Is it what you say, or how you say it? Exploring the effects of email messages for online panel surveys.

Authors: Zoltan Fazekas, Andre Krouwel, Matthew T. Wall

Abstract:

Variation in levels of survey participation and in the quality of individual responses, as well as the problem of attrition in panel studies are ongoing core concerns for all survey researchers. Various reward and incentive structures are employed in order to maximise response rates and response quality. One component of this structure is the content and tone of covering letters or emails that ask respondents to participate (or, in some cases to continue to participate) in surveys. However, there is little evidence of a systematic approach to testing the effects of formulation, tone, and content of the messages that accompany survey requests on response propensity and response quality in the existing literature. We intend to fill this gap using a quasi-experimental research design in which 14,000 respondents to an online survey undertaken by Kieskompas.nl (a Dutch electoral advice website) are re-contacted and asked to participate in an additional survey. We formulate 8 different email messages (as well as using a 'baseline' boilerplate letter) that tap into three dimensions of variation. We contrast altruistic and egoistic appeals, formal versus informal writing styles, and linguistically simple versus linguistically complex formulations. We analyze the differential effects stemming from various formulations on response propensity and over several data quality measures. Our analysis also seeks to uncover how demographic characteristics condition such effects.

1. Introduction

'*In short, a survey involves many decisions that need to fit together and support one another in a way that encourages most people to respond and minimizes inaccurate and inadequate answers'* (Dillman, Smith and Christian, 2008: 13).

In this article, we investigate whether varying one element in the implementation of an online survey can influence both response rates and response quality. The element that we focus on here is the email 'cover letter' that accompanies online invitations to participate. Attempts to assess the effects of the design and contents of cover letters for improving response rates in self-completion surveys have found little consistent evidence of strong effects (Redline, Oliver, and Fesco, 2004). However, online self-completion surveys differ from paper versions in that the link which survey targets must click on to complete the survey is embedded in the 'cover letter' email. As such, we have reason to suspect that the contents of such emails may affect the likelihood that targets respond to the survey, as well as influencing the quality of the responses that they provide.

Survey response and response quality maximization are core issues of public opinion research, and the attention being devoted to these issues by practitioners is on the rise (Curtin et al, 2000). Online recruitment and collection provides potentially enormous benefits in terms of cost and speed of data entry. There also appear to be advantages to be exploited in terms of data quality: Chang & Krosnic's (2009) research indicates that, in terms response quality (as measured by random error, satisficing, and social desirability), internet research outperforms random digit dialling. Within internet research volunteer (i.e. non-probability opt in) samples had the highest quality responses.

However, issues of non-response and bias are even more pronounced for online than for offline research, and methodologies for maximizing response rates + quality in an online-only environment are in their infancy. Furthermore, weighting procedures alone have proven insufficient to redress instances of bias using non-probability recruitment methods (Loosveldt and Sonck, 2008; see also Couper, 2000 for an overall typology of web-based surveys). Sills and Song (2002) conclude argue that 'Low response rates, self-selectivity of Internet users, technological issues with the deployment of the research tool, and concerns over Internet security have troubled recent studies. Yet, for special populations that regularly use the

Internet in their daily lives, the new medium has been found to be a sensible means of achieving meaningful results' (Sills and Song, 2002: 23). As the proportion of the population who regularly use the internet is growing rapidly in many societies, the problems of coverage associated with online surveying should decrease. This argument suggests that, if they can improve their performance in terms of response rate, internet surveys should be increasingly useful for gathering public opinion data.

As such, internet surveying as a technique holds great potential for advancing our capacity to analyze public opinion, with the combination of probability sampling and internet surveying representing a particularly promising avenue (Chang and Krosnic, 2009). There seems little doubt that self-administered online surveys will become a significant element of public opinion research in the years to come (Dillman, 2007). However, assuring high levels of survey response is of particular concern for internet surveys. We seek to investigate here whether the contents of the email soliciting survey response, a cost-free design feature of an online survey, can be manipulated to alter response rates and response quality.

We focus on three aspects of the email message that online survey respondents receive. These are: 1) the type of appeal made to the respondent 2) the complexity of the message, in terms of writing style and 3) the tone of the email. We dichotomize each dimension and composed 8 email 'cover letters' which comprise all possible combinations of these dimensions. We also deployed 2 'control' or 'baseline' letters – which comprised the standard text sent out by the Dutch polling company Synovate. A large panel of internet survey volunteers (N=14,000) were then randomly assigned (with assignment stratified by age, gender, and education to minimize significant differences across groups with regard to these characteristics) to one of 10 groups. The treatment in this experiment is the type of letter received, and each group was assigned a specific letter. In the next section, we seek to situate our research with regard to existing studies of areas of survey implementation that have been found to reduce survey error. Having done so, we describe the treatment letters and provide quantitative estimates of the extent to which they capture variation on our conceptual dimensions of 'appeal', 'complexity', and 'tone'. We then analyse the results of our experiment, before concluding by drawing out the consequences of this study for practitioners engaged in online surveying.

2. Research Contribution

Fundamentally, of course, the goal of survey design is to permit the surveyor to make valid inferences about the population that she is studying on the basis of a sub-sample of that population. The entire basis for making of such inferences is a set of statistical models which assume that all elements of the target population have a random chance of being included in the survey sample. To the extent that this is not the case, and particularly to the extent that certain sections of the population are *systematically* more or less likely to be sampled – a survey will generate less reliable, and potentially biased measures of tendencies in the population.

The classic approach to survey error involves a division into four basic types of error: coverage, sampling, measurement, and non-response (Dillman, Smith and Christian, 2008). Sampling error relates to the difficulty of drawing reliable inferences for the population on the basis of a small sample of observations from that population. Even assuming that the sampling procedure is entirely randomized, we must estimate population characteristics with considerably wider margins of error when samples are small than when they are large. Coverage error occurs when not all members of the population have an equal known probability of being included in the survey sample – this is a problem common to internet surveys, as internet access and regular use is not universal across Dutch or other societies.

Measurement error arises when poorly worded question design makes it difficult to interpret findings, or when social cognitive mechanism lead to respondents 'satisificing' their way through surveys – rather than communicating their opinions. Biemer (2010) also notes that measurement error can arise in the processing of respondent data, due to errors in coding, keying, and editing of survey datasets. Finally, non-response error is a particular *bête noire* of survey researchers. Non-response comes in two forms – the first is survey non-response, where a sampled individual simply refuses to respond to the survey, the second is item non-response, where the respondent refuses to register a response to a particular item – either by skipping that item completely, by refusing to respond explicitly, or by stating that they 'don't know' the answer to a question about their own opinion.

While internet surveying makes it easy to contact people in larger numbers and minimize sampling error, non-random 'opt-in' online panels face problems of significant coverage error,

high response error relative to other modes, and potentially, measurement error (Couper, 2000). As such, to fully capitalize on the advantages offered by online public opinion data collection, it is important that academics and practitioners be intent on maximizing response rates and response quality in every aspect of the design of online surveys.

Historically, there is little evidence that content of survey cover letters has a strong effect on response rates in off-line survey environments (Harvey, 1987) although some research has pointed to small, but significant effects due to varying letter content (Brennan, 1992; Redline, Oliver, and Fesco, 2004). Dillman's (1978) 'Total Design Method' focuses on maximizing survey response and quality at every stage of the survey design and implementation. Built on the premises of social response theory, the Total Design Method seeks to orientate survey implementation and design process towards inclining the voter to perceive that the social and tangible rewards of participating in the survey outweigh the losses in terms of time and effort.

To be sure, the design and content of the cover letter that accompanies self-administered surveys is one of a range of elements to be considered in an overall survey design, and this research should be seen as only one of many necessary contributions to improving response rate and answer quality. Several studies have found that techniques such as monetary incentives, advance (snail mail) letters and telephone follow ups, questionnaire design, can serve to improve response rates and quality (Dillman, 2007; Dillman, Smith, and Christian, 2008; Rao, Kamiska and McCutcheon, 2010). However, gains from these methods come with associated costs in terms of expense and time for public opinion researchers. Altering the content of the cover email, on the other hand, is virtually free of charge to practitioners and therefore should be a early step in designing any online survey.

Our research is guided by the following simple argument. A major difference between online and other survey formats is that digitized email messages take over the role of classic cover letters or of human interviewers. Such email 'cover letters' contain the link where the potential respondent can access the online survey. Thus, we anticipate that the content of email messages is more important for online surveys than cover letters were for paper questionnaires. This article provides objective evidence of the scale of the effects of choices concerning cover letter content on response rates and answer quality for a non-probability online panel survey. We also seek to draw lessons for practitioners on how best to tailor their cover emails to maximize response and quality on the basis of this evidence.

In the next section, we outline our approach to varying letter content, elaborating the dimensions of variation considered and the theoretical justification for focusing on these elements. We also provide quantitative estimates of the extent to which the contents of the survey letters that we drew up for the experiment differed along the dimensions of interest in this study. We then elaborate how the survey experiment was carried out, with respondents randomly assigned into 10 different survey groups, with each group receiving a different letter. In our analysis we investigate the overall effects of letter type on the response rates and response qualities of each group. In our conclusion we extrapolate practical recommendations for online survey practitioners on the basis of our findings.

3. Variations in message, tone, and complexity of email 'cover letter' messages.

In our letter formulation process, we conceptualized the online 'cover letter' as a persuasive document, designed to influence the motivations of respondents in such a way that they were more likely to respond to the survey. This approach builds on Dillman's (2007) contention that the interaction that takes place between a surveyor and a survey respondent is best conceptualized as a social exchange. Dillman's model of this exchange focuses on three dimensions – *reward*, *cost*, and *trust*. His formula for maximizing survey quality in implementation involves deploying measures which build trust (for instance, accompanying a survey request with a pre-paid sum, as well as via multiple contacts) and improve respondents' perception of the rewards that they will accrue from participating (such rewards can be either material or immaterial) while reducing the perceived costs of participating (in terms of time, effort, and, potentially, the risk of confidentiality breaches). Overall, Dillman's approach leads us to focus on the motivations of respondents, and to treat respondents as intelligent social beings – meaning that one must maximize those elements that will lead respondents to positively evaluate participating in the survey, and will motivate them to complete each survey diligently.

In terms of an online cover letter – we sought to find a method for manipulating each of these three dimensions. The first dimension that we considered was *reward*. Given that no monetary incentives were offered in our design, it appears unlikely that a survey cover letter, in and of

itself, can be manipulated to affect the material reward/cost structure of the survey request. However, Hansen's (1980) 'self perception' model of survey response indicates that survey requests can offer internal motivators to encourage survey response by associating survey completion with either personal or societal rewards. This is done by weaving a specific appeal into the text of the letter – these appeals do not ask respondents to fill out the survey to maximize survey accuracy – but rather they emphasize the intangible personal benefits that come from expressing one's opinions, or the benefit that the research represents to society as a whole. The former appeals can be considered 'egoistic' (appealing to the respondents' sense of self) and the latter 'altruistic' (appealing to respondents' sense of social obligation) (Redman, Oliver, and Fesco, 2004).We therefore divided letters into two types of underlying appeal: egoistic and altruistic. The 'egoistic' letters focus on the respondent as individual, emphasizing how important, reliable and valuable the respondents' opinion is. The altruistic letters, on the other hand, build on the idea of reciprocity, contribution to research and society being achieved through responding to the survey.

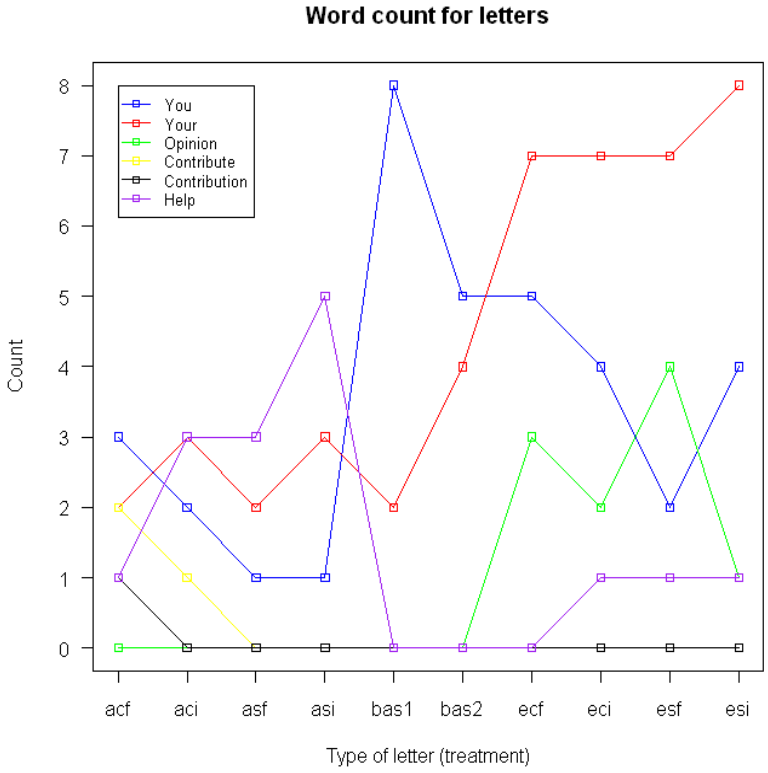
According to these premises, for egoistic messages we employed a vocabulary that emphasized words such as "you", "your", "opinion", whereas for altruistic messages words as "contribute", "contribution", "help". Of course, all of these framing efforts take place within the constraints that the email is comprehensible, communicates the topic of the survey, and asking politely for a response: therefore, for instance, the word "help" was not completely dropped from egoistic messages. Using text mining, the appearance and frequency of the previously mentioned set of words was analyzed. Figure 1 displays absolute count values for each in all eight letters, with the altruistic letters displayed on the left hand side, the 'baseline' letters[1] in the centre and egoistic letters on the right. We can from Figure 1 that see that the words 'contribute' and 'contribution' appear only in the altruistic messages (sparsity of 80% and 90%)[2], whereas 'opinion' only appears in the egoistic messages (sparsity of 60%, only appearing in egoistic messages). The use of 'you' and 'your' is much higher for egoistic messages than altruistic messages. 'Help' appears much more often in altruistic messages, but it is also present for the of the egoistic messages, where it appears once. The baseline messages do not contain any of the words specified as altruistic, and they contain a substantial numbers of uses of 'you' and 'your', placing them closer to the egoistic than the altruistic end of our 'appeal' dimension. The

[1] These letters were those used from the previous survey wave, we elaborate on these letters in the design and analysis section.

[2] As they are considered to be "hard" words, they do not appear in the more simple or informal versions.

low frequency counts for the major words on both dimensions reflect the functional requirement that the messages not to be overly long (no letter was longer than 250 words).

Figure 1: Word counts extracted by text mining



The second dimension of Dillman's (1979) social exchange matrix is *cost*. The cost of a survey completion is the time, effort and risk that a respondent perceives to be associated with completing a survey. In terms of an email cover letter, we argue that the linguistic complexity of the survey proposition can act as a proxy for the cost of completion. This is a particularly important element for online surveys, where the content and design of the survey itself is typically not contained in the contact email. All users have to go on in evaluating the survey cost is the email 'cover letter'. We therefore posit that complex language patterns in the cover letter may serve to increase the perceived cost of completion, and consequently depress response rates. Simple messages, on the other hand should minimize the cost perception of respondents, and improve response rates. We therefore divide email messages into simple and complex. Simple messages were written in words with small numbers of syllables, and avoiding multi-clause sentence structures. Complex formulations employed longer wording and complex sentence structures. Again, to maintain realism – the simple formulations were not childishly simple and the complex formulation was not impenetrably complex. As such, integrating this design element into letter composition involved a degree of subjective creativity, however we

were able to examine the extent to which 'complex' letters differ using a computer algorithm designed to distinguish textual complexity.

Table 1: Readability scores for the letters

Letter	Flesh Kincaid Grade Level
ACF – Altruistic-complex-formal	12.09
ECF – Egoistic-complex-formal	10.77
ECI – Egoistic-complex-informal	10.44
ACI – Altruistic-complex-informal	10.23
B1 – Baseline 1	9.24
B2 – Baseline 2	9.33
ESF – Egoistic-simple-formal	9.42
ASF – Altruistic-simple-formal	8.02
ESI – Egoistic-simple-informal	7.56
ASI – Altruistic-simple-informal	6.27

As a benchmark for the difficulty of the text, we present in Table 1 the Flesh Kincaid Grade level scores for each email. Here, the lower scores indicate easier understanding, or more precisely lower number of grades in formal education[3]. Firstly, we can see that the baseline messages line up nicely in the middle of the scale, falling somewhere between the simple and complex formulations. The Flesh Kincaid Grade scores indicate that 'complex' messages are much harder to read. However, tone, which we shall discuss next, is not unrelated to complexity. The combination of a 'simple' formulation and an 'informal' tone generated the easiest messages to read (6.27, respectively 7.56), while 'formal' messages were more difficult to read across the board, meaning that the most complex formulation was the two combining formal tones and complex formulations.

[3] These scores are reported for the English texts – before translation. For the Dutch texts we use the DOUMA score, but for reasons of consistency and easier comprehension we present the results on each dimension for the initial English texts. The scores for the Dutch texts are in line with these scores.

Thirdly, in as much as one can, we seek to manipulate the level of *trust* that the cover letter generates. We do so by manipulating the tone of the letter. Given the novelty of this research in a Dutch cultural context, it is difficult to predict the effects of letter tone. On the one hand, a formal tone designates legitimacy and authority. For instance, Brennan (1992) found that cover letters signed by researchers whose title designated high statsus on the research team generated higher completion rates than cover letters signed by low ranking researchers. On the other hand, a friendly tone can establish a positive and trusting communication stream. It seems most likely that respondents' interpretation of tone are a result of societal or personal factors – some people may place greater trust in formal letters others in informal letters.

In terms of writing the letters the tonal dimension was the easiest of the three to manipulate, since the Dutch language has separate formulations for "You" (formal) and "you" (informal), essentially covering this dimension perfectly. However, to make sure that the letters capture substantively different tones, we also altered the opening and closing sections of the letters. Whereas, in the formal versions we posit the topic of the survey and introduce Kieskompas, the informal messages start with 'Hi' or 'Greetings'. For informal messages, we also modified the first sentence, so that the Kieskompas Director introduces himself by name to induce a more personal, informal, and closer atmosphere. Furthermore, in the closing lines we used "Kind regards" in the informal messages, instead of "Yours sincerely" in the formal messages – followed by signature in both cases.

Overall, then we have three dimensions of letter variance designed to capture the three-dimensional matrix of survey participation developed by Dillman (1979) which focused on the reward, cost, and trust elements of surevey completion as a form of societal interaction. Table 2 below recaps how we operationalise each of these elements in our experimental study. Combining all three dimensions generates 8 letter types (2X2X2) listed in table 2, with a further 'baseline' letter deployed to test these formulations against existing practice. In the next section we elaborate on how the experiment was designed and executed, before providing an analysis of the observed effects of email wording on response rates and quality over 14,000 respondents.

Table 2. Relationship between treatment operationalisations and Dillman's (1979) response theory.

Element of survey proposition (Dillman's 1979 typology)	Corresponding feature of email cover letter	Experimental manipulation of each element (treatments)
Reward	Appeal	Altrusitic versus egoistic
Cost	Lexical complexity	Complex versus simple
Trust	Tone	Formal versus Informal

List of Abbreviations of treatments arising from this approach:

ACF – Altruistic-complex-formal

ACI – Altruistic-complex-informal

ASF – Altruistic-simple-formal

ASI – Altruistic-simple-informal

ECF – Egoistic-complex-formal

ECI – Egoistic-complex-informal

ESF – Egoistic-simple-formal

ESI – Egoistic-simple-informal

A final step in our analysis of the contents of the email letters derived from the approach described in table 2 involves Wordscores. Before running the analyses, one must emphasize that the scores obtained here may be misleading for at least two reasons. First, the texts are relatively short, but more importantly, as we have described, there are 3 underlying dimensions in the texts, and Wordscores is suitable to detect one dimension, based on the reference texts chosen. However, this approach informative, because it uncertainty measures are provided.

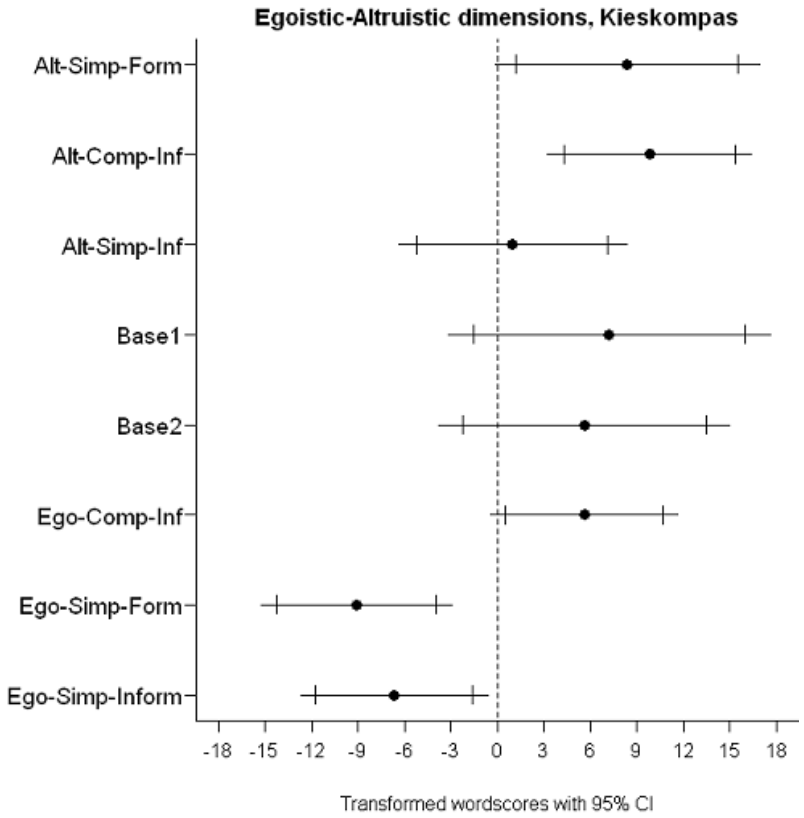
ACF and ECF were chosen as reference texts for this analysis, because ECF is the longest and richest text, and the counterpart on the altruistic end of the scale was needed, to have two similar anchors (even though it is not the longest text). After running the analysis on the raw texts, Figure 2 displays the point estimates with the 95% confidence intervals. The results are mixed, but considering the previously mentioned caveats, they can be considered reasonable[4]. The point estimates on average look acceptable, however the confidence intervals are relatively wide. Both ACI and ASI are different from ESF and ESI, but ASI and ECI behave rather strangely. In case of ASI this is not surprising, since the text is relatively short. When it comes to ECI, a potential explanation could be that a complex but informal combination may be very limiting on word choice, and it is an unusual combination. Furthermore, if ACF and ECF are the reference texts at 10 and 0 – although no confidence intervals are available for these – they place in the middle, making them indiscernible from the other combinations[5].

Finally, the position of the baseline messages is of interest. As we can see in Figure 2, they appear to be more altruistic than egoistic in their appeal. The baseline emails were previously employed as standard recontact emails, thus it is not surprising that they have an altruistic component asking for participation and help in understanding the political arena. This position on the altruistic-egoistic dimension comes against what the word count statistics suggested. From this step it is clear though that they might have words and formulations that are more altruistic, although they are not necessarily overlapping with the words we designated as altruistic in our emails. Again, knowing the position of the baseline messages is important, since they will be included in the multivariate analysis.

[4] After the automatic rescale in Wordscores (Stata) the 0-10 scale was extended, most probably generated by the low number of words and short texts.

[5] We replicated the same Wordscores analysis using the Austin package in R. The results are reported in Appendix X. Although everything was kept the same -- the only difference from the analysis carried out in Stata is that punctuation was removed, but no stemming or stop word deletion was employed -- the results are much more promising. On one hand, this is good news, as even after rescaling, all egoistic and altruistic letters are more discernable, with non-overlapping upper or lower bounds for the confidence intervals. On the other hand, caution is required, as the replication shows different results that are only dependent on the software. This may be given by the fact that only around 60% of the words can be scored, and again, the texts have more underlying dimensions that interact, and may bias the results.

Figure 2: Transformed wordscores, with reference ECF (0) and ACF (10)



We also extracted separate single wordscores for the words of interest mentioned above. These statistics are very much as expected[6]. Looking at the single wordscores based solely on the reference texts, "help", "contribute", and "contribution" all score 10, meaning that they are associated with altruism. "Opinion" on the other hand, has the score of 0, being associated with the egoistic appeal. "You" receives a score of 4.17 and "Your" 2.54, both being closer to the egoistic end of the scale. Nevertheless, since these are used in common sentences in all of the letters – inevitable, considering the "request" nature of the letters – the scores are not that extreme.

Overall, in this section we offered quantitative, text based statistics to describe our letters that will be used as treatments, focusing on how different they are. This step was essential, because we want to be reassured that the treatments are indeed different and do capture something of the theoretical constructs to which they relate, as described in Table 2.

[6] These scores were extracted using the Austin package in R.

4. Study Design

Our quasi-experimental design uses a panel of respondents who left an email and indicated consent for recontact on a Dutch vote advice application website: Kieskompas.nl. Several members of this panel and already been contacted for political surveys designed by Kieskompas and implemented by the public opinion agency Synovate. This approach suggests a possible limitation of our study related to self-selection. Accordingly, our full sample is composed by individuals who were interested enough to previously express their opinions and further help our research, making them examples of politically more interested individuals (on average). However, our goal is not to generalize our findings to the level of Dutch people, but to assess specific effects of email messages in online surveys. Moreover, this aspect of our sample also suggests that we should not anticipate that email content should have a large impact on response rates or response quality, since members of the panel have already had experience with surveys and are obviously interested in the political arena. Nevertheless, as we will see below, there is evidence that letter contents had significant effects.

The pooled dataset had 14,206 respondents. They can be grouped into three categories that are relevant for our study: (1) 6,113 respondents who had not yet been re contacted by Kieskompas and Synovate, (2) 2,269 respondents who had been re contacted once already, and (3) 5,824 respondents who had been re contacted twice already. This categorization is important because depending on the frequency of previous collaboration we may expect different response behavior. The two previous survey waves had elicited high response rates – the first wave (sent to respondents from category 3 above) had a 62% response rate (3,619/5,824) while the second wave elicited at 52% response rate (4,257/8146). Both waves took place very close to the panel sign-up which took place during the Dutch legislative and election campaigns in May and June 2010. The topics studied in these waves were leadership evaluations, media consumption, vote choice, and perceptions of the Kieskompas.nl product. The 6,113 respondents who had not been re contacted at the start of the experiment had left their emails on the Kieskompas.nl local elections websites in March 2010.

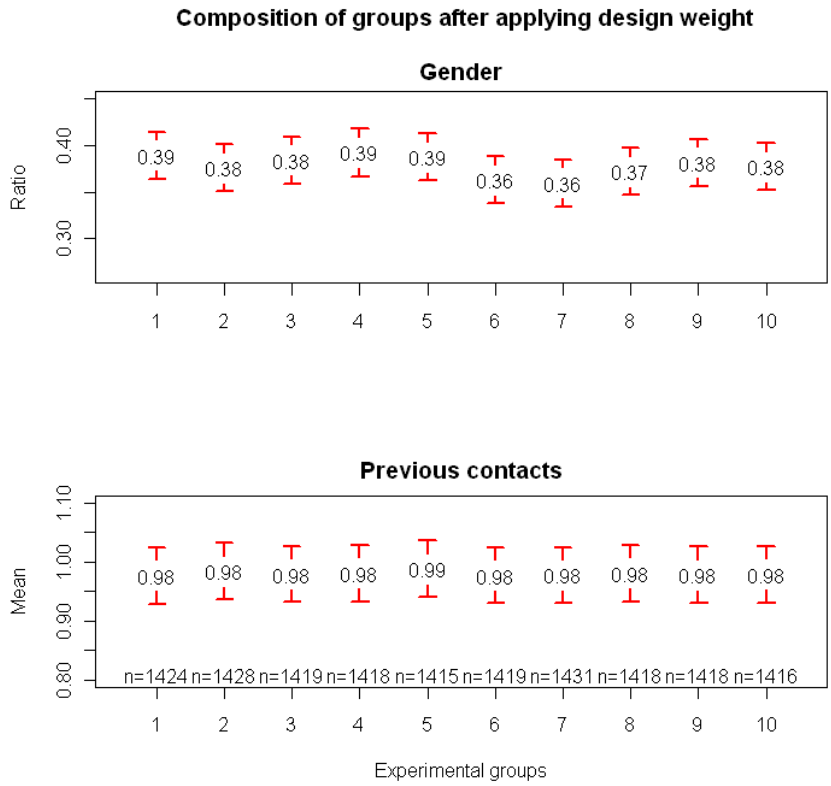
As described above, we have 8 ‘treatment’ letters and a ‘baseline’, which was simply the same letter as used in the previous survey waves (i.e. it was non-specific in terms of the survey’s

content/subject), was deployed to two groups. Overall, this approach meant that we needed 10 experimental groups from our 14,206 respondents. Indiidvuals were randomly assigned to groups, which were generated using stratified random selection to equalize gender, education, and number of previous contacts from Kieskompas/Synovate across each group. We chose this method because we wanted to minimize the between-group variation for: gender, education, and previous contacts. After the stratified random allocation, a letter was assigned to each group – thus our treatment allocation was randomized.

We deployed the baseline to two experimental groups as a control in order to observe the extent of between-group variation in the outcomes of interest when the letter was identical for both groups. Whatever variation we see between these groups on any of our variables of interest can be considered as a default of baseline level of acceptable variance, in a way, measurement error associated with the items. If the between-group variation for the two control groups is not significant or low in magnitude – together with the stratified random allocation – then we can be more confident that the other between-group variation is associated with to our treatment. Secondly, the baseline groups give us an insight on whether there is topic-specific variation compared to previous surveys designed by Kieskompas. As this wave's topic was coalition formation and it was not fielded during an election campaign, we might expect that the topic influences response rates or data quality indicators. In order to have a clear grasp of this, we kept the baseline messages used previously, to be able to isolate only the topic specific effects.

Figure 3 displays the composition of the groups for gender and previous contacts after the stratified allocation (on the covariates of interest). It should be noted that we did not stratify group composition in line with the composition of the general Dutch population, but instead we sought to align each group with the composition of our pooled set of respondents (14,206). There are no major differences in composition on the covariates we stratified; moreover none of these are even remotely close to reaching statistical significance, and hence our experimental groups are balanced for these dimensions.

Figure 3: Mean differences with 95% confidence intervals for experimental groups



The survey concerned coalition formation following the national elections, and was sent to respondents between the 12th and 22nd of December 2011. It is possible that the topic had some deflating effect on respone rate (see next section) however, the considerable time that had elapsed between previous contact (either in signing up via the local election sites in March, or receiving a post election survey in June) probably explains why response rates generally were lower for this wave than previous ones. In the next section we analyze differences in response rates and qualities across the treatment groups.

5. Results

Our analysis is divided into two sections. The first part analyzes **response rates** by comparing them according to the type of cover letter received. The second part focuses on the issue of **response quality.** As we will observe, these two categories do not always conform to identical logics in our analysis.

Response Rates

Following a classic approach when dealing with different control and treatment groups, we are interested in the average treatment effect associated with our manipulations in terms of appeal, complexity, and tone[7]. Figure 4 displays the average response rates for each experimental group (with 95% confidence intervals). We can see that the baseline letters turn out to fare the best of all treatments in terms of response rates (around 33% response rates for each). Also, as there is no significant difference between the 2 groups who received identical baseline letters, we can be more confident that the rest of the between-group variance is associated with the treatment we employed. Altruistic appeals generated higher response rates on average, and by looking at panel 4 in Figure 4, we see that these differences between egoistic and altruistic appeals are statistically significant. Altruistic appeals register 4% higher response rates on average, compared to egoistic messages.

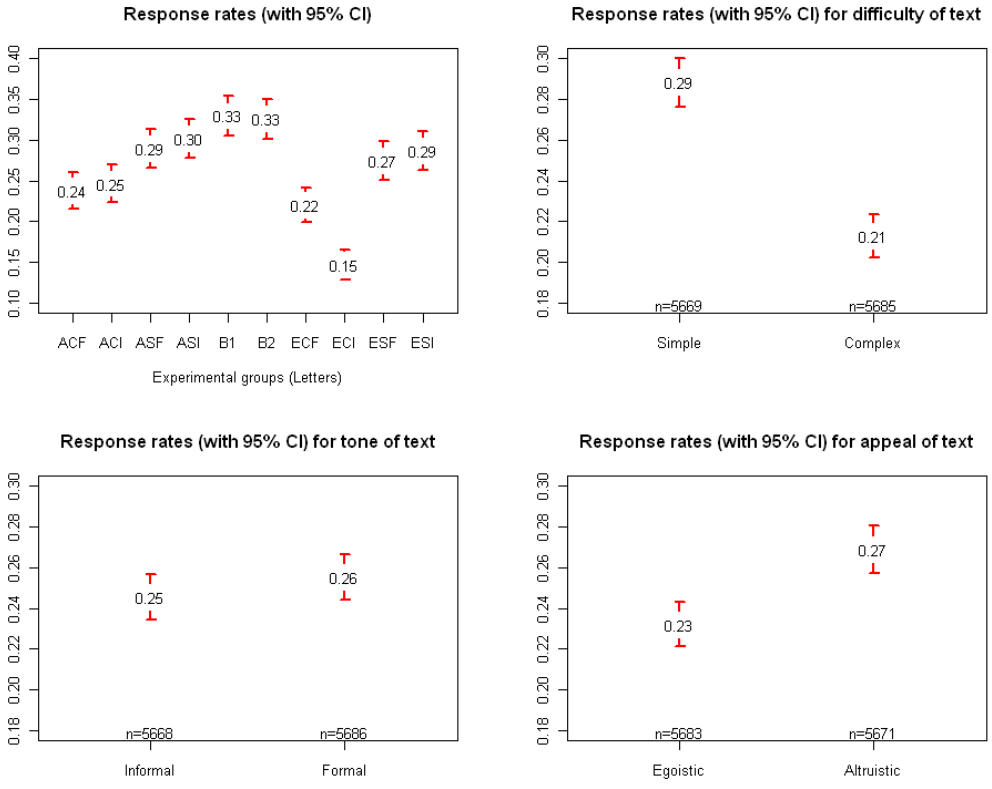
On both sides of the appeal dimension, complex messages perform badly. However, if they are paired with an altruistic appeal the situation is not as grim, but the response rates are still lower. The egoistic-complex-informal message has the lowest response rate, 15%, which is also statistically significantly lower than the rates for any other group. By further analyzing separate batches of the dimensions in Figure 4, we indeed see that simple messages have an average of 29% response rate, whereas complex messages only 21%.

[7] The obvious limitation of this approach is that the initial sample of around 14 000 respondents already makes use only of those people, who once supplied their email address and voiced their interest in participating in upcoming surveys. Thus, our findings should not be generalized for very low propensity individuals, as they already dropped out from the sample.

The difference between formal and informal messages in minuscule, thus we cannot tell with certainty which one fairs better.

A tentative and partial conclusion from this step would be to keep you message simple, as surveyor. Also, if this is not possible or the messages are medium difficulty, paired with altruistic appeal work much better in attracting respondents. An egoistic appeal in a complex message that still tries to keep an informal tone yields 50% lower number of respondents than a baseline a simple altruistic or a simple egoistic message. The difference in response rates illustrates the pitfalls inherent in email message design – poorly designed message can cost significant numbers of respondents in the online space.

Figure 4: Response rates for experimental groups treatment



Response Quality

Turning to our second stage of the analysis, we concentrate only on the quality of responses that we received. In this case, we are interested in analyzing between-individual differences, examining the treatment effect associated with the re-contact email messages. A first indicator on response quality is **item non-response**. For this indicator, we simply compare the average level of item non-response among groups as a first step, and furthermore we employ a regression analysis, where the response variable is the count of items left unanswered for each individual.

Figure 5 displays the group average item non-response (as count for all the items) with 95% confidence intervals. As expected, since all of the panel respondents contacted for this experiment were opt-in volunteers, the non-response counts are very low. Even if none of the between-group differences are statistically significant, however, the groups still display different levels of missingness and we can identify that the 'baseline' groups register the highest average count of missing values. We must also remember that these groups had the highest response rates. Hence, in only descriptive terms, the hypotheses that various external motivators have divergent effect on response rates and on the actual data quality (Davern et al., 2003) seem to be grounded in this case. Nevertheless, as the differences are not significant and the counts are very low, this conclusion cannot be formulated yet.

In order to better understand the driving forces of the item non-response and its between group-variance, we will employ a multivariate analysis.

Figure 5: Item non-response across groups

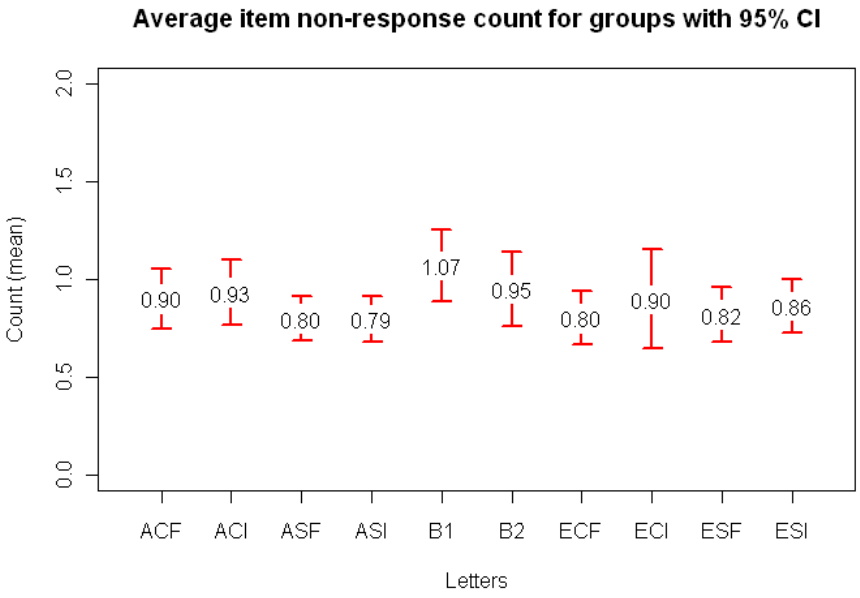
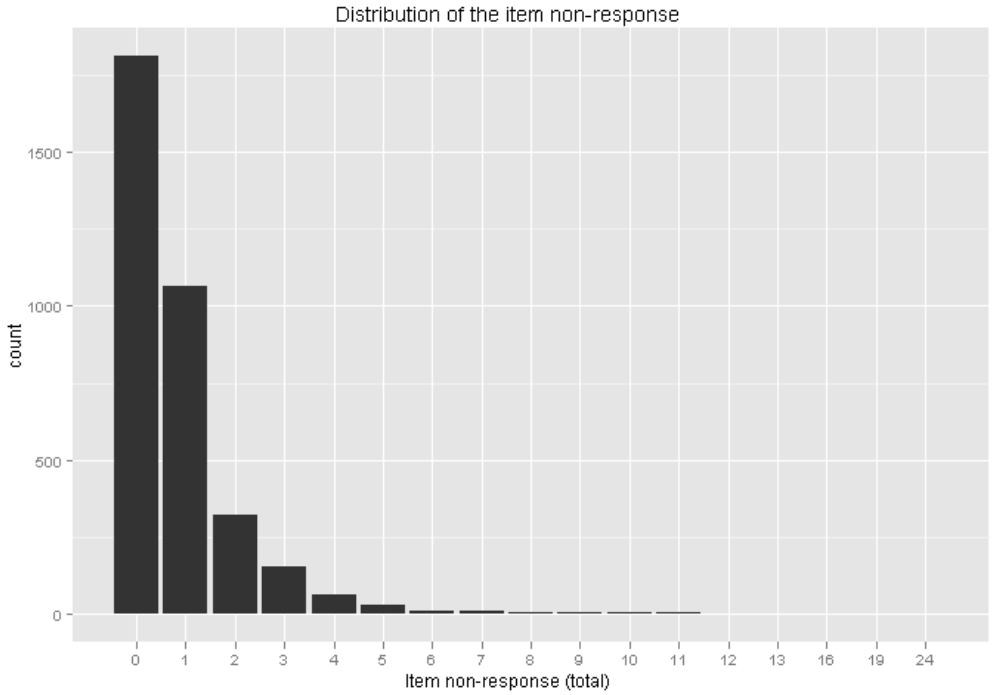


Figure 6: Distribution of item non-response



As we look at Figure 6, the distribution of these values clearly suggests that a classic OLS based linear model is not suitable for modeling the item non-response as dependent variable. First, it

is count data that cannot have negative values -- truncated. Second, there is an extremely high frequency of zeros. Given these properties of the distribution of the dependent variable – total count of missing values for each individual, we employ a zero inflated regression model with a negative binominal count distribution. We specify predictors for the count model and run an intercept only binary model[8], but the estimation is done simultaneously in a mixture model where parameters are estimated via maximum likelihood (Zeileis et al., 2010).

Alongside the group identification, we employ two control variables, gender (variable name *female*, coded 1 for female and 0 for male) and *age* (mean centered for a meaningful intercept). As the group identification is a nominal variable we run two distinct models. The first model includes the nominal group membership variable, but here we only focus on the predicted item non-response count for each type of letter. In the second model, we partition the group variable into 4 dichotomous variables, as follows: *altr*, coded 1 for altruistic messages and 0 for other; *form*, coded 1 for formal tone and 0 for other, *comp*, coded 1 for complex content and 0 for other; and *nobase*, coded 1 if the group did not get baseline message and 0 if it did. This final reversal is important and useful, because we want to estimate the effect of baseline letters in the intercept (when everything is 0). By employing this specification, we can see the effect of content, difficulty, and tone separately.

Table 3 displays the results of the two models where the dependent variable is the count of item non-response for each individual. As the link function is logit, these are log-odds, thus for the interpretation we will transform them (by taking the exponent) into odds.

[8] We do this, because on one hand we are interest in the effects of letters on the missing data count, not how would they predict having zero item non-response. On the other hand, to make sure, we ran the models with identical predictors for the binary component of the mixture, and the model fit was the same, with no additional fit gained by adding additional parameters (Vuong test insignificant, $p<0.17$ and $p<0.32$).

Table 3: Zero-inflated model results for item non-response

	Model 1	Model 2
Female	0.19 (0.05)*	0.18 (0.05)*
Age	-0.002 (0.001)	-0.001(0.05)
Group	-0.02 (0.008)*	
Altruistic		0.009 (0.05)
Formal		-0.03 (0.05)
Complex		0.07 (0.05)
Not baseline		-0.19 (0.07)*
Intercept	-0.09 (0.05)	-0.07 (0.05)
Intercept inflation model	-9.679 (10.6)	-13.12 (60.32)
Log(theta)	-0.06 (0.05)	-0.05 (0.05)
Log-likelihood	-4555	-4552
N	3506	3506

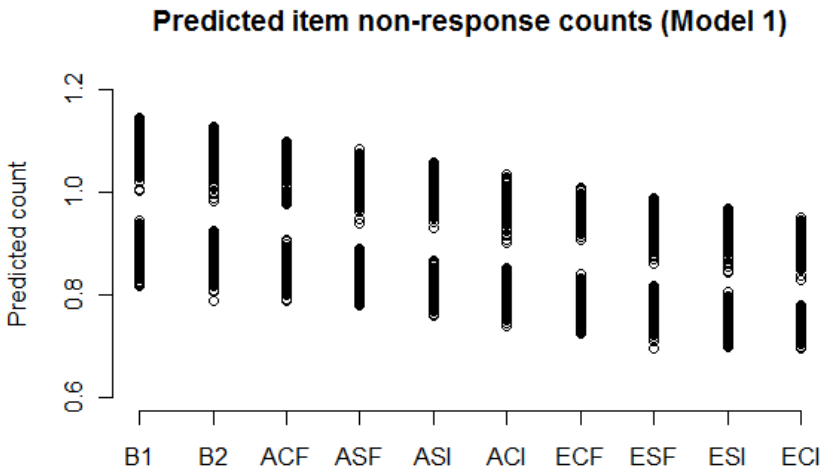
Note: * is $p<0.05$. For both models, the likelihood-ratio tests indicate significantly better fit than the baseline model. Untransformed coefficients with standard errors in parentheses.

As we can see, there is basically no difference in model fit between these two specifications and also the coefficients for the common variables are very similar. We find that gender is a

significant control, suggesting that in case of women we can expect higher item non-response counts – around 8.33% higher compared to men (0.93 vs. 1.2), everything else held constant.

Turning our attention to quantities of interest based on our designed experiment, we see a significant effect from the group identifier, suggesting that when it comes to item non-response, what letter one received indeed matters. In model one – as the variable is nominal – interpreting the coefficient is not meaningful, but the predicted counts linked to each group show how our pre-designed letters fair in relationship to item non-response. Figure 7 displays the predicted counts based on Model 1.

Figure 7:



In line with the difference suggested by the mean comparisons, we can see that although the baseline messages brought the highest response rates, they are also associated with higher item non-response count, or less data quality. Similarly, for the egoistic appeal letters we see, on average, lower item non-response predicted, the opposite relationship compared to what we found in case of response rates, where altruistic messages had a better performance.

Nevertheless, turning to Model 2 we see that if we model the group effect decomposed into dichotomous variables reflecting each dimension (and the baselines), the effect that stems from the letters is captured by the characteristic of not being in the baseline category. Given

our coding, we look at the baseline groups in the intercept, thus based on the results we can say that getting any type of letter – other than the baseline – decreases the expected count of item non-response. However, we do not have the necessary statistical results to discern between the letters that contained various framing. What we can say based on our empirical results is that if we would have repeated samples, in over 95% of the samples the effect of having a specific letter different from the baseline is expected to be different from zero, and increase the quality of the gathered data.

However, item non-response count is only one indicator of the data quality. We also test the effects of letters on how consistent the responses are. For this, we used mutually exclusive formulations in agree/disagree questions. This approach is linked to acquiescence, but in that case the emphasis is on excessive agreeing (or agreeableness) with statements, making the agreement with the formulation *X* more frequent than the disagreement of formulation *non(X)*(Krosnick, 1999; Krosnick et al, 1996; Krosnick, 1999). We analyze here two negative formulations, and look at inconsistency in answers as data quality indicators. The normal expectation is, if there is no acquiescence, to have a correlation of -1.0 minus measurement error (random) between the two items that are reversed pairs, however this does not happen[9] (Krosnick 1999).

A differentiation between the two reversal items has to be done because one of the probable causes of acquiescence and also inconsistency is the increasing fatigue during the interview session (Krosnick, 1999). Alo, reversals differ with 'close' reversals following each other within the same set of questions and 'distant' reversals separated by several intervening questions. We do not have any particular reason to assume that our treatment could be correlated with different levels of fatigue – as the questionnaires were identical for everybody – and we still want to see how our treatments influence both 'close' and 'distant' reversals[10]. Furthermore, the first item reversal refers to preferences related to strong majoritarian or minority governments; the other reversal is concerned about one given party's role in the coalition formation. Hence, we also have topical or difficulty related variation. A 5-scale (mid-point reflecting "neither") was used in the agree/disagree questions, and we use a conservative

[9] Krosnick's (1999) meta-analysis on 41 studies presents a correlation of -0.22 between mutually exclusive items.
[10] The first reversal is the closer one, being part of the same block of questions: 5.1 and 5.1. Both are in the 20% of the questions. The second reversal is between 5.8 and 8.2, where the second item is already over the half of the questionnaire.

measure for the consistency of reversals. This translates into a coding in which we also take into consideration the intensity of the expressed attitude. Thus, a strongly agree given for formulation *X* needs a strongly disagree for *non(X)*, in order to be considered consistent (and so forth). As this measure is used as a measure of data quality linked and introduced as inconsistency, we code our dependent variable as 1 if there is inconsistency, and 0 for consistent answers.

First, we report the correlations between the reversals. As pointed out before, they should be -1.0 minus measurement error, but previous research finds that, on average, they are around -0.22. Our reversals are somewhere in-between. The items of the close reversal are correlated -0.42 ($p<0.001$), whereas the items from the distant reversal -0.33 ($p<0.001$). These correlations already indicate that, indeed, we might see the impact of fatigue. For these two reversals we specify two logistic regression models, where the predictors are the previously employed group identifier dummy variables, with additional control for gender and age. These covariates are identical to the ones from Model 2, previously described. Table 4 reports the coefficients from these models (log-odds), with cluster-corrected standard errors (where clustering is given by groups) and significance levels.

Table 4: Model results for inconsistency

	Inconsistency 1 (close)	Inconsistency 2 (distant)
Female	0.005 (0.006)	0.23 (0.05)**
Age	-0.002 (0.003)	-0.0001 (0.002)
Altruistic	0.001 (0.05)	-0.02 (0.064)
Formal	0.11 (0.05)**	-0.02 (0.06)
Complex	0.15 (0.06)**	-0.11 (0.06)*
Not baseline	-0.06(0.15	0.189 (0.1)*
Intercept	0.48(0.07)**	0.93 (0.08)**
Log-likelihood	-2304.9	-1957.6
N	3505	3505

Note: ** is p<0.05, * is p<0.1. For both models, the likelihood-ratio tests indicate significantly better fit than the baseline model. Untransformed coefficients with cluster-corrected standard errors in parentheses.

As expected, the same model fits differently depending on the choice of inconsistency analyzed, granting further support to the theory that fatigue is an important determinant of this response quality problem. As we saw it previously in the case of correlations between the included items, our model explains much better the inconsistency between distant reversals. Furthermore, by comparing the intercept, we see an almost double default probability of distant inconsistency than close for the close version[11]. Even more interestingly, gender is not related to the inconsistency between closer items, but it is a strong predictor of the distant version, showing that women have higher odds of "committing this error", but only if there is a significant distance between the reversal questions.

[11] The intercept describes the log-odds for males with average age who received the baseline message.

Our messages also display different effects. Getting a formal or also a complex re-contact message influences positively the odds of close inconsistency, a result that is similar to the effects on response rates, and goes slightly against the impact detected for item non-response. Switching to distant inconsistency, the effects are weaker. In this case, being among the receivers of complex messages reduces the odds of inconsistency, but it is only remotely significant. Finally, being in any of the groups that did not get the baseline also increases the odds of distant inconsistency, in which case the results are again pointing to positive effects of baseline letters on not just response rates, but also on data quality.

6. Conclusions

This paper is still progressing at this point, however there are several interesting conclusion to be drawn:

There is ample evidence here that email letters' contents can substantially influence response rates – while we did not advance a superior formulation to the industry 'baseline', we did demonstrate that overly complex language patterns are damaging to response rates, and that, overall, altruistic appeal messages bring slightly higher response rates than egoistic appeal messages. We found little consistent evidence, however, to suggest that the formality or informality of letters exerted much of an influence either way.

The performance of the 'baseline' treatment was puzzling – on the one hand, the baseline letter secured the highest response rates. Furthermore, response rates were near identical for the two baseline groups, indicating that the baseline letter had some positive effect on response rates, relative to the other groups. On the other hand, our measures of data quality indicated that the baseline groups had the lowest quality data.

One speculation as to the superior performance of the baseline letters is that they did not explain the subject of the survey – coalition formation negotiations, whereas this topic was explained in the others. As such, it may be that several respondents in the 'baseline' groups may not have responded, had they known the survey topic. This may also help to explain why the baseline groups provided slightly poorer quality data than the others. As such, future research may find that vaguely phrased solicitation emails elicit higher response rates, but somewhat lower quality responses.

Appendix

Wordscores with R using the Austin package

	Score	Std. Err	Rescaled	Lower	Upper
ACI	6.29	0.208	14.83	13.59	16.079
ASF	5.73	0.257	11.29	9.76	12.825
ASI	5.36	0.278	8.97	7.31	10.634
Bas1	4.75	0.325	5.14	3.20	7.087
Bas2	4.87	0.374	5.91	3.68	8.149
ECI	3.81	0.188	-0.74	-1.86	0.385
ESF	3.11	0.322	-5.19	-7.11	-3.264
ESI	3.48	0.248	-2.83	-4.30	-1.346

Note: 158 of 274 words (57.66%) are score able. As for the previous analysis, ACF and ECF were used as reference texts.