvalTips (3 ± 6) the difference was positive, indicating fewer typed characters than in the final answer.

4.9.2.4 Duration

Duration was also not affected by showing the speech-to-text hint. Again, the question did significantly affect the duration ($p < .001$). Additionally, the stylesheet version had a significant effect on duration ($p = 0.034$). Respondents needed most time for the CountryProb question (111.5 ± 6), followed by CountryPositive ($46.4s \pm 6$) and EvalTips (11.0 ± 6). Furthermore, respondents spent less time on the open questions in the old stylesheet ($44.6s \pm 9$) compared to the new stylesheet ($68.0s \pm 5$).

Table 18 Results multilevel models

	Answer length	Typed characters	Difference	Duration
	N = 5308	N = 5307	N = 5308	N = 5308
Speech-to-text hint	F(1,1683) = 0.43	F(1,1656) = 1.21	F(1,1666) = 0.85	F(1,1598) = 2.01
Question	F(2,3449) = 318.23 ***	F(2,3422) = 109.05 ***	F(2,3443) = 17.56 ***	F(2,3362) = 130.81 ***
Grid	F(3,1685) = 1.87	F(3,1659) = 0.57	F(3,1668) = 0.42	F(3,1600) = 1.76
Smiley	F(2,1684) = 1.08	F(2,1657) = 1.56	F(2,1666) = 1.76	F(2,1599) = 0.71
Order	F(1,1685) = 0.35	F(1,1658) = 0.97	F(1,1668) = 2.04	F(1,1600) = 1.00
Length	F(1,1684) = 0.11	F(1,1658) = 0.03	F(1,1668) = 0.17	F(1,1600) = 0.46
Stylesheet	F(1,1684) = 2.35	F(1,1658) = 3.15	F(1,1667) = 1.87	F(1,1599) = 4.48*
Gender	F(1,1686) = 7.23**	F(1,1659) = 5.46*	F(1,1669) = 3.19	F(1,1601) = 2.22

* = $p < .05$, ** = $p < .01$, *** = $p < .001$

4.9.3 Conclusion speech to text encouragement

Both in the main analyses and in the additional analyses we found no effects of showing the speech-to-text encouragement.

The positive difference between the answer length and the number of typed characters that was found for the EvalTips question, could be caused by the use of the speech-to-text functionality, but also by the use of autocompletion, autocorrection or swipe⁶ for example. Furthermore, an explanation for the longer duration in the new stylesheet is lacking as the answers did not differ between the stylesheet and there was also no interaction between stylesheet and question. Future research should first investigate with the help of different devices, settings and scenarios how different behaviors are logged in the audit trail. With that knowledge, results like the ones we found in this analysis can be better interpreted and explained.

4.10 Short and longer version of the questionnaire

4.10.1 Results main analysis

The effect of questionnaire length was examined in a total of 40 models: 33 response quality indicators and 7 satisfaction indicators (see Tables 13 and 14 for

⁶ Text entry by 'swiping' from letter to letter without lifting the finger from the keyboard.

an overview and Appendix B for a detailed list of all dependent variables). Significant effects of length were found in 6 response quality indicators and 5 satisfaction indicators.

The long questionnaire took significantly longer to complete (Duration: Wald χ^2 (1) = 650.75, $p < .001$; $M = 24.40$, 95% CI [23.49-25.36]) compared to the short version ($M = 16.66$, 95% CI [16.03-17.31]).

The length of the questionnaire had a significant effect on one of the three operationalizations of primacy (PR_SatSoc: Wald χ^2 (1) = 8.95, $p < .01$), a significant effect on one of the two operationalizations of midpoint reporting (MP_all: Wald χ^2 (1) = 6.84, $p < .01$) and a significant effect the only operationalization of acquiescence (AC_all: Wald χ^2 (1) = 16.82, $p < .001$). The effect on acquiescence was also dependent on the device, as indicated by a significant interaction effect of length by device (Wald χ^2 (1) = 4.75, $p < .03$). Primacy occurred less often on the Satisfaction with Society question amongst respondents who completed the long questionnaire ($M = 1.25$, 95%CI [1.17-1.33]) compared to those with the short questionnaire ($M = 1.35$, 95%CI [1.26-1.44]; $p < .01$). Note that in the long version of the questionnaire, respondents received a question on satisfaction with life immediately before the question on satisfaction with society. This was the only difference between the two versions up to that point.

Midpoint reporting occurred more in the long questionnaire ($M = 30.25 \pm 0.82$) compared to the short questionnaire ($M = 28.60 \pm 0.82$; $p < .01$). Acquiescence occurred least often in the long questionnaire on the computer ($M = 52.39 \pm 1.10$) compared to the long questionnaire on the smartphone ($M = 55.65 \pm 1.09$; $p = .03$), and to the short questionnaire both on the computer ($M = 56.80 \pm 1.10$; $p < .001$) and on the smartphone ($M = 57.00 \pm 1.08$; $p < .01$).

There was an interaction effect of length by device on the percentage of no answer (NAGR_all: Wald χ^2 (3) = 11.56, $p < .01$). In the short questionnaire on smartphone the percentage of no answer was lowest ($M = 3.22 \pm 0.46$). This was significantly lower compared to both the short and the long questionnaire on the computer (resp. $M = 4.91 \pm 0.46$; $p = .03$ and $M = 4.71 \pm 0.46$; $p = .01$), but not significantly lower compared to the long questionnaire on smartphone ($M = 4.34 \pm 0.46$; $p = .05$).

Content wise, the length of the questionnaire affected the scale value of satisfaction with society (MeanSatSoc: Wald χ^2 (1) = 13.45, $p < .001$). The longer version of the questionnaire resulted in a lower score ($M = 2.95 \pm 0.03$) for satisfaction with society ($M = 3.03 \pm .03$; $p < .001$). Remember that in the long version of the questionnaire, respondents received a question on satisfaction with life immediately before the question on satisfaction with society.

Finally, the length of the questionnaire had significant negative effects on multiple aspects of satisfaction (MeanSatQst: Wald χ^2 (1) = 91.71, $p < .001$, Satn_letter: Wald χ^2 (1) = 5.64, $p = .02$, Satn_clear: Wald χ^2 (1) = 27.05, $p < .001$, Satn_easy: Wald χ^2 (1) = 46.45, $p < .001$, Satn_layout: Wald χ^2 (1) = 16.23, $p < .001$). The longer version of the questionnaire was evaluated more negatively overall ($M =$

3.99 ± 0.02), with specifically a lower satisfaction with the letter (M = 3.95 ± 0.03), the clarity of the questions (M = 4.18 ± 0.02), the ease of completion (M = 4.16 ± 0.02) and the layout (M = 4.15 ± 0.03) compared to the short version (overall: M = 4.12 ± 0.02; p < .001, letter: M = 4.00 ± 0.03; p = .02, clarity: M = 4.27 ± 0.02; p < .001, ease: M = 4.28 ± 0.02; p < .001 and layout: M = 4.22 ± 0.03; p < .001). Note that the likelihood to agree to be contacted again for a similar survey was not affected by the length of the questionnaire.

4.10.2 Conclusion questionnaire length

Beyond duration, effects of questionnaire length on response quality were modest and not fully consistent. Some indicators (primacy, midpoint reporting, and acquiescence) differed between the long and short versions, these differences do not seem structural and were small in size. This also goes for the device interactions that were observed for acquiescence and no answer. Furthermore, the differences found in the answer on the satisfaction with society question are likely caused by a context effect (with the respondents in the long version of the questionnaire answering this question directly after a satisfaction with life question).

The more robust and consistent findings concern respondent evaluations. The long questionnaire was rated less positively on multiple aspects, including overall satisfaction, clarity, ease of completion, layout, and even the invitation letter. Overall, the evidence suggests that the longer questionnaire increase respondent burden—both objectively, through longer completion times, and subjectively, through lower satisfaction ratings.

4.11 Device effects

4.11.1 Results main analysis

The main and interaction effects of device were examined in a total of 62 models (see Tables 12 and 13 for an overview and Appendix B for a detailed list of all dependent variables). In the models significant effects of device were found in 14 quality indicators and 1 satisfaction indicator.

The significant interaction effects with device and the experimental conditions have already been reported in the previous sections (i.e. layout, grid, smileys and length). Here we report the remaining significant main effects of device (smartphone vs. PC) on the dependent variables.

There was a significant effect of device on break-off in the questionnaire (Breakoff: Wald χ^2 (1) = 14.05, p < .001). There was significantly more breakoff on smartphone (M = 0.16 ± 0.02) compared to the PC (M = 0.10 ± 0.01).

Device significantly affected the reported number of social media accounts (SUM_SocMed: Wald χ^2 (1) = 12.94, p < .001). More social media accounts were selected by respondents who completed the questionnaire on smartphone (M = 2.73 ± 0.05) compared to PC (M = 2.50 ± 0.06). Additionally, device significantly

related to having a Facebook account (SocMed_FB: Wald χ^2 (1) = 6.39, $p = .01$). Smartphone users had a Facebook account more often ($M = 0.67 \pm 0.02$) than computer users ($M = 0.60 \pm 0.02$).

Furthermore, device significantly affected the number of characters that were used to respond to the open questions (Nchar: Wald χ^2 (1) = 31.39, $p < .001$). Smartphone users typed significantly less characters ($M = 108.20 \pm 11.30$) than respondents did when responding via the PC ($M = 177.01 \pm 12.12$).

The mean score on the scale of satisfaction with life significantly varied with the device used when completing the questionnaire (MeanSatLife: Wald χ^2 (1) = 4.18, $p = .04$). Smartphone users scored their life higher ($M = 3.94 \pm 0.03$) compared to PC users ($M = 3.87 \pm 0.03$). Moreover, the score on the extraversion subscale of TIPI and on the mental health scale were device-dependent (resp. EXTRA: Wald χ^2 (1) = 17.32, $p < .001$; MHI: Wald χ^2 (1) = 4.01, $p = .05$). Smartphone users were more extravert ($M = 4.58 \pm 0.08$) and scored higher on mental health ($M = 70.90 \pm 0.94$) compared to PC users (EXTRA: $M = 4.19 \pm 0.09$ and MHI: $M = 68.84 \pm 1.01$).

One of the five operationalizations of mismatch was significantly affected by device (Mismatch_ChildBudg: Wald χ^2 (1) = 11.86, $p < .01$). There were more mismatches on whether child-related budget was received on smartphone ($M = 0.01 \pm 3.16$) compared to PC (0.00 ± 1.43).

At last, satisfaction with the process of logging in significantly differed dependent on the device (Sat_Login: Wald χ^2 (1) = 7.44, $p = .01$). Smartphone users evaluated this process less positive ($M = 4.14 \pm 0.03$) compared to PC users (4.22 ± 0.03).

4.11.2 Conclusions device effects

Device use (smartphone vs. PC) as a main and/or interaction effect was significant in 15 of the 62 analyzed models.

Respondents on smartphones were more likely to break off the questionnaire (16% vs. 10%). Also, they provide substantially shorter answers to open questions (on average 108 characters vs. 177 on PC). The effect of device on other response quality indicators varied and showed no clear patterns. It was notable that device is related to different scores on social media use, satisfaction with life, mental health and extraversion. Device use did not seem to impact satisfaction with the response process, except for a negative impact on satisfaction with logging in.

5. Discussion

This experiment examined the effects of a smartphone-first redesign of the login process and questionnaire layout, as well as several innovations in question formats. The study was carried out in a fresh sample of the Dutch population aged 16 and older, with a gross sample size of 12,000 persons drawn by means of simple random sampling without replacement. The experiment is designed as a randomized block design using nine age categories as block variables. Testing six different treatment factors in one experiment with persons randomly drawn from the Dutch target population increases the generalizability of the observed results. Moreover, the questionnaire design and fieldwork were implemented in a setting that is practically identical to the conditions of CBS's regular household surveys. This ensures that the findings are directly relevant and transferable to CBS practice.

Across a wide range of indicators of response behavior and respondent satisfaction, the results revealed remarkably few significant effects. This finding is reassuring: it suggests that our respondents largely provide answers regardless of the tested design alternatives, reducing concerns about interruptions in time series or systematic measurement bias when implementing new designs.

The effects that were observed were generally small and sometimes mixed in direction. For example, certain grid formats slightly reduced the percentage of 'no answer' or recency effects, but no variant consistently outperformed others across all indicators. Some effects are difficult to interpret—for instance, the impact of questionnaire layout on satisfaction with the letter, but not on satisfaction with the questionnaire layout itself—and may represent a chance finding.

The experiment confirmed the persistence of two well known device-related differences: break-off rates were higher for smartphone respondents and open answers are shorter on smartphones. These differences were not resolved by the new design or other innovations. It is uncertain whether questionnaire design alone can eliminate this gap. Smartphone respondents reported being more often distracted during completion. If distraction is more common among smartphone users who break off, design interventions may have limited impact. Furthermore, substantive differences between devices in outcomes such as social media use, mental health, and extraversion suggest that self-selection may also play a role, with smartphone respondents differing systematically from PC respondents even after controlling for background characteristics. These findings underline the challenge of distinguishing between device effects and selection effects and will need continued attention in future monitoring and research.

This study also highlights the challenges in assessing the effects of variations in questionnaire length. First, the difference in length in this study may have been too small to produce measurable effects on response quality. Second, it is inherently difficult to design questionnaires of different lengths without introducing contextual differences. The lower satisfaction observed among

respondents completing the longer questionnaire may partly reflect the fact that they received different questions that were potentially more burdensome to answer compared to those completing the shorter version. Future studies that aim to investigate the effects of questionnaire length should pay greater attention to ensuring that the questions in the short and long versions are comparable both in format and content.

At the same time, these findings provide a clear basis for CBS's design choices. In a multidisciplinary workshop, it was concluded that the smartphone-first design should be implemented, despite the modest effects, because it aligns with the CBS corporate style guide, improves accessibility and provides a consistent experience over devices. Since major design effects were not found, there is no urgency to migrate all questionnaires to the new design; instead, a gradual rollout is planned. For grid questions, the accordion design is the preferred option, though some technical adjustments are still needed. If these cannot be resolved, the stem-fix variant will serve as a fallback across devices. The results of this experiment indicate that smileys may modestly enhance respondents' perceived experience of the questionnaire. However, they are suitable for only a very limited set of questions, and the results also suggest that their use could increase straightlining on certain items. Therefore, smileys and icons are not implemented and currently, no further research into the use of smileys or icons is planned for CBS questionnaires. The same applies to the speech-to-text encouragement.

This study also showed that the login process clearly requires further improvement. In a separate project, efforts to simplify the login procedure were further developed. To gather input on the user-friendliness of the login page, a range of sources was consulted, including complaints received by the CBS contact center, insights from a panel of people with low literacy, findings from Userlab tests, and literature on login behavior. Based on this input, proposals for improvements were formulated for both the login page (antwoord.cbs.nl) and the login instructions in the survey invitation letters. These proposals were tested on a small scale using iterative qualitative methods. The resulting changes—such as lowercase-only passwords, a caps lock notification, and clearer error messages with step-by-step help instructions—have been implemented and rolled out since January 2025. Since then, the number of login errors has declined significantly.

In summary, while the experiment did not reveal clear improvements in data quality or respondent satisfaction, the smartphone-first layout offers clear advantages in alignment with the CBS corporate style guide, accessibility, and consistency across devices. Its implementation at CBS will therefore proceed step by step, supported by ongoing monitoring to detect any unintended consequences in production surveys.

Acknowledgments

We would like to acknowledge the contribution of many colleagues to this project, among others Michelle Creemers, Madelon Cremers, Mark Durlinger, Maarten Pouwels, Bas Huijts, Stefan Theunissen, Darren den Boer, Duco Hoogendoorn, Vivian Meertens, Anne Reedijk, Marina Veldhuizen, Jeroen Schröder, Barry Schouten and Vera Toepoel.

The views expressed in this paper are those of the author(s) and do not necessarily reflect the policies of Statistics Netherlands

References

- Antoun, C., Couper, M. P., & F.G. Conrad (2017). Effects of mobile versus PC web on survey response quality: A crossover experiment in a probability web panel. *Public Opinion Quarterly*, 81(S1), pp. 280-306.
- Couper, M., Antoun, C., & Mavletova, A. (2017). Mobile Web Surveys: A Total Survey Error Perspective. In P. Biemer, E. d. Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. Lyberg, & B. W., *Total survey error in practice* (pp. 133-154). Hoboken, NJ.: John Wiley & Sons.
- Cremers, M., Giesen, D., & Meertens, V. (2022). *Smartphone First Experiment Vragenlijst*. Heerlen: Statistics Netherlands
- Diedenhofen, B., & Musch, J. (2016). cocron: A web interface and R package for the statistical comparison of Cronbach's alpha coefficients. *International Journal of Internet Science*, 11, 51–60.
- Giesen, D., Bakker, J., Meertens, M., & Luiten, A. (2019). *Gebruiksvriendelijkheid vragenlijsten persoonswaarnemings op tablet, smartphone en PC*. Heerlen: Statistics Netherlands.
- Leeuw, E. d., Hox, J., Silber, H., Struminskaya, B., & Vis, C. (2019). Development of an international survey attitude scale; measurement equivalence, reliability, and predictive validity. *Measurement Instruments for the Social Sciences*, pp. 1-10.
- Menold, N., & Toepoel, V. (2022). Do Different Devices Perform Equally Well with Different Numbers of Scale Points and Response Formats? A test of measurement invariance and reliability. *Sociological Methods & Research*, 53(2), pp. 898-939.
- Montgomery, D.C. (1991) *Design and Analysis of Experiments*. Third edition. Wiley & Sons, New-York.
- Pouwels, M., & Giesen, D. (2025). Harmonized matrix layout for different devices. . *Paper presented at the 21th International Blaise User Conference IBUC*. Rotterdam, April 8-10 2025.
- Revilla, M., & Ochoa, C. (2016). Open narrative questions in PC and smartphones: is the device playing a role? *Qual Quant* 50, pp. 2495–2513.
- Roberts, A., & Bakker, J. (2018). *Mobile device login and break-off in individual surveys of Statistics Netherlands*. Statistics Netherlands Discussion Paper.
- Schouten, B., Blanke, L., Gravem, D., Luiten, A., Meertens, V., & Paulus, O. (2018). *Assesment of fitness of ESS surveys for smartphones. WP5 deliverable 1. Cooperation on Multi-Mode Data Collection (MMDC) Mixed Mode Designs for Social Surveys – MIMOD*. EU Grant agreement 07112.2017.010-2017.786 .
- Toepoel, V., & Lugtig, P. (2014). What happens if you offer a mobile option to your web panel? Evidence from a probability-based panel of internet users. *Social Science Computer Review*, 32(4), pp. 544-560.
- Toepoel, V., Vermeeren, B., & Metin, B. (2019). Smileys, stars, hearts, buttons, tiles or grids: influence of response format on substantive response, questionnaire experience and response time. *Bulletin of Sociological Methodology/Bulletin de Méthodologie* 142(1), pp. 57-74.

van Hees, R. (2019). *Herontwerp van de CBS vragenlijst*. Maastricht: Hogeschool
Zuyd Bachelor Thesis.

Appendix A Details Grid Questions

Question	Question stem	Items	Answer options	Remark
News-Interest	If you follow the news, to what extent are you interested in	<ol style="list-style-type: none"> Domestic news? Foreign news? News about politics? 	5 point unipolar, fully labeled, very interested - not interested / not interested - very interested (direction randomized)	<p>* Only shown if respondents answered that they follow the news in a previous question.</p> <p>* Relative assessment of items may be more important for this question.</p>
Device-Internet	The following questions are about your Internet use in general. On which devices have you used the Internet in the past month, that is, since <date 1 month ago shown>?	<ol style="list-style-type: none"> Personal Computer (PC) or desktop Laptop Tablet Mobile phone Other device, such as a Smart TV, game console, e-reader or smartwatch 	2 options: yes / no	
MobPhone-Use	Do you ever use your mobile phone to:	<ol style="list-style-type: none"> Make purchases over the Internet? Pay contactless at a store or restaurant? Do online banking? Stream music, movies or videos? Play games? 	2 options: yes / no	Only shown if answered yes on question about using the Internet on a mobile phone in the past month.

Question	Question stem	Items	Answer options	Remark
		6. Read or post messages on social media?		
SatLife (long only)	The following questions are about satisfaction. How satisfied or dissatisfied are you with:	1. The life you lead now 2. The financial situation of your household 3. The education you have had so far 4. Your job* 5. The amount of free time you have 6. Your circle of friends and acquaintances 7. Your health	5 point bipolar, fully labeled, very satisfied - very dissatisfied (depending on experimental group text only, text and smiley or smiley only)	* item "your job" only shown to respondents who indicated that they had a job. * Relative assessment of items may be more important for this question.
SatSoc	The following questions are about your satisfaction with society and politics. How satisfied or dissatisfied are you with:	1. Dutch society 2. The Dutch economy 3. The administration of your municipality 4. National politics 5. European politics	5 point bipolar, fully labeled, very satisfied - very dissatisfied (depending on experimental group text only, text and smiley or smiley only)	* For respondents in long qnre shown after question on satisfaction about life. Respondents in short version of the qnre did not receive the satisfaction about life question. Up until this question the short and long qnr were identical. Potential context effect. * Relative assessment of items may be more important for this question.

Question	Question stem	Items	Answer options	Remark
Trust	(\$1: We will now show you a number of institutions or organizations in the Netherlands. \$2: Below is a number of institutions or organizations in the Netherlands). Can you indicate on a scale of 1 to 10 how much trust you have in them? A 1 means “no trust at all” and a 10 “full trust”.	<ol style="list-style-type: none"> 1. The big corporations / companies 2. The trade unions 3. The newspapers, both paper and websites / apps 4. The news on TV 5. The judiciary 6. The House of Representatives 7. The government 8. The police 9. Politics in my municipality 10. European politics; The Central Bureau of Statistics (CBS) 	10 point unipolar, endpoint labeled, 1 no trust at all - 10 full trust	<p>* Intro 1 shown for carousel format, intro 2 for all other grid formats.</p> <p>* Grid with many items and answer options</p> <p>* Relative assessment of items may be more important for this question.</p>
Resilience (long only)	The following questions are about how you feel about life. To what extent do you agree or disagree with the following statements?	<ol style="list-style-type: none"> 1. I tend to bounce back quickly after difficult times. 2. I find it difficult to endure stressful events. 3. I don't take much time to recover from a stressful event. 4. It is difficult for me to move on when something unpleasant happens. 5. I usually have little trouble getting through difficult times. 6. I tend to take a lot of time to get over setbacks in my life. 	5 point bipolar fully labeled, completely agree-completely disagree	<p>* Potentially sensitive question;</p> <p>* Reverse worded items (brief resilience scale)</p>

Question	Question stem	Items	Answer options	Remark
Personality (long only)	Below are some statements about traits a person may have. To what extent do you agree or disagree with these statements when it comes to yourself? I see myself as: >>Is one description more applicable to you than the other? Then try to give an average rating of both.<<	<ol style="list-style-type: none"> 1. Extraverted, enthusiastic 2. Critical, belligerent 3. Reliable, disciplined 4. Anxious, easily upset 5. Open to new experiences, complex 6. Reserved, quiet 7. Sympathetic, warm 8. Careless, careless 9. Calm, emotionally stable 10. Restrained, not creative 	7 point bipolar, fully labeled, completely agree-completely disagree	<p>* Potentially sensitive question;</p> <p>* Introduction sentence on small screen on separate screen, on large screen on top of screen;</p> <p>*Reverse worded items (ten item personality inventory)</p>
MentalHealth (long only)	The following questions are about how you have felt in the past 4 weeks. How often:	<ol style="list-style-type: none"> 1. Were you feeling very nervous? 2. Were you so depressed that nothing could cheer you up? 3. Did you feel calm and quiet? 4. Did you feel dejected and gloomy? 5. Did you feel happy? 6. Did you feel stressed? 	6 point bipolar, fully labeled, all the time-never.	<p>* Potentially sensitive question;</p> <p>*Reverse worded items (mental health inventory 5 + stress question)</p>
Allowances	The following questions are about allowances. Did (\$1: you \$2: your household) receive any of the following allowances from	<ol style="list-style-type: none"> 1. Health care allowance 2. Rent allowance 3. Child care * 4. Child budget * 	3 options: Yes 2022, Yes 2021, No, Do not know	* Due to technical reasons, for the Carousel and Accordion experimental group this question was contrary to the other grids

Question	Question stem	Items	Answer options	Remark
	the taks authority in 2022 and/or in 2021? >>You can choose multiple answers.<<			implemented as a stem fixed grid * Questions on Child care and Child benefit only asked if children in the household.
SurvAt_L (long only)	Finally, we would like to ask you some questions about participating in surveys in general and your experiences with this questionnaire. To what extent do you agree or disagree with the following statements?	<ol style="list-style-type: none"> 1. I enjoy filling out questionnaires. 2. I enjoy sharing my opinion or experience in a questionnaire. 3. Questionnaire research in itself is interesting. 4. Questionnaire research is important for society 5. Questionnaire research can provide a lot of useful information. 6. Filling out questionnaires for research is a waste of time. 7. I get invited to participate in questionnaire research too often. 8. Questionnaire research is an invasion of your privacy. 9. It is exhausting to answer many questions in a questionnaire survey. 	7 point bipolar, fully labeled, completely agree-completely disagree	*Reverse worded items; * Item 2 was adapted from the original scale. The original item was "I really enjoy being interviewed for a survey" was considered not relevant in the context of a web survey.

Question	Question stem	Items	Answer options	Remark
SurvAt_S (short only)		<ol style="list-style-type: none"> 1. I enjoy filling out questionnaires 2. Questionnaire research is important for society 3. I get invited to participate in questionnaire research too often. 		
Evaluation	How satisfied or dissatisfied are you with this survey in terms of:	<ol style="list-style-type: none"> 1. The invitation letter? 2. Logging into the questionnaire? 3. The clarity of the questions? 4. How easy it is to complete the questionnaire? 5. The design of the questionnaire (how the questionnaire looks)? 6. The length of the questionnaire? 7. The topics that were covered? 	5 point bipolar, fully labeled, very satisfied - very dissatisfied (depending on experimental group text only, text and smiley or smiley only)	

Appendix B Operationalizations of dependent variables

Analysis nr	Name ⁷	Operationalization ⁸	Interpretation
1	Login	Dummy indicating if sample unit logged in in the questionnaire.	Desirable
2	Breakoff	Dummy indicating if sample unit logged in in the questionnaire but did not complete.	Undesirable
No answer		Percentage of items for which 'no answer' is given as response (NA)	Undesirable: high percentage of no answer indicates satisficing
3	NA_all (LN)	% NA 33 items (all)	
4	NA_lo (LN)	% NA 34 items (long only). All items in grid questions, 5 grids.	
5	NA_SatSoc	% NA 5 items with smiley options (all) 1 grid on satisfaction with society and politics	
6	NA_SatLife (LN)	% NA 6 items with smiley option (long only) 1 grid on satisfaction with life	
7	NAGR_all	% NA 24 items in 4 grid questions (all)	
Number of answers in multiple select questions		Total number of selected answers for items where multiple answers were possible.	Desirable: Higher number of selected answer options indicates optimizing
8	SUM_multiple	Sum of the number of selected answer options for the four multiple select questions that all respondents received (all).	
9	SUM_SocMed	Sum of the selected social media accounts (all).	

⁷ LN indicates that dependent variables were natural log-transformed because significant results on Levene's test for equality of variances.

⁸ All: items that were shown to all respondents; Long only: items that were only shown to the respondents with the long questionnaire.

Analysis nr	Name ⁷	Operationalization ⁸	Interpretation
		Analyzed separately because of icon condition in this question.	
Primacy		Percentage items for which the first response option is chosen (PR)	Undesirable: high percentage primacy indicates satisficing
10	PR_all (LN)	% PR 31 items (all) No PR possible for two open questions)	
11	PR_lo	% PR 34 items (long only) All items in grid questions, 5 grids.	
12	PR_SatSoc (LN)	% PR 5 items with smiley option (all) 1 grid on satisfaction with society and politics	
13	PR_SatLife	% PR 6 items with smiley option (long only) 1 grid on satisfaction with life	
14	PRGR_all (LN)	% PR 24 items in 4 grid questions (all)	
Recency		Percentage items for which the last (substantial) response option is chosen (RE)	Undesirable: high percentage recency indicates satisficing
15	RE_all (LN)	% RE 31 items (all) No PR possible for two open questions.	
16	RE_lo	% RE 34 items (long only) All items in grid questions, 5 grids.	
17	RE_SatSoc (LN)	% RE 5 items with smiley option (all). 1 grid on satisfaction with society and politics	
18	RE_SatLife	% RE 6 items with smiley option (long only). 1 grid on satisfaction with life.	
19	REGR_all (LN)	% RE 24 items in grid questions (all)	
Response order effects in two questions with variation in the direction of the response options			Undesirable: response should not depend on the direction of response options
20	NwsIntMean	Average interest in domestic, international and political news.	

Analysis nr	Name ⁷	Operationalization⁸	Interpretation
21	SocMed_FB	Dummy indicating respondent reports to have a Facebook account	
22	SocMed_LinkedIn	Dummy indicating respondent reports to have a LinkedIn account	
Midpoint reporting		Percentage items for which the middle response option is chosen (only for items with an uneven number of response options) (MP)	Undesirable: high percentage midpoint reporting indicates satisficing
23	MP_all	% MP 8 items in 2 grid questions (all)	
24	MP_lo (LN)	% MP 28 items in 4 grid questions (long only)	
25	MP_SatSoc (LN)	% MP 5 items with smiley option (all) 1 grid on satisfaction with society and politics	
26	MP_SatLife	% MP 6 items with smiley option (long only). 1 grid on satisfaction with life	
Straightlining		Dummy indicating that within a grid for all items the same answer is chosen (SL)	Undesirable: high levels of straightlining may indicate satisficing.
27	SL_NwsInt	SL in 3 items on interest in news (all – if indicated that they follow the news in previous questions); 5 point unipolar scale.	
28	SL_DevInt	SL in 5 items on device use (all); yes/no answer options	
29	SL_MobUse	SL in 6 items on purpose use of mobile phone (all– if indicated that they use internet on the phone in past month); yes/no answer options	
30	SL_SatLife	SL in 6 items on satisfaction with life (smiley option; long only); 5-point bipolar scale	
31	SL_SatSoc	SL in 5 items on satisfaction with society and politics	

Analysis nr	Name ⁷	Operationalization ⁸	Interpretation
		(smiley option; all); 5-point bipolar scale	
32	SL_Trust	SL in 10 items on Trust in institutions (all); 10-point unipolar scale	
33	SL_BRS	SL in Brief Resilience Scale 6 items (long only); 5 point bipolar scale	
34	SL_TIPI	SL in 10 items Personality questions (long only); 7 point bipolar scale	
35	SL_MH	SL in 6 items on Mental Health (long only); 6 point bipolar scale	
36	SLS_BRS ⁹	SL -strict definition in Brief Resilience Scale (long only)	
37	SLS_MH	SL -strict definition in Mental Health grid (long only)	
	Acquiescence	Percentage items in which the 'agree'-answers are chosen (AC)	Undesirable: high percentage indicates satisficing
38	AC-all	% AC in 3 items in grid question	
39	AC_lo	% AC in 22 items of 3 grid questions (long only)	
40	Nchar	Sum of the number characters of the responses at the two open questions on 'Biggest issues in the country' and 'What goes well in the country'.	Desirable: longer answers indicate optimizing
41	Durاتم_co (LN)	Duration in minutes between time first login and time last login, only for questionnaires completed on the same day, cut off at mean + 2 x SD.	Multi interpretable: a longer duration can indicate optimizing as well as issues with completing the questionnaire
	Substantial results		Undesirable to have differences that depend on the design.
42	MeanSatSoc	Mean satisfaction with society; 5 items ,5-point scale, bipolar: very satisfied	

⁹ The strict definition excludes cases with 'no answer' responses and is applied solely to two scales containing reversed items. Strict straightlining indicates inconsistent response behavior. In contrast, for other grids, straightlining may represent a valid response pattern rather than satisficing.

Analysis nr	Name ⁷	Operationalization ⁸	Interpretation
		– very unsatisfied. (all) smiley option.	
43	MeanSatLife	Mean satisfaction with life; 6 items, 5-point scale, bipolar: very satisfied – very unsatisfied. (long only) smiley option.	
44	MeanTrtInst	Mean trust in institutions; 11 items, 10-point scale, unipolar: no trust at all – all trust (all).	
45	BRS	Score on Brief Resilience Scale; 6 items, 5-point scale, bipolar: completely agree – completely disagree (long only).	
46	OPEN	Mean of openness TIPI subscale (long only).	
47	CONS	Mean of conscientiousness TIPI subscale (long only).	
48	EXTRA	Mean extraversion TIPI subscale (long only).	
49	AGREE	Mean of agreeableness TIPI subscale (long only).	
50	NEUR	Mean of neuroticism TIPI subscale (long only).	
51	MHI	Score on Mental Health; 5 items, 6-point-scale, bipolar continuously – never (long only).	
	Mismatch with register data		Undesirable: mismatch can indicate input errors
52	Mismatch_house	Dummy indicating reported housing type does not match register.	
53	Mismatch_HCAI	Dummy indicating that report on whether healthcare allowance was received does not match register info.	
54	Mismatch_RentAll	Dummy indicating that report on whether rental allowance was received does not match register info.	

Analysis nr	Name ⁷	Operationalization⁸	Interpretation
55	Mismatch_ChildCareA II	Dummy indicating if that report on whether childcare allowance was received does not match register info.	
56	Mismatch_ChildBudg	Dummy indicating if that report on whether child-related budget was received does not match register info.	
	Respondent satisfaction		Desirable to score high on satisfaction items
57	MeanSatQst	Mean of 5 satisfaction-items concerning various aspects of the research (the 5 items detailed below)	
58	Sat_Letter	Score at satisfaction item on letter.	
59	Sat_Login	Score at satisfaction item on process of logging in.	
60	Sat_Clear	Score at satisfaction item on clarity of the questions.	
61	Sat_Easy	Score at satisfaction item on ease of completing the questionnaire.	
62	Sat_Layout	Score at satisfaction item on layout of the questionnaire.	
63	Recruit	Dummy indicating respondent answered yes to recruitment question if they may be contacted again for similar research.	Desirable to receive positive response at recruitment item

Appendix C Response by treatment combination

Response by treatment combination; 5.a: no encouragement of use of speech, 5.b: encouragement of use of speech. See Section 3 for more details on the treatment combinations.

			Regular login screen				Experimental login screen			
			Regular qnr. layout		Exp. qnr. layout		Regular qnr. layout		Exp. qnr. layout	
Length	Grid Des.	Graph. Symb.	5.b	5.a	5.b	5.a	5.b	5.a	5.b	5.a
Short	3.a	4.a	128	87	-	-	116	84	-	-
		4.b	-	-	-	-	-	-	-	-
		4.c	-	-	-	-	-	-	-	-
	3.b	4.a	128	90	35	33	127	83	36	29
		4.b	-	-	52	32	-	-	39	25
		4.c	-	-	41	29	-	-	41	17
	3.d	4.a	-	-	45	31	-	-	38	22
		4.b	-	-	39	32	-	-	38	27
		4.c	-	-	45	27	-	-	41	26
	3.c	4.a	-	-	41	27	-	-	38	31
		4.b	-	-	53	29	-	-	36	25
		4.c	-	-	41	27	-	-	34	28
Long	3.a	4.a	109	73	-	-	108	87	-	-
		4.b	-	-	-	-	-	-	-	-
		4.c	-	-	-	-	-	-	-	-
	3.b	4.a	111	95	34	24	131	92	45	26
		4.b	-	-	38	27	-	-	42	33
		4.c	-	-	34	27	-	-	40	25
	3.d	4.a	-	-	43	20	-	-	50	20
		4.b	-	-	40	32	-	-	35	25
		4.c	-	-	42	23	-	-	41	27
	3.c	4.a	-	-	42	36	-	-	34	30
		4.b	-	-	38	28	-	-	41	28
		4.c	-	-	43	26	-	-	41	29

Appendix D Descriptive statistics dependent variables

Variable	N	Minimum	Maximum	Mean	SD
Login	12060	0.00	1.00	0.38	0.49
Breakoff	4586	0.00	1.00	0.11	0.31
NA_all	4088	0.00	96.97	5.28	9.72
NA_lo	2015	0.00	100.00	2.01	9.84
NA_SatSoc	4088	0.00	100.00	5.96	16.82
NA_SatLife	2015	0.00	100.00	1.10	8.00
NAGR_all	4088	0.00	100.00	4.96	11.50
SUM_multiple	4088	0.00	20.00	8.76	2.70
SUM_SocMed	4088	0.00	7.00	2.77	1.68
PR_all	4088	0.00	65.63	23.03	7.21
PR_lo	2015	0.00	55.88	9.60	8.98
PR_SatSoc	4088	0.00	100.00	2.35	9.97
PR_SatLife	2015	0.00	100.00	24.51	28.05
PRGR_all	4088	0.00	70.83	16.48	7.86
RE_all	4088	0.00	56.41	9.04	5.00
RE_lo	2015	0.00	38.24	6.33	6.69
RE_SatSoc	4088	0.00	100.00	7.16	17.23
RE_SatLife	2015	0.00	100.00	1.32	6.04
REGR_all	4088	0.00	58.33	10.78	7.22
NwsIntMean	3856	1.00	5.00	3.13	0.81
SocMed_FB	4088	0.00	1.00	0.61	0.49
SocMed_LinkedIn	4088	0.00	1.00	0.44	0.50
MP_all	4088	0.00	100.00	30.01	20.30
MP_lo	2015	0.00	77.78	20.49	11.93
MP_SatSoc	4088	0.00	100.00	36.96	27.69
MP_SatLife	2015	0.00	100.00	16.87	19.75
SL_NwsInt	3896	0.00	1.00	0.29	0.45
SL_DevInt	4088	0.00	1.00	0.12	0.32
MobUse_SL	3912	0.00	1.00	0.27	0.45
SL_SatLife	2015	0.00	1.00	0.10	0.30
SL_SatSoc	4088	0.00	1.00	0.11	0.32
SL_Trust	4088	0.00	1.00	0.02	0.15
SL_BRS	2015	0.00	1.00	0.05	0.23
SL_TIPI	2015	0.00	1.00	0.02	0.13
SL_MH	2015	0.00	1.00	0.03	0.17
SLS_BRS	2015	0.00	1.00	0.01	0.09
SLS_MH	2015	0.00	1.00	0.01	0.11
AC_all	4088	0.00	100.00	53.30	24.90
AL_lo	2015	0.00	100.00	48.32	12.09

Nchar	4088	0.00	7991.00	175.31	263.89
Durat_co	3959	3.35	80.72	20.95	14.07
MeanSatSoc	3934	1.00	5.00	2.97	0.70
MeanSatLife	2002	1.00	5.00	3.92	0.56
MeanTrtInst	3911	1.00	10.00	6.05	1.35
BRS	1962	1.00	3.00	1.81	0.51
OPEN	1966	1.00	7.00	4.96	1.06
CONS	1974	1.00	7.00	5.61	0.97
EXTRA	1966	1.00	7.00	4.47	1.35
AGREE	1969	1.00	7.00	4.26	0.79
NEUR	1976	1.00	7.00	5.05	1.25
MHI	1958	4.00	100.00	70.59	15.99
Mismatch_House	3949	0.00	1.00	0.22	0.42
Mismatch_RentAll	3099	0.00	1.00	0.03	0.17
Mismatch_ChildcareAll	3099	0.00	1.00	0.05	0.21
Mismatch_ChildBudg	3099	0.00	1.00	0.05	0.23
MeanSatQst	4032	1.00	5.00	4.04	0.52
Sat_letter	3997	1.00	5.00	3.95	0.70
Sat_login	4024	1.00	5.00	4.18	0.71
Sat_clear	4033	1.00	5.00	4.19	0.63
Sat_easy	4033	1.00	5.00	4.24	0.64
Sat_layout	4017	1.00	5.00	4.17	0.65
Recruit	4088	0.00	1.00	0.72	0.45

Colophon

Publisher

Centraal Bureau voor de Statistiek
Henri Faasdreef 312, 2492 JP Den Haag
www.cbs.nl

Prepress

Statistics Netherlands, CCN Creation and visualisation

Design

Edenspiekermann

Information

Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contactform: www.cbs.nl/information

© Statistics Netherlands, The Hague/Heerlen/Bonaire 2025.

Reproduction is permitted, provided Statistics Netherlands is quoted as the source.