Article

Social Science Computer Review
1-13
© The Author(s) 2015
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0894439315588736
ssc.sagepub.com
SAGE

Thinking Inside the Box: Visual Design of the Response Box Affects Creative Divergent Thinking in an Online Survey

Alicia Hofelich Mohr[1], Andrew Sell[1], and Thomas Lindsay[1]

Abstract

While the visual design of a question has been shown to influence responses in survey research, it is less understood how these effects extend to assessment-based questions that attempt to measure *how*, rather than just *what*, a respondent thinks. For example, in a divergent thinking task, the number and elaboration of responses, not just how original they are, contribute to the assessment of creativity. Using the Alternative Uses Task in an online survey, we demonstrated that scores on fluency, elaboration, and originality, core constructs of participants' assessed creative ability, were systematically influenced by the visual design of the response boxes. The extent to which participants were susceptible to these effects varied with individual differences in trait conscientiousness, as several of these effects were seen in participants with high, but not low, conscientiousness. Overall, our results are consistent with previous survey methodology findings, extend them to the domain of creativity research, and call for increased awareness and transparency of visual design decisions across research fields.

Keywords

survey research, visual design, answer box size, creativity, divergent thinking

The effects of modifying the visual layout or design of survey questions have received considerable attention in the field of survey methodology. Researchers have found that changing aspects of the visual design of a question can have significant effects on participants' answers to the same question, impacting responses on topics as varied as past behavior, attitudes, or opinions (e.g., Christian, Dillman, & Smyth, 2007; Couper, Conrad, & Tourangeau, 2007; Dillman, Smyth, & Christian, 2009). Visual aspects of a question are thought to influence how participants cognitively process the question, including how they decide to express their responses (Groves et al., 2009; Tourangeau, Couper, & Conrad, 2004). The size of a response text box, for example, can cue participants to the researcher's expectations about the type, structure, or length of the desired response.

[1] University of Minnesota, Minneapolis, MN, USA

Corresponding Author:
Alicia Hofelich Mohr, College of Liberal Arts Research Support Services, 257 19th Ave. S, Minneapolis, MN 55455, USA.
Email: hofelich@umn.edu

However, the extent to which these effects extend to questions outside of traditional surveys is less understood. For example, while such effects are well studied in the context of survey methodology and opinion measurement, less is known about how they translate to questions designed to assess underlying psychological constructs latent in the response. In these types of questions (which we will refer to as *construct based*), researchers often use multiple aspects of the response to assess not only *what* participants are thinking but also *how* they are thinking. For example, in a popular creativity task, the measurement of a participant's divergent thinking ability is based not only on the originality of the response (content) but also on the number of responses given and how elaborative they are (Guilford, 1967). Similarly, in writing therapy studies that assess constructs such as health, the topic of the written response is often less informative than the linguistic style of the response (e.g., Campbell & Pennebaker, 2003). In construct-based questions such as these, the detail or elaboration of the response, rather than just the semantic content, directly affects the measurement of the construct. Effects of visual design are known to influence the elaboration and detail of responses in survey questions. Hence, it is important to test whether these effects translate to construct-based questions, as visual design may affect the actual measurement of an individual's trait or ability. Visual design decisions made by researchers could be confounding factors in these measurements, especially when comparing results across different studies. The present study examines this question in the context of a construct-based creativity task.

Visual Design Effects

Many researchers have found participants provide longer responses when given a large, versus small, text box (Christian & Dillman, 2004; Israel, 2010; Smyth, Dillman, Christian, & McBride, 2009; Stern, Dillman, & Smyth, 2007) and when given more, versus fewer, lines (Spörrle, Gerber-Braun, & Försterling, 2007). The number of answers provided by participants is also influenced by the size or number of text boxes, with participants listing more items when given several smaller text boxes versus one large text box (e.g., Keusch, 2014) or when given a greater number of text boxes of the same size (e.g., Smyth, Dillman, & Christian, 2007). However, increasing the size or number of text boxes can also increase the nonresponse rate for that item (Smyth et al., 2007; Zuell, Menold, & Korber, 2015).

With the popularity of online survey tools, researchers across a variety of fields are moving traditionally paper-based tasks to an online environment. Although visual design has been studied both online and on paper, the impact of visual design manipulations within these modes can differ. For example, when presented with longer text boxes on paper, but not online, participants answered frequency questions with more elaborative alphanumeric responses (Fuchs, 2009). However, in general, visual design manipulations appear to produce consistent findings across modes. For example, the number of list-style text boxes influences the number of responses participants provide online (Keusch, 2014), just as the number of lines does in paper studies (Spörrle et al., 2007). Studies in both modes have also found that visual design effects may be qualified by participant characteristics, such as whether they were early or late responders (Smyth et al., 2009 [online]) or had certain demographic attributes (Stern et al., 2007 [paper]).

In addition to potential differences the same visual design may produce in paper versus online surveys, moving traditional paper forms online can also present opportunities for researchers to make different decisions about the visual design. In the same modality, many widely used measures or tasks have standardized forms, producing consistent visual layouts across studies. If visual design is changed and not reported, it may impact researchers' ability to compare results across studies, affecting perceived replicability and reliability of the measures. Thus, it is important to study the potential consequences of different visual design decisions in various tasks or questions outside of traditional survey research, where these effects are less studied. In psychological research, for example, much attention has been paid to the characteristics and response quality of online participant pools (such as Amazon Mechanical Turk [Mturk]; Buhrmester, Kwang, & Gosling, 2011), but less attention has been paid to the consequences of visual design decisions made in the transition from paper to online.

Construct-Based Questions: Creativity and Divergent Thinking

One example of a construct-based task is the Alternate Uses Task (AUT; Guilford, 1967), a popular assessment used to measure an aspect of creativity called divergent thinking. In this task, participants are asked to list as many different uses as they can for a common object (such as a brick or bucket) in a set amount of time. Traditionally, this task is given on paper, with space below the instructions for participants to provide their responses. Responses are typically scored on the number of uses listed (fluency), the number of different categories included in their list (flexibility), the uniqueness of uses (originality), and how much detail is given about each use (elaboration), resulting in a composite or several scores that are taken to reflect divergent thinking ability.

Although typically given on paper, researchers giving the task online may choose one of several reasonable options for formatting the AUT. For example, it may be appropriate to give participants one large essay box in which to write their responses or to display several smaller boxes for participants to write each item on their list. If the findings from survey methodology studies extend to this task, then this decision, as well as choices about the number or size of the boxes, will influence the divergent thinking scores of respondents. However, these decisions are not typically considered in the creativity literature. Recent papers that use the AUT online have given one large box (Griffin & Jacob, 2013), several small boxes (Lewis, Dontcheva, & Gerber, 2011), or did not describe the format at all (Wiseman, Watt, Gilhooly, & Georgiou, 2011).

Given the rising popularity of online participant pools and the increased expectation for reproducible results, it is important to determine whether these visual design choices have an effect on research outcomes. If the visual design does contribute systematic variance to divergent thinking scores, it could affect the perceived reliability and replicability of these scores across studies and groups. This would demonstrate a need for researchers to clearly state and explain the visual design decisions made in their methodology. If we do not find evidence that visual design manipulations affect responses in this task, it would suggest that question content and research objectives may mitigate the effects of visual design choices, possibly demonstrating a boundary condition for visual design manipulations.

Present Study

The present study investigates whether manipulating the type of response box (one large *unsegmented* box vs. several small *segmented* boxes) and number/size of response boxes (5, 10, or 15 lines/boxes) influences the fluency, flexibility, elaboration, and originality of participant responses in the AUT. Based on the findings from survey research, we predict that participants who receive more/larger response boxes will provide more responses (i.e., higher fluency scores) than those who receive fewer/smaller boxes. We also predict that participants who receive segmented boxes will provide more responses (i.e., higher fluency scores), write less detailed responses (i.e., lower elaboration scores), and list items from more categories (i.e., higher flexibility scores) than those who receive unsegmented boxes, controlling for the number/size of the visible textbox(es). We also examined whether the number/size of boxes influenced the number of nonresponses on the AUT or the number of participants that ended the task early. While we did not have any specific hypotheses about how our visual design manipulations would affect originality scores, we did exploratory analyses to see whether these scores differed by type or number/size of the response boxes.

We also hypothesized that individual differences in personality, specifically conscientiousness, would affect susceptibility to our visual design manipulations, such that participants who are high in conscientiousness will be more susceptible to the visual design effects.

Although a previous study found that visual manipulations most affected late responders, who were presumably less motivated to complete the study (Smyth et al., 2009), we predicted that the nature of the AUT as an assessment-based task would produce high engagement overall, and that differences would be more likely to appear among those who pay more attention to the layout of the question. For example, if highly conscientious participants are concerned with performing the task well or meeting the researchers' expectations, they may be more likely to take cues from the visual display of the response box and thus be more susceptible to these effects than participants low in conscientiousness. To test this, participants were given a personality questionnaire after completing the AUT.

In addition to the hypotheses described earlier, which have the greatest impact on meta-analysis or examination of results across studies, we were also interested in potential impacts our manipulations may have on individual studies. For example, researchers may use creativity scores such as those produced by the AUT to distinguish or differentiate among participants, or to correlate individual differences in creativity with other measures. In these cases, it may be useful to know whether different visual layouts of the same task can produce greater differentiation among individuals. One potential measure of such differentiation is the variance. Designs that produce higher variance across individuals could be more useful in these situations. To this end, we also tested whether variance in fluency, elaboration, flexibility, or originality scores differed by type or number/size of boxes.

Method

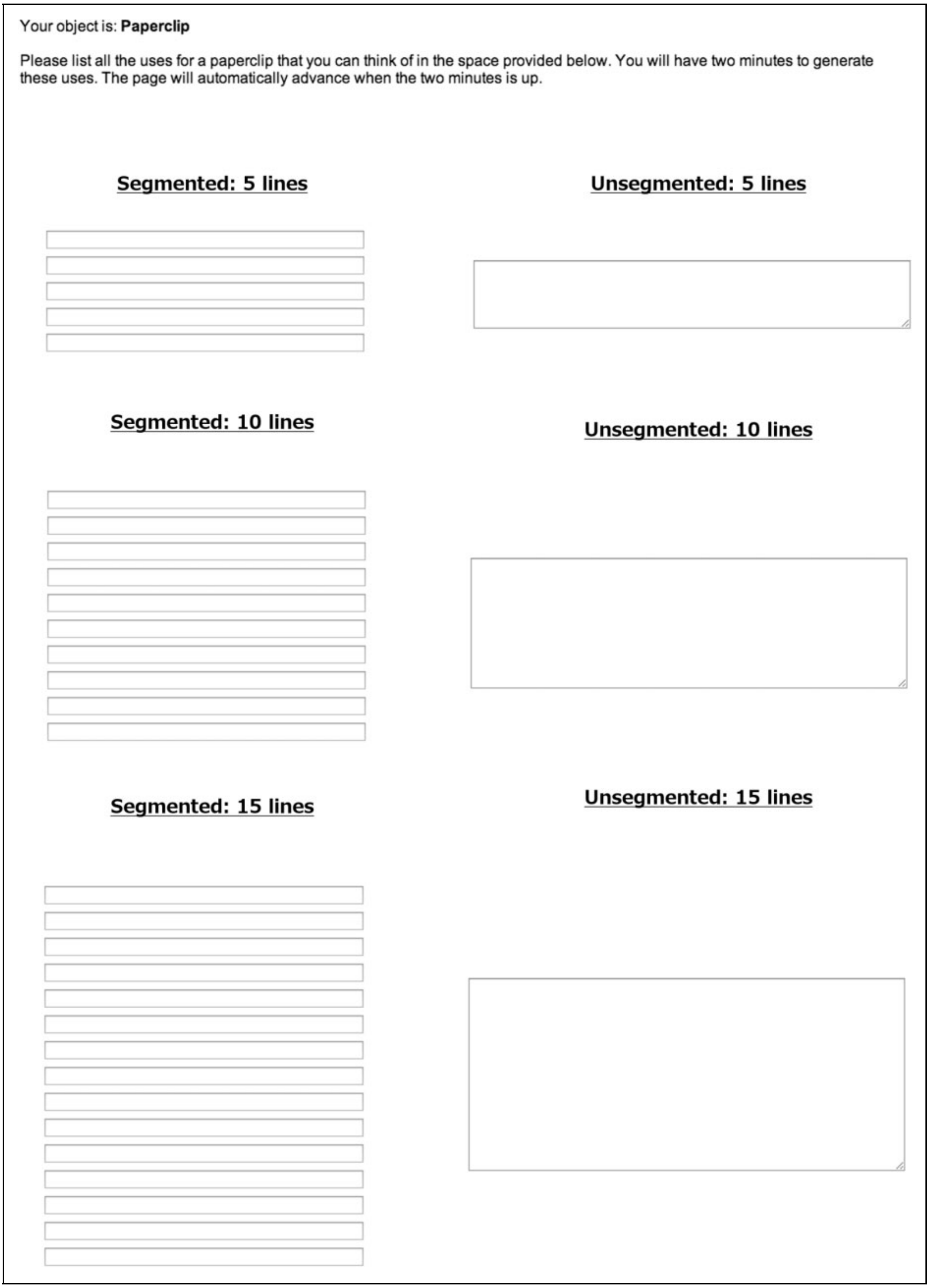
Participants and Procedure

Participants were recruited via Mturk.[1] A total of 619 participants (231 men, 382 women, 6 other/nonresponse, age mean = 35.07, range 16–82) completed the study for USD0.25 compensation. Recruitment was restricted to participants located in the United States with a Human Intelligence Task (HIT) approval rate greater than or equal to 90%. One participant was removed for invalid responses on the creativity task (e.g., nothing and none), one was removed for reporting an age less than 18, and 21 participants were removed for abnormally long response times (at least 20 s over the time at which the page would have automatically advanced[2]). This left a total of 596 participants in the analysis (223 men, 369 women, and 4 other/nonresponse).

After consenting, participants were randomly assigned to one of six experimental conditions that determined the type of response field (a single unsegmented box, or multiple segmented boxes) and number of boxes/lines (5, 10, or 15) in the response field. Additionally, the item (paper clip and brick) seen in the divergent thinking task was equally and randomly distributed across participants. Participants were given 2 min to complete the task (with a pop-up message appearing with 30 s remaining) and then were automatically advanced to a personality inventory and demographic questions. The study was conducted online using an in-house survey tool.

Measures and Design

Guilford's AUT. In the AUT, participants were asked to provide as many uses for an item (paperclip or brick) they could think of in 2 min. All participants saw the same instructions (with the exception of item—half saw brick and half saw paper clip), but the response box formats were manipulated according to a 2 (type: *segmented/unsegmented*) × 3 (size: *5, 10, 15 boxes/lines*) design, with participants assigned to one of the six possible cells (see Figure 1).



Your object is: **Paperclip**

Please list all the uses for a paperclip that you can think of in the space provided below. You will have two minutes to generate these uses. The page will automatically advance when the two minutes is up.

Figure 1. Example Alternate Uses Task instructions and response box displays for each of the six experimental conditions.

To avoid imposing an artificial ceiling on the number of responses a participant could give, another box appeared when participants began typing in the last visible segmented text box. This allowed participants to report as many uses as they wanted, while still manipulating the initial number of visible boxes. Similarly, the unsegmented boxes accepted characters beyond the perceived size, with a scroll bar appearing when participants typed outside the viewable field. All participants were instructed to generate as many responses as possible. Participants were not allowed to go back to previous pages in the study.

Big Five Inventory (BFI). To explore whether individual differences in personality, specifically in conscientiousness, are associated with differences in susceptibility to the effects of visual display, participants completed the 44-item BFI (John & Srivastava, 1999) after the AUT. Additionally, we included demographic questions related to age, sex, and level of education.

Coding of Responses

Responses on the AUT were scored for fluency, flexibility, elaboration, and originality. Fluency was measured by the number of complete responses made by each participant. Incomplete responses were excluded from analysis. Due to limited availability, the greatest number of judges were assigned to ratings that were more subjective (e.g., originality) and fewer were assigned to ratings that were more objective (e.g., elaboration). Flexibility was rated by three judges as the number of categories present within each participant's set of responses. The judges' scores were averaged to form a single flexibility score per participant. Elaboration scores were given to each response in two ways; first, the word count of each response was taken, and second, each response was given a 0, 1, or 2 score by two judges based on the level of detail included in the response (see Hommel, Colzato, Fischer, & Christoffels, 2011). Originality scores from 1 (*least original*) to 5 (*most original*) were given to each response by four judges based on how uncommon, remote, and clever the response was (procedure described in Silvia et al., 2008). All judges were blind to participant condition. After checking interrater reliability, the judges' ratings for elaboration and originality were averaged for each response. Finally, all elaboration and originality scores were averaged across the responses given by each participant, resulting in a single score per participant. The BFI was scored according to standard practices (John & Srivastava, 1999), resulting in five personality scores—conscientiousness, openness, extraversion, agreeableness, and neuroticism. Given our hypotheses, we only examined the conscientiousness scores.

Data Analysis

Separate analysis of variance (ANOVA) models using type II sums of squares with variance heteroskedastic correction (Long & Ervin, 2000) or Welch's *t*-tests were run to test the four main hypotheses, depending on the variables of interest. All preprocessing and analysis were done in R (R Core Team, 2014). Because we were interested in assessing the linear effects of increasing the number/size of text boxes, this variable was treated numerically, rather than categorically. Exploratory ANOVAs were used to assess the effects of box type and number/size on creativity. We used a median split to produce high- and low-conscientiousness groups, and added this variable to retest significant models to explore whether conscientiousness interacted with any significant effects of visual design. To test whether variance differed by visual design, Bartlett or Levene tests were used for variables that were less or more skewed, respectively. An α of .05 was used for all tests. All data and code underlying this article are available online (Hofelich Mohr, Sell, & Lindsay, 2015)

Table 1. Means (Standard Deviation) for Scores by Response Box Type and Number/Size.

	Segmented			Unsegmented		
	5	10	15	5	10	15
Fluency	5.55 (2.50)	6.43 (2.64)	6.36 (3.10)	5.84 (2.40)	5.97 (2.62)	6.10 (2.74)
Average word count	3.63 (2.16)	3.56 (1.86)	3.22 (1.76)	4.28 (2.32)	4.23 (2.19)	4.26 (2.08)
Elaboration	0.11 (0.21)	0.09 (0.14)	0.08 (0.16)	0.13 (0.21)	0.12 (0.18)	0.16 (0.18)
Creativity	2.08 (0.42)	1.93 (0.35)	1.96 (0.43)	1.96 (0.48)	1.98 (0.40)	2.06 (0.45)
Flexibility	4.64 (2.01)	5.09 (2.05)	5.07 (2.45)	4.67 (2.04)	4.80 (1.86)	5.03 (2.40)

Results

Does a Greater Number/Size of Response Boxes Produce Higher Fluency Scores?

Supporting our hypothesis, we found that fluency scores increased linearly as the boxes that were initially visible increased in number/size, $F(1, 594) = 3.90$, $p = .048$, estimate $\beta = .27$. Examining the means, there was a larger difference between 5 and 10 boxes/lines than between 10 and 15 boxes/lines (see Table 1). This was confirmed with follow-up t-tests, as participants who received five boxes/lines had lower fluency scores than those who saw 10 or 15, $ts(387.94) > 1.95$, $ps < .05$, while there was no difference between 10 and 15 boxes/lines, $t(392.41) = -0.11$, $p = .91$. This effect did not differ by type of box, interaction with segmented versus unsegmented, $F(1, 592) = 1.03$, $p = .31$.

Do Segmented Boxes Produce Higher Fluency Scores?

When controlling for the number/size of the boxes initially visible, there was no difference in fluency scores between segmented and unsegmented boxes, $F(1, 592) = 0.26$, $p = .61$ (Table 1). Although the number/size of boxes influenced the number of responses provided by participants (as shown earlier), whether they saw a single box or multiple smaller boxes did not make a difference.

Do Segmented Boxes Produce Lower Elaboration Scores?

Elaboration ratings between the two judges agreed on 92% of the cases and correlated strongly with the word count measure of elaboration, $r(594) = .74$, $p < .001$. Supporting our hypothesis, both measures of elaboration were lower for participants who saw segmented boxes than those who saw an unsegmented box, ratings: $t(590.09)= -3.07$, $p = .002$; word count: $t(585.82) = -4.69$, $p < .001$ (Table 1). This effect did not differ by the number of boxes/lines seen by the participant for either measure of elaboration, $Fs(1, 592) < 2.01$, $ps > .16$.

Do Segmented Boxes Produce Higher Flexibility Scores?

The interrater reliability for the three judges' flexibility ratings was high, Cronbach's $\alpha = .97$. Because the number of categories was constrained by the number of responses given, we controlled for participants' fluency scores when we tested the effects of box type on flexibility. Flexibility scores did not differ between the segmented and unsegmented conditions, $F(1, 593) = 0.0002$, $p = .99$.

Does Visual Design Affect Item Nonresponse or Early Termination Rates?

Among participants who completed the study ($n = 619$), only one did not respond to the AUT task. Therefore, to examine whether the number/size or type of boxes were associated with differences in survey noncompletion rates, we used the data from all participants who consented to participate, including those who did not complete the study (total $n = 759$). Chi-square tests revealed no difference in survey termination rates for those participants who had made it to the AUT between segmented ($n = 62$) and unsegmented ($n = 53$) conditions, $\chi^2(1, N = 759) = 0.63$, $p = .43$, nor among the number/size of the boxes conditions (5, $n = 33$; 10, $n = 43$; and 15, $n = 39$), $\chi^2(2, N = 759) = 1.52$, $p = .47$.

Does Visual Design Influence Originality Scores?

The interrater reliability for the judges' originality scores was .73 (Cronbach's α). Although neither the number/size of boxes nor the type of response box were related to originality scores, Fs(1,592) < 0.15, ps > .7, a significant interaction was found between the two, $F(1, 592) = 6.24$, $p = .01$. Follow-up tests revealed that for participants who saw segmented boxes, originality scores decreased when more boxes were visible, $F(1, 295) = 3.74$, $p = .05$, estimate $\beta = -.06$. This was not the case for participants who saw unsegmented boxes, $F(1, 297) = 2.61$, $p = .11$, estimate $\beta = .05$, as originality tended to increase with the size of the box in this condition, although this did not reach statistical significance.

Are Highly Conscientious People More Susceptible to Visual Design Manipulations?

The reliability for the conscientiousness items on the BFI was high in our sample and similar to published findings (John & Srivastava, 1999), Cronbach's $\alpha = .81$, overall $M = 3.71$, $SD = 0.66$. Fifty-eight participants did not provide enough data to calculate a conscientiousness score and were excluded from further analysis. A median split created high-conscientiousness ($M = 4.26$, $SD = 0.38$) and low-conscientiousness ($M = 3.12$, $SD = 0.36$) groups (those with scores equal to the median were excluded, $n = 30$), which were used to determine whether any of our significant findings were qualified by this personality trait.

When included as a term in each model, conscientiousness did not significantly interact with the visual design effects found for fluency, $F(1, 504) = 0.001$, $p = .97$, elaboration, Fs(1, 504) < 2.77, ps > .10, or originality, $F(1, 500) = 0.20$ $p = .65$. However, because we had an a priori hypothesis that highly conscientious people would be more susceptible to visual design manipulations, we tested each of the significant findings separately for participants low and high in conscientiousness.

When examined separately, highly conscientious participants elaborated less in segmented compared to unsegmented conditions, ratings: $t(261.83) = -2.51$, $p = .01$; word count: $t(260.66) = -4.48$, $p < .001$ (see Table 2). Participants low in conscientiousness did not show this effect for elaboration ratings, $t(241.93) = -0.91$, $p = .36$, and showed a trend effect for word count, $t(240.14) = -1.90$, $p = .06$. Similarly, in highly conscientious participants, the interaction between box type and number/size of boxes significantly influenced originality scores, $F(1, 260) = 4.04$, $p = .045$, with increasing number/size of boxes leading to reduced originality scores in the segmented condition ($\beta = -.07$) and increased originality in the unsegmented condition ($\beta = .07$). This was not seen for less conscientious participants, $F(1, 240) = 1.55$, $p = .21$. Although the relationship between fluency and number/size of boxes was positive in both groups, high: $\beta = .22$ and low: $\beta = .23$, this effect did not reach significance in either group alone, Fs(1, 242) < 1.23, ps > .27.

Table 2. Means (Standard Deviation) for Participants High and Low in Contentiousness.

		Segmented			Unsegmented		
		5	10	15	5	10	15
High	Fluency	5.56 (1.96)	6.30 (2.72)	6.54 (3.61)	6.44 (2.18)	5.94 (2.75)	6.40 (3.16)
(n = 264)	Average word count	3.68 (2.03)	3.44 (1.60)	3.27 (1.73)	4.51 (1.69)	4.69 (2.38)	4.35 (2.15)
	Elaboration	0.11 (0.23)	0.07 (0.13)	0.11 (0.16)	0.13 (0.18)	0.15 (0.21)	0.18 (0.20)
	Creativity	2.10 (0.43)	1.87 (0.35)	1.97 (0.43)	1.94 (0.42)	2.04 (0.45)	2.07 (0.42)
	Flexibility	4.73 (1.79)	4.96 (2.10)	5.16 (2.71)	5.10 (1.98)	4.89 (1.96)	5.17 (2.65)
Low	Fluency	5.93 (2.68)	6.42 (2.41)	6.44 (2.86)	5.53 (2.50)	5.92 (2.48)	5.90 (2.62)
(n = 244)	Average word count	3.62 (2.50)	3.61 (1.41)	3.13 (1.80)	4.08 (2.46)	3.68 (1.66)	3.96 (1.84)
	Elaboration	0.13 (0.21)	0.09 (0.13)	0.06 (0.18)	0.11 (0.20)	0.10 (0.14)	0.13 (0.15)
	Creativity	2.08 (0.38)	2.03 (0.30)	1.97 (0.45)	1.95 (0.52)	1.89 (0.34)	2.02 (0.50)
	Flexibility	4.81 (2.10)	5.25 (1.87)	5.19 (2.37)	4.39 (2.11)	4.59 (1.76)	4.91 (2.42)

Does Variance Differ by Visual Design for Each Score?

In general, variance in each score differed by our manipulations in the same directions as the means reported earlier. Variance of fluency scores increased with the number/size of boxes, Bartlett's $\chi^2(2) = 6.33$, $p = .04$, but did not differ by segmentation, Bartlett's $\chi^2(1) = 1.71$, $p = .19$. The variance of elaboration scores was higher when participants saw unsegmented (rating $s^2 = 0.36$; word count $s^2 = 4.81$) compared to segmented boxes (rating $s^2 = 0.30$; word count $s^2 = 3.74$), Levene's test Fs(1, 594) > 7.15, ps < .008, but did not differ by the number/size of boxes, Fs(2, 593) < 1.29, ps > .28. Originality score variance differed by the number/size of boxes, Bartlett's $\chi^2(2) = 8.23$, $p = .02$, and tended to be higher in unsegmented ($s^2 = 0.20$) than in segmented conditions ($s^2 = 0.16$), $\chi^2(1) = 3.20$, $p = .07$. Variance in flexibility scores (controlling for fluency) did not differ by the number/size of boxes nor by segmentation, Fs(2, 593) $\leq$ 0.74, ps > .48.

Discussion

The present study extends findings from survey methodology and opinion measurement, demonstrating that the visual design of the response box can influence how participants respond to a construct-based question in a creativity task, which ultimately affects the measurement of their creative ability. Our findings have important implications for creativity and other research that assesses latent constructs using multiple aspects of a response, as we demonstrate that visual design choices, which are not often reported in the literature, affect these assessments.

Specifically, participants' fluency, elaboration, and originality scores on a divergent thinking task were influenced by variations in the size, number, and type of response box. Participants provided more uses for an item as the number or size of the boxes increased and provided more elaborative responses when the boxes were unsegmented, rather than segmented. These results are consistent with previous research manipulating the number of lines in a response box (Spörrle et al., 2007) and the type of response box (Smyth et al., 2007). Furthermore, we demonstrate that visual design influences the quality of responses in this creativity task, as measured by originality. Participants who saw more segmented boxes gave responses that were less original than those who saw fewer segmented boxes, while the opposite was true for those who saw larger versus smaller unsegmented boxes.

Our findings for fluency scores are not explained by any artificial ceiling effects our manipulation may have imposed, as participants in each of the segmented box conditions gave more responses than the number of initially visible boxes (number exceeding visible boxes in 5 line condition:

$n = 39$ (40.0%), 10 line condition: $n = 7$ (7.3%), and 15 line condition: $n = 1$ (1.0%). Additionally, although a technical issue caused the width of the segmented and unsegmented boxes to differ depending on participants' browsers, differences in width between the two conditions did not account for the elaboration effects we found between segmented and unsegmented types of boxes.[3] Contrary to our hypotheses, we did not find evidence that the type of box influenced the number of responses participants provided, nor that the type or number/size of response box(es) affected the number of categories participants included in their responses. Item nonresponse also did not appear to differ by visual design.

We found individual differences in the susceptibility to visual design effects, which varied by conscientiousness. For highly conscientious participants, visual design manipulations reliably influenced how elaborative and original their responses were, while this was not the case for participants low in conscientiousness. As one of the defining characteristics of conscientiousness is performing tasks thoroughly and efficiently, a potential explanation for this result is that highly conscientious participants were more motivated to perform well on the divergent thinking task and were more likely to use cues from the visual display of the response box to guide their responses. For example, segmented boxes may have served as a cue that list-style responses were preferred. When participants took this into account, they elaborated less than if given a larger unsegmented box. More research is needed to understand the ways personality and situational characteristics, such as conscientiousness and motivation, may differentially affect respondents' susceptibility to visual design.

As an effort to determine whether certain types of response box design would produce greater differentiation between individuals, we tested whether variance of each score differed by our manipulations. We found that variance tracked with the means for fluency and elaboration, fluency variance highest for larger/more boxes, and elaboration variance highest in unsegmented boxes. This may not be surprising given the count-like nature of many of these measures (which are often modeled as poisson distributed with a single parameter for mean and variance[4]). However, these results still have applicability; in a single study seeking a wide range of scores to differentiate respondents, our results suggest the best response box to use would be an unsegmented box 10–15 lines long.

Overall, our results suggest that decisions made about the visual design of a response box influence the number, length, and quality of the responses participants give in a creativity task. Unlike questions about opinions or experiences, these responses do not simply provide the researcher more information, they change how the underlying construct is assessed—in this case, modifying the fluency, elaboration, and originality of the output a participant gives during creative thinking. While previous research indicates that visual design serves to cue respondents about the type and length of response desired, our research extends this to suggest it also can affect the processes people engage in during creative thinking. Our results support the importance of transparency of visual design decisions in published studies, especially because many of these decisions are not typically reported.

Acknowledgement

The authors thank Yi Ying Sack and Kayla Tatro for assisting with data scoring and Lilian Li for help with piloting.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project was funded by the Associate Dean for Research and Graduate Programs in the College of Liberal Arts, University of Minnesota.

Notes

1. Amazon's Mechanical Turk is a crowdsourcing platform that has become popular for the online recruitment and testing of research participants. While this sample is more demographically diverse than college samples, the data quality has been shown to be comparable (Buhrmester et al., 2011; Casler, Bickel, & Hackett, 2013).
2. Participants could be on the page for more than 120 s if they did not click "OK" on the confirmation box that warned them that they had 30 s left to complete the task. The page would then advance 30 s later unless the pop-up window was still open, in which case the page would advance when "OK" was pressed.
3. Even though the underlying HTML code for the segmented versus unsegmented text boxes was set at the same size (size = 62), browsers rendered the segmented (list-style) and unsegmented (essay-style) text boxes differently. This difference was noticed after data collection was completed. Particularly noticeable was that the unsegmented box appeared wider than the segmented boxes, even though they were set to the same size. While this did not affect our manipulation of box number/size (as the height of the box in the unsegmented condition relative to the number of boxes in the segmented condition was the same), it is possible the variable width of the boxes may have influenced participant responses.

   Specifically, we were concerned that the differences in box width could account for the differences we found in elaboration scores between the two conditions, as participants were more elaborative in the unsegmented compared to segmented boxes (same direction as the width error). Previous research looking at response length has found text box width (described in the literature as textbox length for one-line list-style boxes) affects the length of participant responses on paper (Fuchs, 2009; Israel, 2010) but not online (Fuchs, 2009). Therefore, we determined it was important to test whether the unintended width differences between the segmented and unsegmented boxes were contributing to or explaining the differences in elaboration between conditions, rather than the "list-style" versus "essay style" nature of the boxes. To test this, we first gathered the browser and operating system (OS) version information from each respondent (which were collected by our survey software), and used Sauce Labs (https://saucelabs.com/) to test the survey in each of the most popular browser and OS combinations used by respondents (covering combinations used by 75% of the participants). Textbox width was measured in pixels using a screenshot crosshair with pixel readout (Mac OS). Widths were estimated for segmented and unsegmented conditions of each of the browser/OS combinations. These width estimates were then combined with the score data for each participant and used as a covariate in the analysis of variance, examining elaboration scores between segmented and unsegmented conditions. When width estimates were included in the model, the effects of segmentation were still significant for both rating, $F(1, 443) = 4.14$, $p = .04$, and average word count measures of elaboration, $F(1, 443) = 3.91$, $p = .048$, indicating that the differences in widths did not account for the tendency in people to elaborate more when given unsegmented, compared to segmented, text boxes.
4. We modeled these variables as normally distributed in our sample, as the sample size was large and the means were consistently unequal to the variance. Reanalysis with Poisson models yielded the same results as the linear models reported in the results.

References

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3–5. Retrieved from http://doi.org/10.1177/1745691610393980

Campbell, R. S., & Pennebaker, J. W. (2003). The secret life of pronouns: Flexibility in writing style and physical health. *Psychological Science*, *14*, 60–65. Retrieved from http://doi.org/10.1111/1467-9280.01419

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, *29*, 2156–2160. Retrieved from http://doi.org/10.1016/j.chb.2013.05.009

Christian, L. M., & Dillman, D. A. (2004). The influence of graphical and symbolic language manipulations on responses to self-administered questions. *Public Opinion Quarterly*, *68*, 57–80. Retrieved from http://doi.org/10.1093/poq/nfh004

Christian, L. M., Dillman, D. A., & Smyth, J. D. (2007). Helping respondents get it right the first time: The influence of words, symbols, and graphics in web surveys. *Public Opinion Quarterly*, *71*, 113–125. Retrieved from http://doi.org/10.1093/poq/nfl039

Couper, M. P., Conrad, F. G., & Tourangeau, R. (2007). Visual context effects in web surveys. *Public Opinion Quarterly*, *71*, 623–634. Retrieved from http://doi.org/10.1093/poq/nfm044

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys: The tailored design method* (Third). Hoboken, NJ: John Wiley & Sons, Inc.

Fuchs, M. (2009). Differences in the visual design language of paper-and-pencil surveys versus web surveys: A field experimental study on the length of response fields in open-ended frequency questions. *Social Science Computer Review*, *27*, 213–227. Retrieved from http://doi.org/10.1177/0894439308325201

Griffin, G., & Jacob, R. (2013). Priming creativity through improvisation on an adaptive musical instrument. In *Proceedings of the 9th ACM Conference on Creativity & Cognition* (p. 146). Yi-Luen Do, E., Dow, S., Ox, J, Smith, S., Nishimoto, K., and Tan, C.T. (Eds.). New York, NY: ACM Press. Retrieved from http://doi.org/10.1145/2466627.2466630

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). Questions and answers in surveys. In R. M. Groves, G. Kalton, J. N. K. Rao, N. Schwarz, & C. Skinner (Eds.), *Survey methodology* (Second, pp. 217–257). Hoboken, NJ: John Wiley & Sons, Inc.

Guilford, J. P. (1967). *The nature of human intelligence*. New York, NY: McGraw-Hill.

Hofelich Mohr, A., Sell, A., & Lindsay, T. (2015). *Thinking inside the box: Data from an online Alternative Uses Task with visual manipulation of the survey response box [dataset]*. Data Repository for the University of Minnesota. Retrieved from http://hdl.handle.net/11299/172116

Hommel, B., Colzato, L. S., Fischer, R., & Christoffels, I. K. (2011). Bilingualism and creativity: Benefits in convergent thinking come with losses in divergent thinking. *Frontiers in Psychology*, *2*. Retrieved from http://doi.org/10.3389/fpsyg.2011.00273

Israel, G. D. (2010). Effects of answer space size on responses to open-ended questions in mail surveys. *Journal of Official Statistics*, *26*, 271–285.

John, O. P., & Srivastava, S. (1999). The big-five trait taxonomy: History, measurement, and theoretical perspectives. In L. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102–138). New York, NY: The Guilford Press.

Keusch, F. (2014). The influence of answer box format on response behavior on list-style open-ended questions. *Journal of Survey Statistics and Methodology*, *2*, 305–322. Retrieved from http://doi.org/10.1093/jssam/smu007

Lewis, S., Dontcheva, M., & Gerber, E. (2011). Affective computational priming and creativity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11). ACM, New York, NY. Vancouver, BC, Canada, 735–744.

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, *54*, 217–224. Retrieved from http://doi.org/10.1080/00031305.2000.10474549

R Core Team (2014). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria. URL http://www.R-project.org/.

Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., … Richard, C. A. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts*, *2*, 68–85. Retrieved from http://doi.org/10.1037/1931-3896.2.2.68

Smyth, J. D., Dillman, D. A., & Christian, L. M. (2007). *Improving response quality in list-style open-ended questions in web and telephone surveys*. Paper presented at Annual Conference of the American Association for Public Opinion Research, Anaheim, CA.

Smyth, J. D., Dillman, D. A., Christian, L. M., & McBride, M. (2009). Open-ended questions in web surveys: Can increasing the size of answer boxes and providing extra verbal instructions improve response quality? *Public Opinion Quarterly*, *73*, 325–337. Retrieved from http://doi.org/10.1093/poq/nfp029

Spörrle, M., Gerber-Braun, B., & Försterling, F. (2007). The influence of response lines on response behavior in the context of open-question formats. *Swiss Journal of Psychology*, *66*, 103–107. Retrieved from http://doi.org/10.1024/1421-0185.66.2.103

Stern, M. J., Dillman, D. A., & Smyth, J. D. (2007). Visual design, order effects, and respondent characteristics in a self-administered survey. *Survey Research Methods*, *1*, 121–138.

Tourangeau, R., Couper, M. P., & Conrad, F. (2004). Spacing, position, and order: Interpretive heuristics for visual features of survey questions. *Public Opinion Quarterly*, *68*, 368–393. Retrieved from http://doi.org/10.1093/poq/nfh035

Wiseman, R., Watt, C., Gilhooly, K. J., & Georgiou, G. (2011). Creativity and ease of ambiguous figural reversal. *British Journal of Psychology*, *102*, 615–622. Retrieved from http://doi.org/10.1111/j.2044-8295.2011.02031.x

Zuell, C., Menold, N., & Korber, S. (2015). The influence of the answer box size on item nonresponse to open-ended questions in a web survey. *Social Science Computer Review*, *33*, 115–122. Retrieved from http://doi.org/10.1177/0894439314528091

Author Biographies

Alicia Hofelich Mohr is a research data manager in the College of Liberal Arts at the University of Minnesota. She received her PhD in psychology and MA in statistics from the University of Michigan in 2012. Her interests include research methodology, reproducible research, and open data. She can be contacted at hofelich@umn.edu.

Andrew Sell is a research project designer in the College of Liberal Arts at the University of Minnesota. He received his MBA at the University of Minnesota in 2013. His research interests include questionnaire design, visual design theory, and survey nonresponse. He can be contacted at sell0136@umn.edu.

Thomas Lindsay is the coordinator of research support services in the College of Liberal Arts at the University of Minnesota. He received his MA in history from the University of Minnesota in 2003. His interests include experimental research design, research regulation, and data management. He may be contacted at lindsayt@umn.edu.