

# Ensuring Data Integrity for Business Decisions

David Day  
Jenny Koo  
Birgi Martin



The use of online research through online access panels has increased dramatically. Today, online research is widely accepted as a sound method to evaluate, understand, and examine public opinion for diverse purposes. However, while using online panels can reduce the cost and time it takes to undertake research, there is often little understanding of how the quality of the panel affects data integrity and importantly the ability to make reliable business decisions based on the data collected.

This paper will identify some of the components of 'quality' in online data collection and the effect these can have on the data integrity. It will review how components of the online data collection process, particularly the invitation process, can have a dramatic impact on the actual quality of the data collected.

# Introduction

The growth of online research continues unabated with more and more people switching research from the more traditional methodologies to online. Helping to feed this growth is the increasing number of online access panels companies that are able to provide clients with large multi-national panels. These panels are able to provide clients with easy access to thousand, if not millions, of potential respondents for research.

While online panels can bring significant benefits to users, in terms of speed and cost, as well as the ability to reach complex target groups and use interactive stimuli, there seemed to be less awareness of some of the quality issues that could occur through using online access panels. Some questions that need to be asked include: Does it make a difference to the quality of the data collected using one panel compared to another, what actually affects the quality of the data collected and what are the best procedures that should be used?

Understanding the drivers of quality within access panels is critical for the continued development and growth of online research. If data collected is untrustworthy or even incorrect, then the future of online panels and even online research could be in doubt. This paper will review some of the drivers and the role of the panellist invitation process, in particular the effects the time a survey is sent can have on the data integrity.

# Defining an effective online panel

While there are many online panels in the market, how does a client know which is going to provide it with 'quality' data that it can trust to base its business decisions on? While all panel providers promise a good quality service and data that is reliable, how can a client understand these drivers of quality and the potential impact it can have on the data they receive?

ESOMAR recently published the "25 Questions to Ask your Panel Provider" as the first step to helping the users of panels to understand the various aspects of an online panel. These questions helped stimulate the discussion on what makes a good online panel and identified the core areas that can affect the 'quality' of an online panel.

The next stage of the discussion is to look at how the whole process of developing and using an online panel can impact the effectiveness of the panel and the quality of the data it provides. How can we ensure that the data a client receives is reliable and will allow them to confidently base their business decisions upon? In essence, how can we ensure an effective panel that provides quality data that clients can trust?

Ensuring data integrity can not just be achieved through one quick fix, it needs to be ingrained throughout the panel organisation. Every aspect of an online access panel has the potential to affect the quality of the end data a client receives, from the recruitment of the panellists to the survey invitation methods and the management of the fieldwork.

# Understanding the Interview process

Online access panels provide clients with a ready resource to undertake research and contact a range of pre-identified panellists. However the systems and procedures, from panel recruitment to the fieldwork, can impact on the data. Through reviewing the interview process, we can identify the impacts on data integrity.

## Profiling the panellist

The basis of any research should be an understanding of the characteristics of the people who are both being invited to undertake the survey, as well as those that partake. Online access panels often provide clients with the benefit of comprehensive panellist information already defined, but what is the impact on data integrity for clients if this data is incorrectly recorded or even wrong.

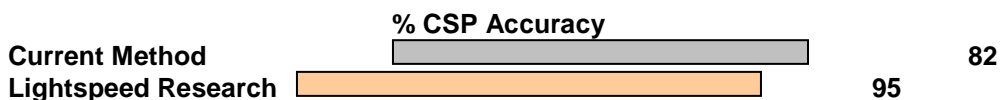
### Objective

Often information is gathered through asking the panellist to provide information about themselves, but there are occasions when it could be difficult for them to accurately provide a response. This could be due to them lacking the knowledge to make the decision or even not having an understanding of the complexities of the choices. Basing business decisions on this information could be detrimental to the client as the data might be representing the wrong type of respondents. To try and understand these complexities, Lightspeed Research undertook two studies to determine the accuracy self provided data had on the integrity of the profiling. The two studies covered Social Class and Habitat in France.

## Case Study 1: Social Class

While not all panels record social grade (some provide only income), when it is recorded it is normally calculated through the use of one or two standard questions. Due to the complexity of the social class classification system it can be difficult to accurately calculate a respondent's classification from asking simple set questions.

Lightspeed Research undertook an analysis to see how people had been classified on the French panel by social grade (SEC). The analysis was based on numerous tests carried out in accordance with the various levels of the INSEE socio-professional categories (Version PCS-2003) and asked respondents a combination of an open-ended and a closed question. This included asking the respondent to enter their exact job title. Once this information was collected the responses to the open-ended question were compared to the answers from the closed question to verify that the respondent has been classified correctly.



*Improvement based on Lightspeed Research's CSP verification process*

Overall it was found that approximately 18% of the respondents who answered, classified themselves incorrectly and consistently in the same categories. Through manually re-coding each respondent across the whole panel we were able to redefine the respondents more accurately to match the actual classifications and improve the reliability of the sample.

## Case study 2: Habitat

City size is an important selection criteria for undertaking online research studies, with a majority of European studies using habitat size as means of selecting respondents. If the panellists code themselves incorrectly then it is likely that the sample provided for research will be made up with the wrong people, potentially invalidating the research undertaken.

Lightspeed Research undertook an analysis of how accurately people classified themselves by city /town size. 31,000 French panellists were asked to supply their postcode, which was then analysed against the self selected city size response originally provided through the registration survey.

The study showed that when people were asked to identify the size of town/city they resided in, less than half of people accurately identified their habitat size. The accuracy of the self coding was shown to also be directly related to the size of the city or town they resided in. People that came from smaller towns (less than 20,000 inhabitants) tended to be the most accurate, with 90% of respondents correctly selecting their town size. However, as the size of the town increased the percentage of correct responses decreased, and in cities with over 200,000 inhabitants, a massive 78% of respondents incorrectly classified themselves.

### % of Panellists who correctly identified their Habitat size

	Correct %	Incorrect %
Less than 20,000 inhabitants	90	10
From 20,001 to 50,000 inhabitants	48	52
From 50,001 to 100,000 inhabitants	34	66
From 100,001 to 200,000 inhabitants	26	74
More than 200,000 inhabitants	22	78
<b>Total</b>	<b>47</b>	<b>53</b>

These two studies highlight the importance of understanding the role of the panellists in collecting data and how critical accurate profiling information can be on data integrity. Even the process for identifying and profiling panellists can be seen to impact the quality of the information collected. Without an understanding of the dynamics of panellists and how people in the panel respond to questions, there is a real danger that profiling questions, such as Habitat and Social Class, can become misleading or even inaccurate.

## Invitation wording

How panellists are invited to a survey can also have an impact on the integrity of data collected through online research. Even the very beginning of any survey, the invitation wording, can significantly impact the type of people that respond and the quality of the data collected. Lightspeed Research conducted a study to better understand how the invitation process could impact survey results.

### Objective

As previously published in 2005, Lightspeed Research conducted a study in order to understand the impact survey invitation wording could potentially have on survey results. A short survey was conducted on fishing and panellists were asked whether they fished, how often they went fishing and, if they did, which brands of rod they were familiar with and used. To qualify for this survey, respondents had to fish at least once a month.

A sample for the research was selected and divided into two groups. While the actual questionnaire taken by both groups was identical, the invitation process was altered. One group was pre-warned about the subject matter and the qualifying criteria, while the other group was not. The sample was balanced by gender and age across both invite segments.

		Non targeted		Targeted
	Count	%	Count	%
Aware of 'Real' Brand	371	91%	449	79%
Aware of 'Fake' Brand	38	9%	119	21%

The results clearly showed that when respondents were pre-warned regarding the subject matter and the qualifying criteria, the number of incorrect responses to the brand of rod they know and used increased. It can therefore be surmised that a reasonable number of the respondents did not actually participate in fishing. If this was the case, the results from this targeted sample would be potentially invalid as far as obtaining the opinions of people who fish, highlighting the importance of understanding the impact of the invitation process on data integrity.

## Invitation timing

With the continued pressure by clients and research agencies to shorten the length of time research takes, it is important to understand the dynamics of research to ensure timely and accurate fieldwork. There is often little understanding of the effect the time a survey invitation is sent on the type of people who respond and also the effect that fast responders can have on the data collected.

Lightspeed Research conducted a study in order to explore how the timing, as well as the length of time a survey is in field influenced respondents' participation and behaviour in online surveys, and to determine any potential impact that could occur.

## Objective

Lightspeed Research undertook a study to explore the impact the timing of survey invitations had on survey response rates. The specific objective of this research study is to understand how response rates differ if survey invitations are sent at different times of the day and/or on different days of the week and also if the data received differed for people who responded quickly compared to slower responders.

## Methodology

To undertake this study, Lightspeed Research selected a sample of 7,440 panellists from its GB panel which was then divided into 12 groups (620 each). Each of the groups were matched demographically to ensure consistency. While the questionnaire taken by each group was identical, the invitation process differed for each group, with each group being sent out at a different time. Survey invitation emails were sent to each group at 8:30am, 11:30am, 2:30pm and 5:30pm on Monday, Wednesday and Friday as shown in the table below. The survey was in the field for 6 days for each of the 12 groups.

## Survey Invitations

Day	Date	Time
Monday	20 <sup>th</sup> Feb	8.30am
		11.30am
		2.30pm
		5.30pm
Wednesday	22 <sup>nd</sup> Feb	8.30am
		11.30am
		2.30pm
		5.30pm
Friday	24 <sup>th</sup> Feb	8.30am
		11.30am
		2.30pm
	17 <sup>th</sup> Feb	5.30pm

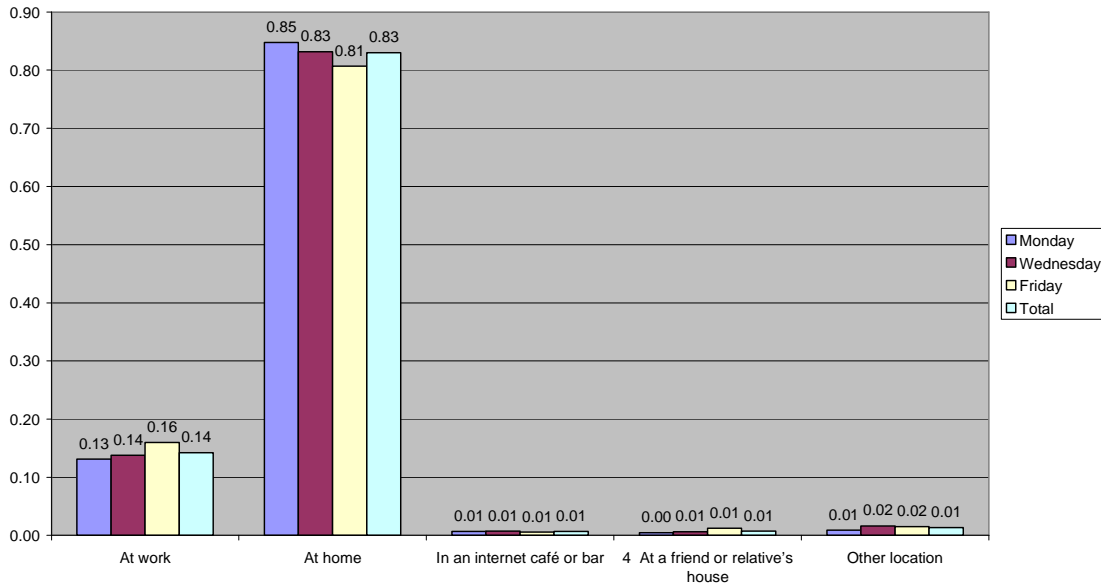
## Survey location

The majority of people who undertook the survey stated that they did so from home (83%), with only 14% stating they took the survey at work.

Location	Total
At work	14%
At home	83%
In an internet cafe or bar	1%
At a friend or relative's house	1%
Other location	1%

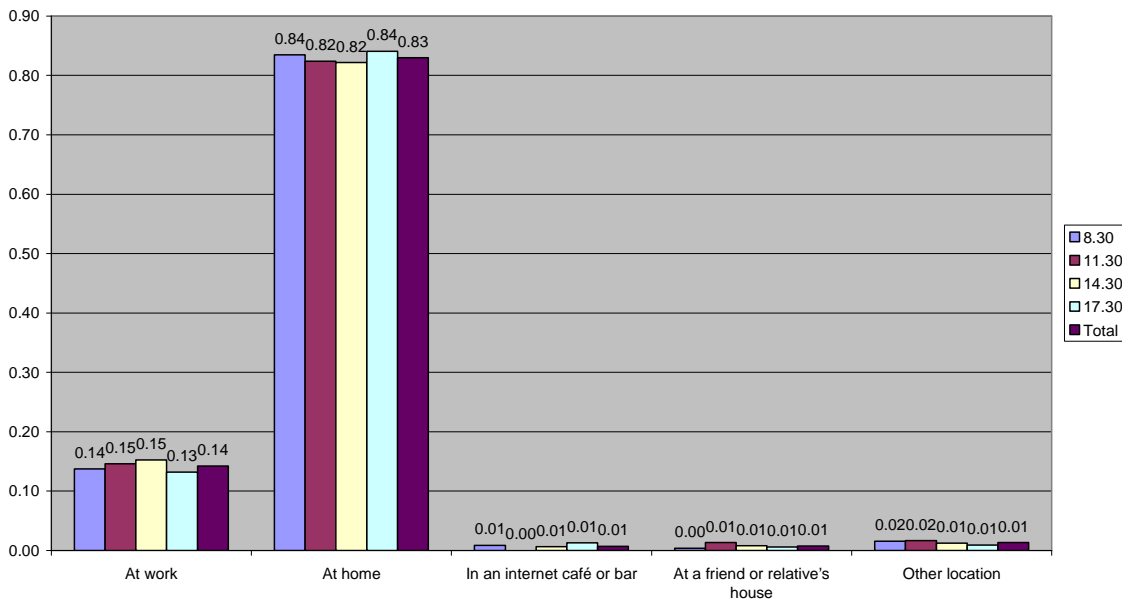
It was also highlighted that there was a slight tendency (but not significant) for people to take the survey at work when the invitation was sent on a Friday. The results actually show a steady (if small) increase in people taking the survey at work as the week progresses, which also corresponds to a slight decrease in people taking it at home.

Location of survey by day of invitation



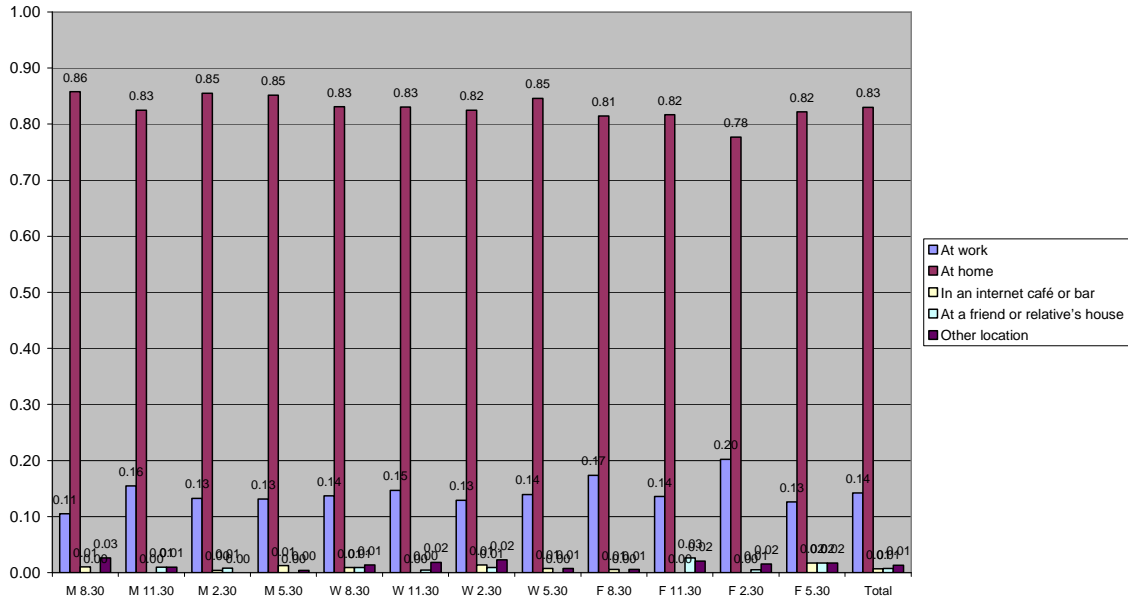
The location of where respondents undertook the survey was also assessed by the time the invitation was sent. There did not seem to be much variation by the time the invitation was sent out, with only a 2% variance recorded between the times.

Location of survey by time of invitation



When reviewing the location by batch, the highest proportion of people responded at home on invitations sent out on Monday at 8.30am (86% of respondents who responded), while the lowest occurred on the batch sent out on Friday at 2.30pm (78%). However the analysis showed that the percentage of people responding from home did not vary greatly by batch, with most of the batches having between 82% and 85% of respondents taking the survey at home.

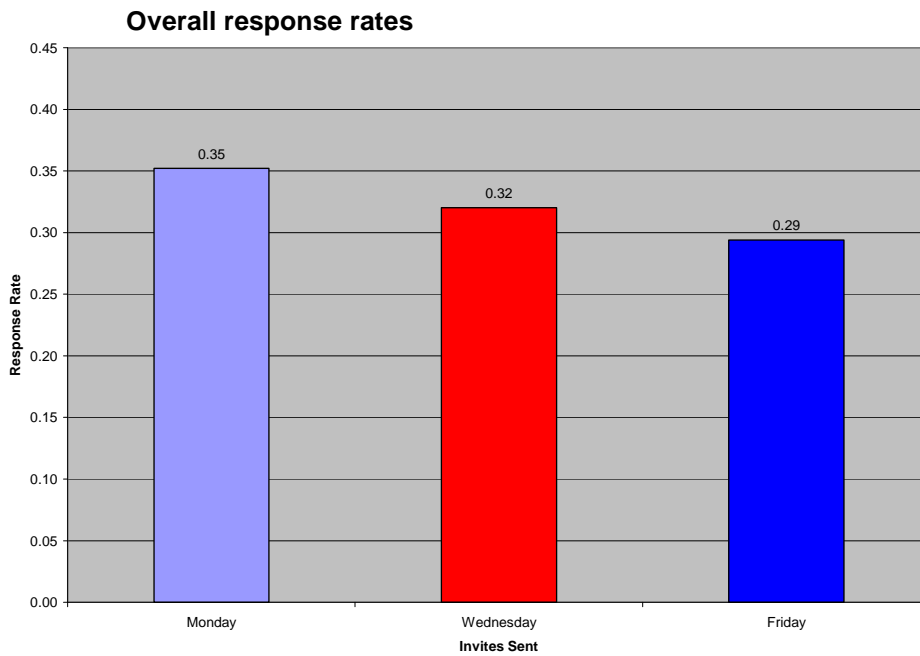
Location of response by batch



# Overall effect on response rates

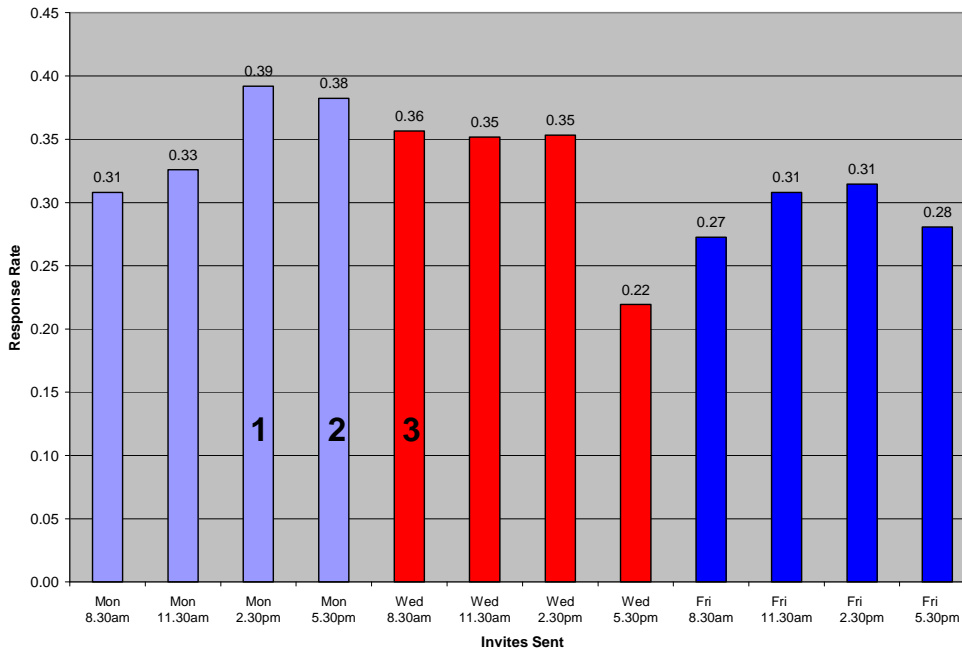
The survey was analysed in terms of overall response rates and also the response rates for different demographics groups within each time slot. The results showed a varying degree of responses by day and also time within days.

While the total response rate for the survey was 32% at the end of the fieldwork period (day 6) there were seen to be differing response rates by invitation time and day. The day that had the highest overall response rate was Monday, with a total response rate of 35%, followed by Wednesday (32%), while Friday had the lowest response rate (29%).



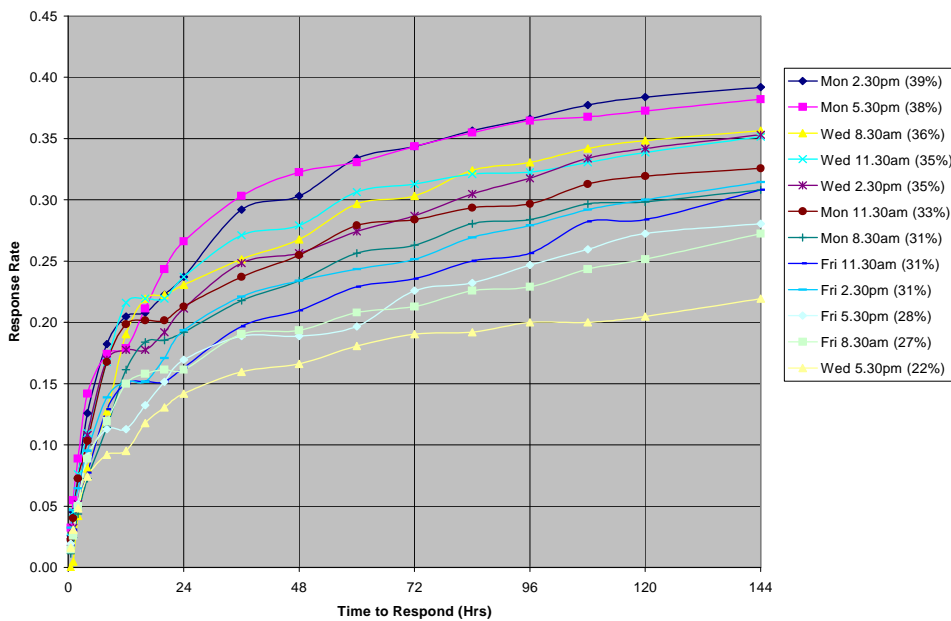
When reviewing the invitation batches it was seen that the Monday 2.30pm batch had the highest response rate (39%) of all the groups, followed by Monday 5.30pm (38%) and Wednesday 8.30am (36%). The lowest response rate by batch was seen to be Wednesday 5.30pm (22%), Friday 8.30am (27%) and Friday 5.30pm (28%).

## Response Rates – Total



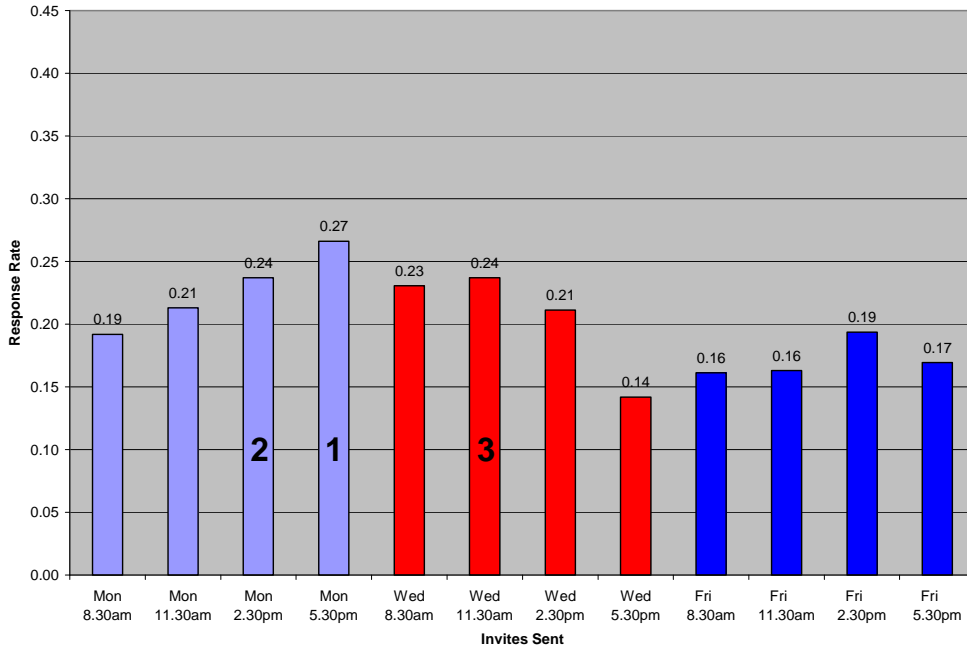
The responses were also analysed to see if there was a pattern for the speed of response (i.e. how long after the email invitation was sent did people actually respond to the survey). Overall 11% of all the respondents who responded to the survey did so in the first hour after the invitations were sent. After 2 hours this figure had increased to 18%, while 52% responded in 12 hours, and 75% within 2 days.

## Response Rates v Time to Respond

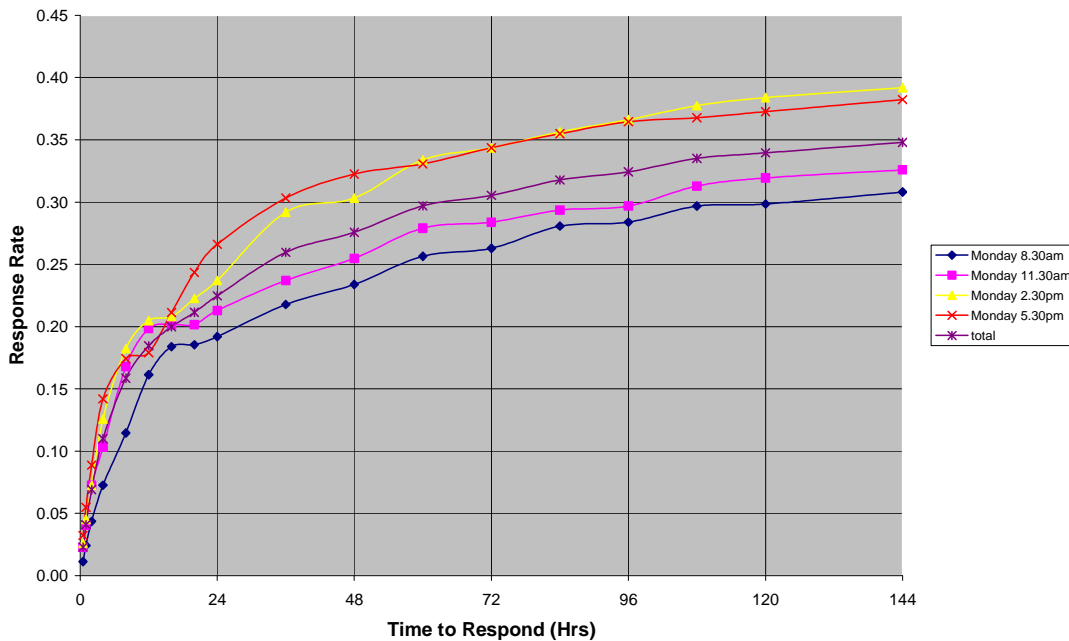


The response rate by batch was also analysed after the first 24 hours to determine if any particular day had a higher level of 'speedy' responses. At the end of the first 24 hours, Monday 5.30pm batch had the highest response rate (27%) followed by Monday 2.30pm (24%) and Wednesday 11.30 (24%), which while similar did rank slightly differently from the final results at the end of the study. The lowest batches were Wednesday 5.30pm (14%) and Friday 8.30am, and 11.30 am (16%).

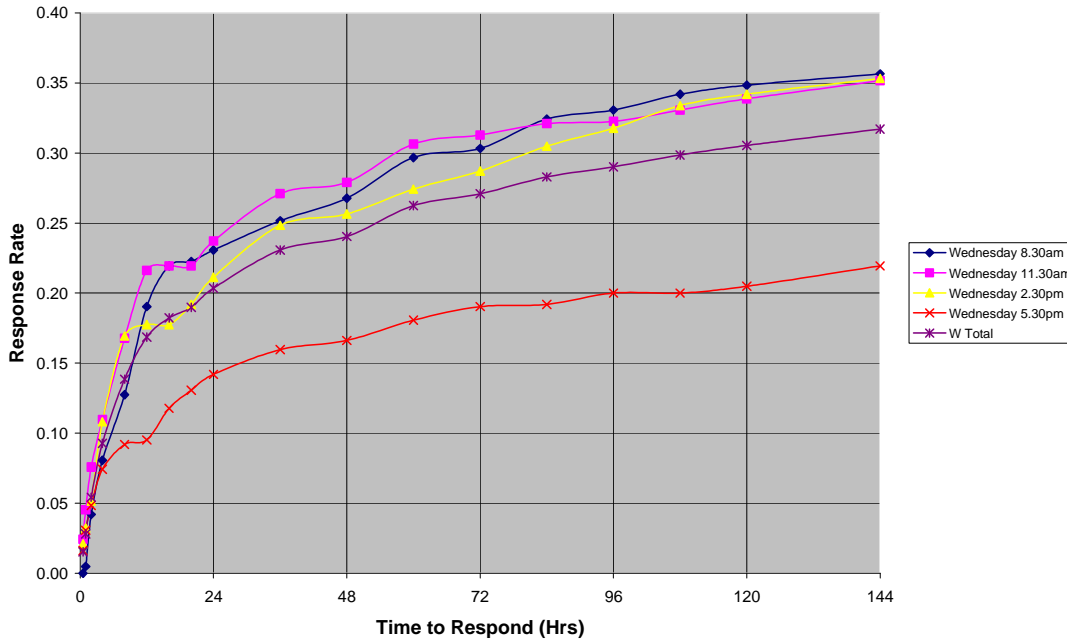
**Response Rates – First 24 Hours**



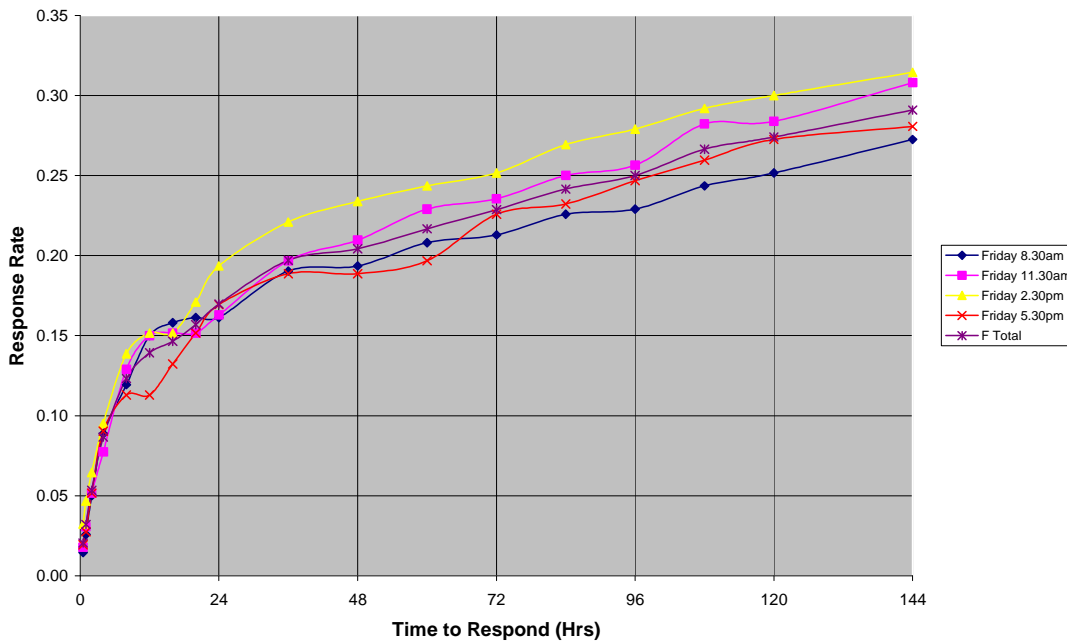
**Monday**



### Wednesday



### Friday



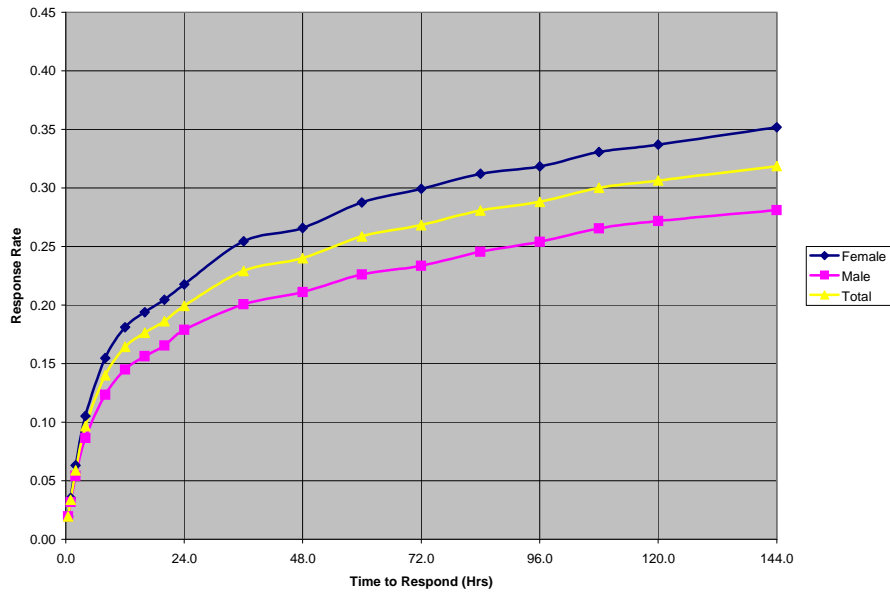
### Response rate by gender

The response to the invitation was analysed by gender to see if there was any disparity between males and females, both in terms of the level of response but also the speed at which they responded.

At the end of the study, females turned out to be better responders to the survey with the overall response rate for females being 35% against 28% for males. Within the first 24 hours, females had a 22% response while males

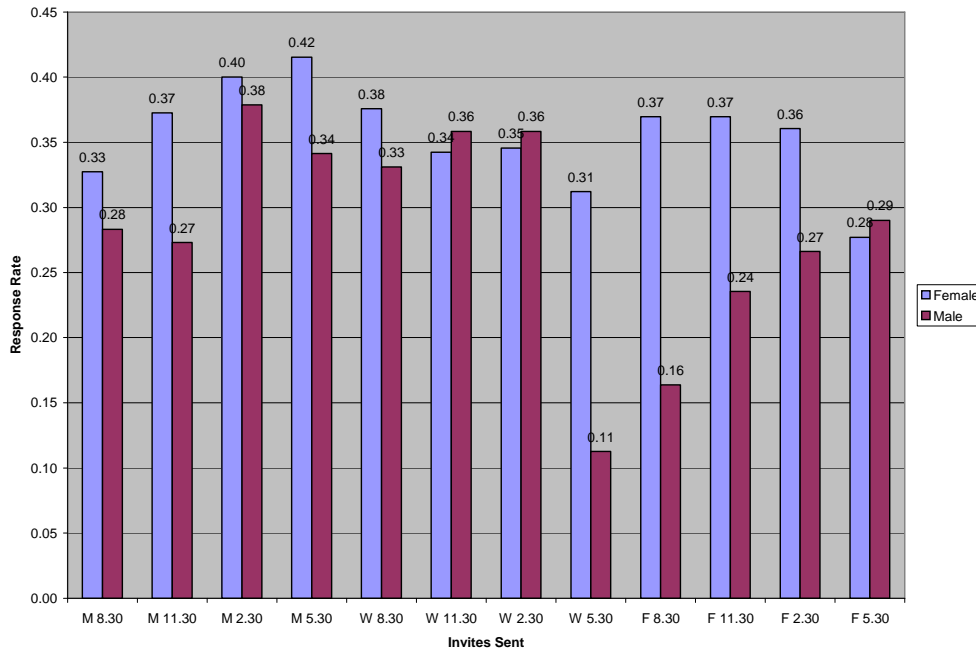
were at 18%. The only time that the response was the same for males and females was within the first 30 minutes. From that point onwards, a gap opened up between the response rates of female and male respondents.

### Response Rates by Gender



The results were again analysed to determine the differences in response rate by individual batches. It was found that females had the best response rate when invitations were sent out on Monday at 5.30pm (42%), while the best time to invite males was Monday 2.30pm (38%). The worst times for inviting males was shown to be Wednesday at 5.30pm (11%), while females had the lowest rate on Fridays at 5.30pm (28%) (see graphs below).

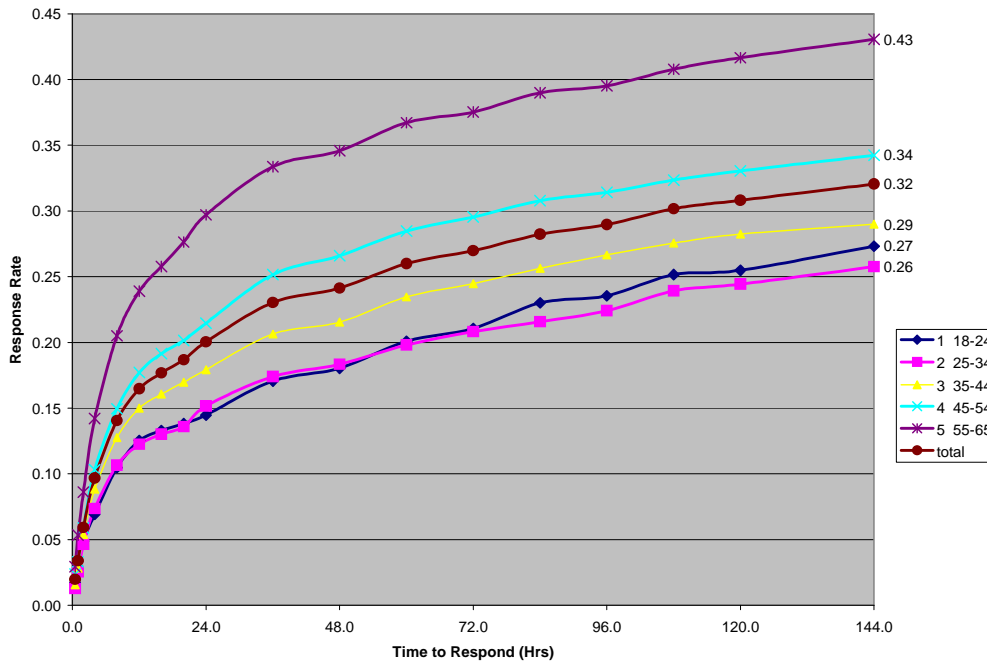
## Response Rates by Gender/Batch



## Results by Age

When analysing the data by age, it was seen that 55-65 year olds were the best responders, both within the first 24 hours (30%) and after the full 6 days (43%). The lowest response rates came from the 18-24 and 25-34 year olds. Within the first 24 hours 18-24 year olds exhibited the lowest response (14%) closely followed by 25-34 year olds (15%). After 6 days, 18-24 and 25-34 year olds are still the worst responders but this time 18-24 year olds are marginally ahead of the 25-34 year olds (27% compared to 26%).

## Response Rates by Age



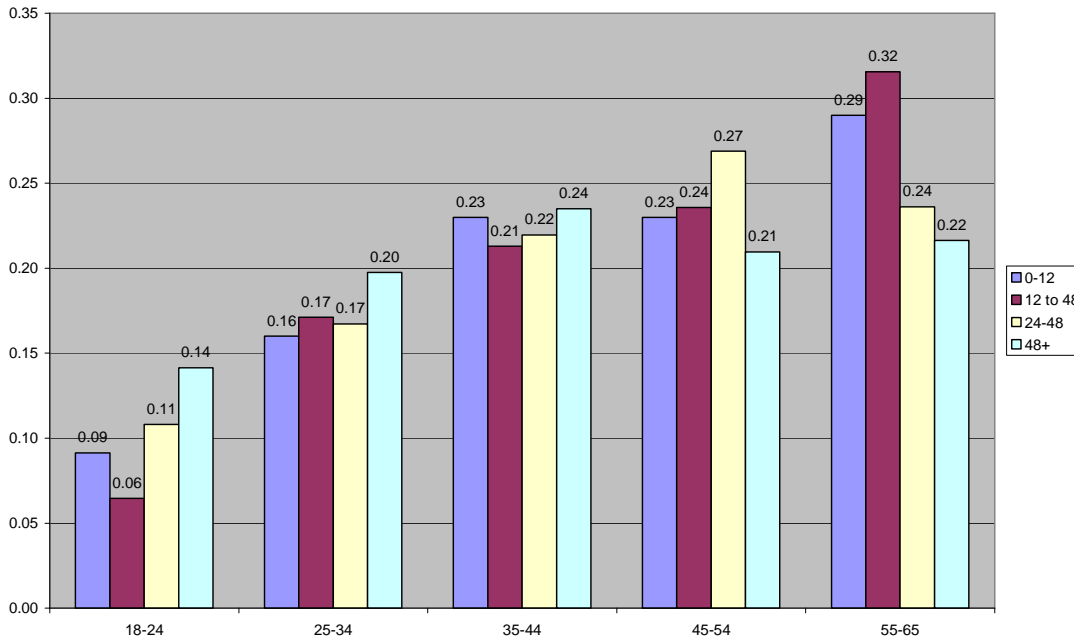
The response rate by age group was again analysed by batch to determine if there was a pattern to the best times to send out invitations. While Monday afternoon was seen to provide the best response rates for most of the age groups, the 25-34 year olds had the best response from invitation sent out on Friday at 11.30pm, and the 55+ on Wednesday 2.30pm.

Age groups	Highest response rate	Lowest response rate
18-24	Monday 2.30pm	Friday 5.30pm
25-34	Friday 11.30am	Wednesday 5.30pm*
35-44	Monday 2.30pm	Wednesday 5.30pm*
45-55	Monday 5.30pm	Wednesday 5.30pm*
55+	Wednesday 2.30pm	Wednesday 5.30pm*

\*This coincided with a major European sporting event and this is in part responsible for the low response rate.

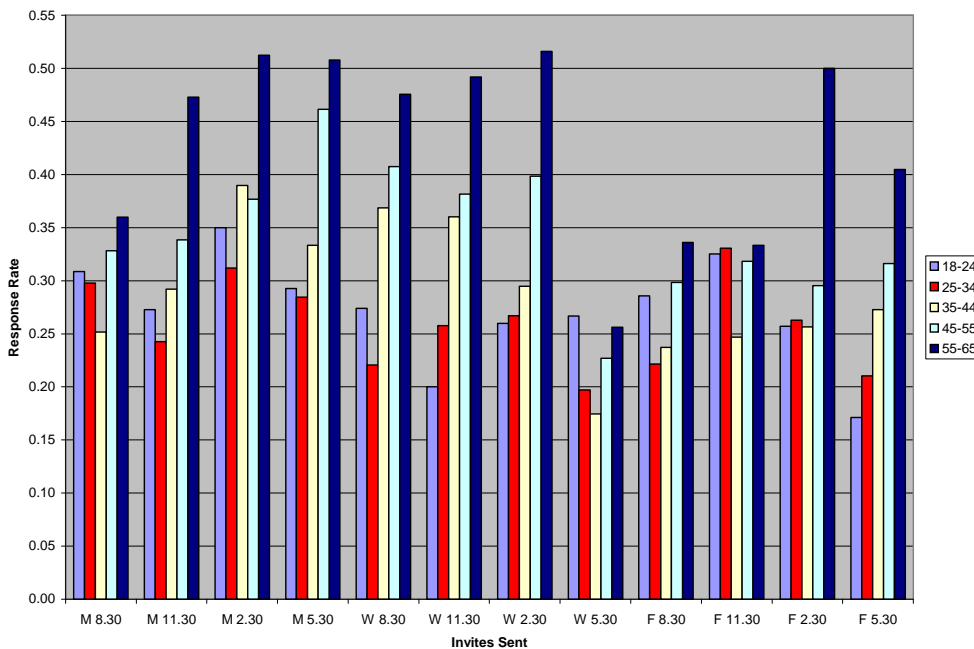
It is key to note that the 18-24 year old panellists are less likely to respond to an invite on a Friday evening than on a Wednesday evening.

## Time to respond by age



The time to respond also shows an interesting difference between younger responders against the older responders. While 18-24 year olds overall had a lower response rate the ones that did respond tended to take longer than the 55-65 year olds. For example 18-24 year olds have a nine percent response rate between 0 and 12 hours after the invitation is sent, rising to 14% for 48 plus. The opposite trend can be seen for the 55-65 year olds, with a response rate of 29% in the first 12 hours dropping to 22% for 48 hours plus.

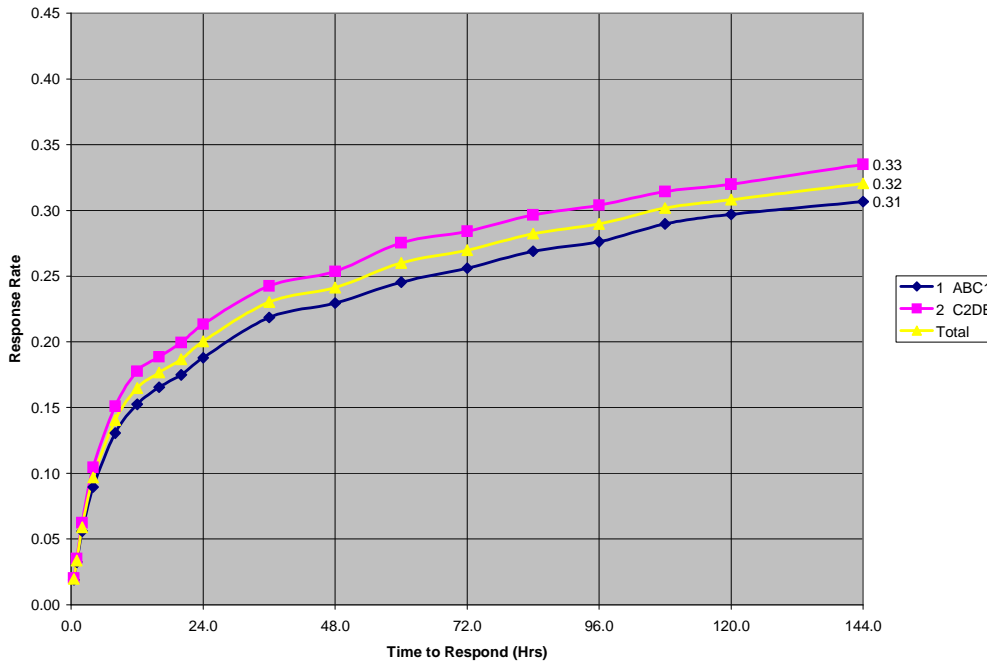
## Response Rates by Age/Batch



## Results by SEG

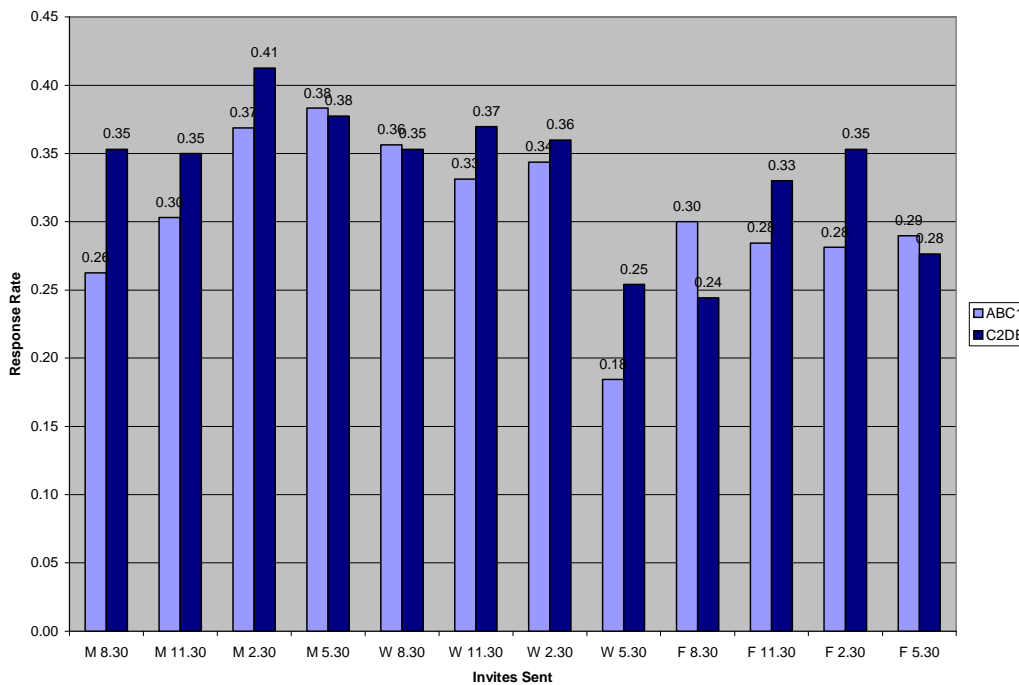
When the responses were analysed by social grade, there was not a great deal of difference between the respondents in the ABC1 social group compared to the C2DE. However respondents in the C2DE group are slightly better responders than those in the ABC1 group. Over the first 24 hours C2DEs had a response rate of 21% compared to 19% for the ABC1s. At the end of the study, C2DE's had a 33% response rate against 31% for ABC1s.

**Response Rates by SEG**



When analysing by batch, the best response rates for ABC1 was seen to be from invitations sent out on Monday 5.30pm (37%), while for C2DE's was Monday 2.30pm (41%). Conversely the worst responses came from invitations sent out on Wednesday 5.30pm (ABC1 – 18%) and Friday 8.30am (C2DE – 24%).

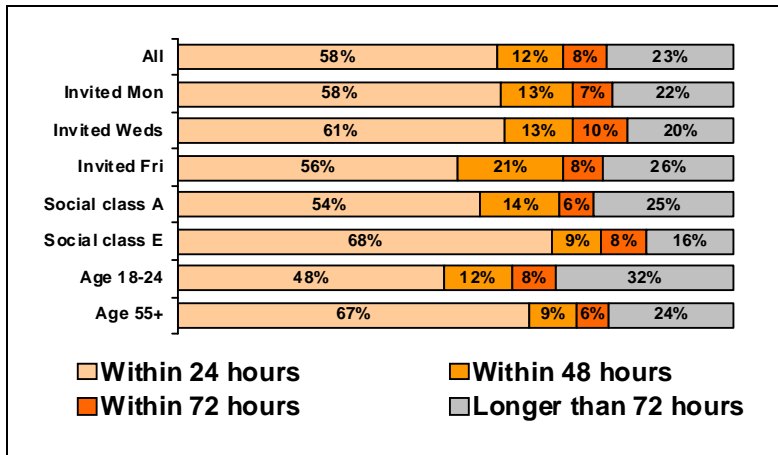
**Response Rates by SEG/Batch**



## Conclusion

Previous work Lightspeed Research has conducted has shown the importance of the classifying questions when profiling online consumer panels with respect to social class and location. This research has been designed to focus on the impact of the timing of launch of field work on both the overall response rate and therefore the feasibility and duration of the fieldwork. It also indicates the importance of quota controls as response rates do vary both by demographic and timing of fieldwork launch.

This study shows that launching fieldwork early on in a working week enhances overall response rates and quota achievement. Overall, response rates fall the later in the week the field work is started. This is because the first 12-24 hours are the most important in any online project with approximately 2/3 of panellists responding in this period. It was also found that extraneous events can lead to very significant temporal reductions in response rates which, given the importance of the initial 24 hours, can significantly (if not permanently) delay the completion of the study. It was further shown that the demographics of the early completers are different from the overall respondent profile with older age groups indexing higher.



In the normal course of events, it appears that launching online fieldwork on a Friday afternoon will render the lowest response rates overall. This would appear to be the effect of a reduction in response rates at the weekend - though this was not explicitly studied here – as many panellists are unavailable to complete the study during the golden 24 hours as they are otherwise occupied. This does suggest that if the time for fieldwork is short, then deploying studies later in the week should be avoided, and this does have implications for certain industries that favour weekend completion. It also does suggest that it is important to understand if the ‘missing’ responders have a different set of characteristics to those available to respond beyond the simple demographic differences outlined in this work which can be managed using quota controls.