Qualitative Testing for Official Establishment Survey Questionnaires

Mojca Bavdaž – University of Ljubljana, Faculty of Economics, Kardeljeva pl. 17, SI-1000 Ljubljana, Slovenia. mojca.bavdaz@ef.uni-lj.si

Deirdre Giesen – Statistics Netherlands, Postbus 4481, 6401 CZ Heerlen, The Netherlands. d.giesen@cbs.nl

Danna L. Moore – Social and Economic Sciences Research Center, Washington State University, Pullman, Washington, WA 99164-4014, USA. moored@wsu.edu

Paul A. Smith – S3RI, University of Southampton, Highfield, Southampton, SO17 1BJ, UK. p.a.smith@soton.ac.uk

Jacqui Jones – Australian Bureau of Statistics, Locked Bag 10, Belconnen ACT 2616, Australia. jacqui.jones@abs.gov.au

ABSTRACT

Best practice research for testing establishment (businesses and other organisations) survey questionnaires is largely the province of official statistics and has developed more slowly than the corresponding research in household surveys. With a focus on the development and testing of establishment survey question(naire)s, this paper: reviews what we know; makes recommendations; reports survey results on the practical application in National Statistical Institutes; and assesses the levels of maturity in the application of approaches.

KEYWORDS

questionnaire testing, best practice, establishment surveys, business surveys, cognitive testing.

1. INTRODUCTION

Qualitative research is used to study *'things in their natural settings, attempting to make sense of, or interpret, phenomena in terms of the meanings people bring to them'* (Denzin & Lincoln, 2005, p. 3). In the survey context, qualitative research methods shed light on the meaning people bring to survey questions and how they interact with the questionnaire. They explore the response process in order to identify measurement errors and explain how and why these errors occur. In establishment surveys the response process evolves at two levels; people take part in organisational processes while going through their own cognitive processes (Bavdaž, 2010a; Edwards & Cantor, 1991; Lorenc, 2006; Sudman, Willimack, Nichols, & Mesenbourg, 2000; Willimack & Nichols, 2010).

Cognitive research methods use cognitive processes and models to gain in-depth understanding of particular issues (Blair & Presser, 1993). These methods are widely applied to the development and testing of all kinds of surveys although their conduct may vary with regard to the type and sensitivity of questions, administration mode, and target population (Willis, 2005). Palmisano (1988) from the US Bureau of Labor Statistics was among the first to report on the application of cognitive research methods to establishment surveys. Other early examples from official establishment surveys include DeMaio &

Jenkins (1991) for the US Census Bureau; Bureau (1991) for Statistics Canada; Eldridge, Martin, & White (2000) for the UK Office for National Statistics; Snijkers (2002) for Statistics Netherlands; and Davidsson (2002) for Statistics Sweden.

Cognitive interviewing is one of the main qualitative methods used for the testing and evaluation of data collection instruments. This is because the knowledge of cognitive processes used in answering survey questions is the first step in determining questioning strategies which lead to more accurate answers (Forsyth & Lessler, 1991). However, survey questions do not operate in isolation and may activate a network of associations beyond the intended question content, which can affect the survey response, and calls for an evaluation of survey responses in a broader context (Gerber, 1999). With this in mind, survey qualitative testing often investigates how respondents relate survey questions to their experiences, circumstances and sociocultural contexts (Miller, 2011). To study the broader context in which the survey questions are posed, cognitive interviewing often includes expansive probes (Beatty, Schechter, & Whitaker, 1997; Willis, 2005) and ethnographic interviewing (Gerber, 1999; Willis, 2005). The broader context is especially relevant in establishment surveys where individuals perform the survey task in an organisational setting and contribute to an organisational response, mainly on abstract economic and business concepts. The follow up on concepts and organisational setting can be exploratory, unstructured and take a substantial portion of interviewing time, thus adding many elements of an in-depth interview to cognitive interviewing. Moreover, the growing use of web questionnaires contributes some elements of usability testing that are practically inseparable from cognitive aspects (Blake, 2015).

In official statistics, some sort of questionnaire pretesting has been considered indispensable for new or revised questions or other changes to a data collection instrument (European Statistical System Committee, 2011; Office of Management and Budget, 2006). Using some of the range of qualitative methods is considered good practice as it allows us to gain more understanding of how elements of the data collection design (e.g. an introductory letter, a survey item, a questionnaire) work. Known relationships between the measurement errors, causes of these errors, resulting effects on reported answers and, ultimately, data quality serve as a necessary foundation for providing suggestions on improving survey and questionnaire design, and can also be insightful in themselves, for example, in the interpretation of survey results.

How National Statistical Institutes (NSIs) implement qualitative testing of establishment surveys is largely undocumented. The community of researchers working with establishment survey questionnaires is small and largely based in NSIs, with a few academics. The regular international Business Data Collection Methodology workshop (initiated in 2006) provides a forum for discussion and exchange of information but involves a relatively small number of organisations and does not systematically address the penetration of qualitative testing methods. To fill this gap, we designed and ran an international survey that collected information from NSIs on how widely these methods were being put into practice in the development of new and/or existing business or establishment survey questions. We focused on qualitative methods suitable for self-administered questionnaires and involving direct contact with businesses, namely interviews, focus groups, observations, and record keeping and usability studies.

In the paper, we set out the challenges which are specific to establishment surveys (Section 2), then review the literature for qualitative studies in questionnaire development, testing and evaluation of establishment surveys (this literature is quite scattered, and although we have searched extensively, there may be further examples; in a few cases we have used research from population surveys as the basis of our deliberations), with the focus on cognitive interviewing, and derive a list of recommendations (Section 3). We then contrast these recommendations with the reality of 32 NSIs that responded to the International Survey of Qualitative Testing Practice for Business and Establishment Surveys (Sections 4 and 5). We conclude by discussing the implications for NSIs (Section 6).

2. CHALLENGES IN QUALITATIVE TESTING FOR ESTABLISHMENT SURVEYS

Compared to quantitative research, fewer hard guidelines appear to exist on what are considered good practices in qualitative research, and the range of available techniques is wide. Unsurprisingly, the same

is true for how these qualitative methods should be used when testing and evaluating data collection methods for surveys. However, in the last decade we have seen a few papers and books that provide guidelines, especially on cognitive interviewing (e.g. Collins, 2015b; Economic and Social Commission for Asia Pacific Region, 2010; Miller, Willson, Chepp, & Padilla, 2014; Office of Management and Budget, 2016; Willis, 2005). Most of this literature focuses on surveys of households and individuals. Applying these guidelines to establishment surveys is not straightforward because establishments and establishment surveys have specific characteristics (see e.g. Cox & Chinnappa, 1995; Rivière, 2002; Snijkers & Bavdaž, 2011) that considerably influence the response process (Willimack, Lyberg, Martin, Japec, & Whitridge, 2004). However, Willimack (2013) provides a good overview.

To begin with, respondents in surveys of households and individuals typically answer questions about themselves generally based on available information that can be retrieved from their memories, while establishment surveys require a person to speak on behalf of an organisation and access information often in organisational systems. As indicated in the Multidimensional Integral Business Survey Response (MIBSR) model (Bavdaž, 2010a), the response process evolves at two levels: at the individual level people involved in the survey response engage in mental processes as they go about attempting to comprehend and answer (or support in some way the answering of) the survey questions; and at the organisational or business level, the implementation of the survey task is organised, authorised and provided with information support. The main differences from household surveys concern involvement of several people with different roles in the survey response (e.g. response coordinator, data provider, authority), retrieval of necessary data from the organisational business records, different impact of individual units on population estimates, and, in the case of official statistics, the mandatory and recurring nature of surveys. Testing of establishment survey questionnaires thus has to address specific questions, e.g. whether the instrument design and communication work for all relevant actors (e.g. external accountants), whether the requested data already exist in the business records, or can be derived or estimated from the available data, and whether expectations, policies and procedures on surveys are in place within the establishment.

Establishment surveys often measure technical concepts with precise definitions. Because of this, data collection is dominated by self-administered modes, and many instructions accompany survey questions. Studies on the content and quality of business records may be necessary before drafting the questionnaire, and content matter knowledge is needed for development and beneficial for testing. Furthermore, finding the respondent who knows most about the requested data is important but can be difficult, and testing procedures may improve identification of the correct reporter for establishments of varying types for the actual survey. Besides the nature of the data, testing procedures have to take into account the burden that testing imposes on the organisation, and that completing a questionnaire can be very labour intensive and difficult, if not impossible, to fully observe or replicate (Giesen, 2007; Willimack*, et al.*, 2004).

Therefore, we propose that a more tailored set of good practice guidelines is needed for establishment surveys, and that their use would improve testing of question(naire)s and therefore their design, particularly in organisations where there is little knowledge of these approaches (as demonstrated in section 4 below). It is possible to design questionnaires effectively without guidelines, but the codification of knowledge provides both a solid foundation for the development of expertise, and a standard against which an organisation's (usually an NSI's) practices can be benchmarked.

In the next section we identify a set of recommendations for testing establishment survey question(naire)s. The goal of these is to increase the validity, credibility and generalisability of the qualitative research results, and ultimately to ensure that survey questions are understood consistently and in the way the researcher intended, and to establish that respondents are willing and able to provide the information being sought. The qualitative research methods, and the recommendations drawn from them here, are broadly accepted and widely used amongst survey researchers and questionnaire developers, and are, therefore, accepted as best practices. There are, however, few examples of quantitative assessment of the effects of qualitative analyses. Therefore the selection of recommendations has inevitably been partly based on our judgement of their merits.

3. THEORY AND RECOMMENDATIONS FOR TESTING ESTABLISHMENT SURVEY QUESTION(NAIRE)S

In this section we review the literature to identify recommendations for conducting qualitative studies in questionnaire development, testing and evaluation of establishment surveys. The review follows typical research stages: overall design, sampling, recruitment, data collection, analysis and reporting.

3.1 Overall design

The use of qualitative research methods necessarily involves interaction with establishments and causes additional burden. Burden represents a constraint to testing and leads to careful selection of when and what to test (Willimack, *et al.*, 2004). Major changes in statistics production – be it new or revised topics – are typically occasions when the merits of qualitative research methods cannot be overlooked.

Working qualitatively with small samples from very heterogeneous populations to inform quantitative research might be particularly challenging. One potentially useful strategy in this situation is triangulation – checking the consistency of findings generated by different data collection or analytical methods, comparing data from different sources and times, and comparing theories or perspectives from different traditions and positions (Patton, 1999); where qualitative and quantitative methods are combined we have mixed research (e.g. Baena & Padilla, 2014). Using multiple sources of evidence has also been recommended to shed light on the response process from different perspectives. In an establishment survey this could be from a description given by the people involved in the response process in the business and the statistical organisation, from an observation of the response process, and from several experts with different experiences (Bavdaž, 2009). Persson, Björnram, Elvers, & Erikson (2015) suggest that methods based on individual judgement should be combined with empirical methods, and qualitative research methods should be combined with quantitative ones as part of a general risk-based questionnaire-testing strategy.

Using different testing methods is typical for major redesigns of official establishment surveys. In such projects, different testing methods have been found to provide both corroborating and complementary findings (e.g. Giesen & Hak, 2005; Tuttle, Morrison, & Willimack, 2010). As an example, cognitive interviewing was found to be complementary to expert appraisal (Forsyth, Weiss, & Miller Anderson, 2003; O'Brien, Fisher, Goldenberg, & Rosen, 2001). Conducting the interview immediately after the observation minimises the elapsed time between the actual response process and reporting about it and enables a comparison of the interviewee's and researcher's perceptions of the same issue, such as the invested effort and time (Bavdaž, 2009). For complex data collection from establishments, combining diverse methods can provide better and more insightful results than any one type of testing method in isolation. Further applying methods sequentially builds evidence and reliability of results as each successive methodology (e.g expert review, cognitive interviews, field tests, or field tests with experiments) gains in intensity and breadth, while iteration within the application of each method allows for retests of changes and can provide complementary information that can be used with more confidence for finalising questionnaires (McCarthy *et al.*, 2018); the approach can also be used for evaluating changes to existing question(naire)s (Jones, 2003).

Even in question testing for surveys of individuals, the comparisons have not been numerous, but support the use of multiple methods, e.g. combining observation and interviewing (Gerber, 1999), cognitive interviewing with behaviour coding (Stapleton Kudela, Forsyth, Levin, Lawrence, & Willis, 2006) or eye tracking (Neuert & Lenzner, 2016), etc. Some studies found one approach to be more productive than the other one, e.g. face-to-face cognitive interviews compared to debriefing questions by mail (Davis, DeMaio, & Zukerberg, 1995) or telephone reinterview compared to unmoderated, online cognitive testing (Mockovak & Kaplan, 2015). Any divergent results about identified question problems (e.g. Yan, Kreuter, & Tourangeau, 2012) remind us that much of the data in qualitative research is context bound, which makes generalisation difficult (Miller & Fox, 2004). Successively applying different methods with increasing intensity enables identification of the pervasiveness or implications of previous findings for measurement problems, enabling researchers to prioritize results and recommendations (Tuttle *et al.*, 2010). More important than perfect convergence is to understand the origin of inconsistencies (Patton, 1999). Using multiple qualitative methods together – a multi-method approach in the sense of Roller & Lavrakas (2015, p. 288) – helps overcome the individual weaknesses

of each testing method (McCarthy, *et al.*, 2018). This is distinct from mixed research which combines qualitative and quantitative methods.

Diagnosing problems with survey questions generally does not provide direct solutions to them; the burden of interpretation lies on the researcher's shoulders (Groves, 1996). Solutions intended to eliminate the problems should also be tested. Development, testing and revision should preferably be an iterative process (Brancato *et al.*, 2005). Whenever feasible, quantification with field and experimental tests should provide evidence of the effects of questionnaire changes (McCarthy, *et al.*, 2018).

With respect to the overall design, we thus propose for establishment surveys:

Recommendation 1: Pre-test new components of a survey with appropriately chosen qualitative methods to get insights into complexities of underlying cognitive and organisational response processes

Recommendation 2: Use more than one testing method

Recommendation 3: Test iteratively

3.2 Sampling and Recruitment

Sampling and recruitment are the first implementation steps in a qualitative research study. Because the goal of qualitative research is to obtain in-depth, detailed information, studies typically have relatively few cases. There is often some information on the types of cases which can be recruited, but it is not clear how to use this information to best effect in designing a qualitative study. Most of the relevant literature uses population survey examples, but two characteristics of establishment survey populations prevent their direct application, namely strongly skewed population distributions and detailed classifications (Rivière, 2002). Tourangeau (2004) reviews experimental design approaches to qualitative testing, which aim to balance samples over some known characteristics and provide a framework for producing generalisable results. However, practical considerations (such as cost) along with the need to target cases with specific characteristics, limit the number of units that can be investigated, which often has a strong influence on what is actually done. Nonetheless, balance is an important principle.

Guidelines on the sample size and composition for qualitative research (mostly rendered for population surveys) are generally qualitative as well, and even where indications of size are given (a review is given by Guest, Bunce, & Johnson, 2006) they need to be adapted critically to the particular research approach. Nielsen and Landauer (1993) found that (for over 11 examples) the detection of usability problems by different testers is well modelled by a Poisson process. Using their approach, when the rate of issue identification is estimated from the first few testers (cognitive interviews in our context) this can allow an assessment of the number of interviews needed to identify a given proportion of the total issues.

One common guideline is to continue collecting data until saturation is reached. Saturation, however, is difficult to justify (Charmaz, 2005), and its definition varies according to the research (O'Reilly & Parker, 2013). Guest *et al*. (2006) characterise saturation as no new data, and at a higher level no new themes, and measure it by following the development of coding of an interview study. We can translate these characterisations for testing questions and questionnaires as: continue testing until no new insights into the cognitive and organisational response processes are obtained, and no new problems with questions and questionnaires emerge. In this article we mainly focus on the second part – the target of the procedures is to identify issues with the questions and questionnaire, and the cognitive and organisational processes are mainly of interest for how they help us to do this. General research on response processes and models obviously makes use of the first part, but is not usually the goal of cognitive interviews in NSIs.

The idea of saturation also contributes to guidelines for testing questionnaires. Beatty & Willis (2007) suggest covering as much of a questionnaire's conceptual terrain as possible, exploring as many paths as possible when skip patterns are used, covering a variety of circumstances relevant to the topic and thus also getting some demographic variety. Collins & Gray (2015) call for a full evaluation of the test questions by recruiting participants who reflect the target population and by including a variety of different kinds of participants. The aim of attaining saturation thus implicitly covers both sample size

and sample composition. In the context of experimental design Tourangeau (2004) suggests that a sample should be chosen to cover a range of characteristics. Bavdaž (2009), discussing application to establishment survey questionnaires, recommends focusing primarily on a variety of establishment sizes but paying attention to other organisational characteristics likely to influence the response process (e.g., kinds of economic activity, outsourcing of the survey task, legal form of operation, origin of capital, geographical location, group participation, involvement in international activities, organisational culture). She also calls for inclusion of both respondents new to the survey task and those familiar with it. NSIs typically have data on business size and economic activity as well as some aspects contributing to organisational complexity (e.g. number of geographical locations etc.), and they typically use these characteristics in sampling designs. Databases with information on business response behaviour and respondent characteristics are expected to offer more sampling characteristics.

In exploratory interview studies, the sample sizes tend to be small, i.e. 'around 15±10' (Kvale, 1996). When aiming for variety, deliberate choices are made about what specific characteristics to pursue when selecting establishments, thus actually employing purposive sampling. Such an approach assumes that researchers have thought about sampling criteria, namely the characteristics that should vary in the sample and how they might influence the findings. Other aims are also possible. Willimack (2013), for example, mentions targeting key establishments for exploratory studies, though the characteristics that make a respondent 'key' are left to the researcher.

Several sampling procedures may be used in qualitative studies. Random sampling from a suitable population, even for rather small sample sizes, has the benefit of representativity (a concept with multiple facets, Kruskal & Mosteller, 1979), so that the results are generalisable by virtue of the randomisation mechanism, although for small sample sizes the variability will be large. However, random selection does not allow easy control of costs (e.g.travel), and in a situation where recruitment rates may be quite low (e.g. Ursachi & Jones, 2005), these costs may be substantial. Quotas provide control over defined characteristics of sampled businesses, and are easily implemented, particularly if characteristics are available from a frame. They provide some pseudo-randomisation, but may miss important features, particularly those associated with businesses which do not participate. In purposive sampling, units are selected based on particular, specific characteristics, usually related to questionnaire features that need testing in order to meet research goals. Although this is not a randomised procedure, so that results cannot be generalised to a broader target population, a purposive sampling strategy offers efficiency. Likewise, when the need to control costs is very strong, a convenience sample may be used.

The balance between a detailed classification of the population to be sampled and the use of replication (which improves generalisability as it makes conclusions more robust to unusual observations) is a delicate one. There are typically more characteristics available than can easily be incorporated into a sampling scheme, but including as many as possible guarantees that the sample (which may be rather small) is well spread (or balanced) over the used characteristics. On the other hand, replication may allow a variety of other characteristics, not available on the frame, to be covered. Recruitment may be more difficult for random samples, because willingness is either not included in the design, or included in such a way that unwilling businesses are also included in the design. If willingness to participate is related with other business features, including businesses which are harder to recruit may cover a wider set of characteristics and therefore help to achieve saturation.

For correct interpretation of the results, the process of sampling and recruiting should be well documented. Designed and achieved samples should be compared to determine whether there was any selectivity in the likelihood that businesses were willing to participate in the test and how this may have affected the outcome. Moreover, it is important to check if the hypothesised influence actually occurred and if any other unforeseen characteristic has influenced the results. For example, if some small businesses only keep mandatory records for tax purposes and others have excellent information systems, then size obviously does not discriminate and it is necessary to examine other characteristics to assess whether they contribute to the difference in data availability (e.g. regulated vs. non-regulated economic activity; a lot of internal reporting because of a distant owner, evidence-based management, etc.). It may be possible to go further than checking for the effects of recruitment on the cognitive testing outcomes – considering the types of businesses with failed recruitment may also give insights into the response processes and likely issues.

With respect to sampling and recruitment for qualitative testing in establishment surveys, we thus propose:

Recommendation 4: Sample for the greatest variety in as many relevant characteristics as possible, starting with business size, economic activity and organisational complexity

Recommendation 5: Use sample sizes that allow as complete an evaluation as possible

Recommendation 6: Document sampling decisions, recruitment processes and outcomes, especially with regard to relevant business characteristics

Recommendation 7: Assess the representativity of the achieved sample relative to the research goals and consider what this may mean for the interpretation of the findings.

3.3 Data Collection

An important guideline for testing in general is to come as close to field conditions as possible (Willis, 2005). As the response process in establishment surveys involves the use of organisational infrastructure, it seems indispensable to focus on qualitative research methods that in some way relate to the response process in the organisational setting (e.g. by interacting with respondents or by studying their records), to conduct the study onsite and to request actual filling in of the questions. Unlike household surveys, the establishment surveys environment is not accurately or easily replicated for testing. To fully understand respondents' cognitive questionnaire answering process within complex reporting structures requires testing with real establishment respondents who have the profound detailed technical knowledge needed for response (McCarthy, *et al.*, 2018).

Expert review or appraisal of a questionnaire is an important evaluation method that effectively builds on previous testing findings, content-matter knowledge and experience, given that many behaviours in the business survey response process have been well studied and still seem pervasive. However, expert review lacks a direct connection with observed units. Evaluating the actual response process may be difficult or sometimes impossible, but it does provide crucial extra information above any hypothetical discussion, as respondents often cannot foresee all the problems they might have with a question before they actually start answering it.

Input from expert reviews should be obtained early in the questionnaire evaluation process so appropriate time can be devoted to vetting and testing recommendations (McCarthy, *et al.*, 2018). Another aspect of field conditions is also the survey mode. The literature lists as many as 27 survey modes (Mohorko & Hlebec, 2016) but most are not typical of establishment surveys that are predominantly based on self-administered paper or web questionnaires. As these questionnaires rely completely on visual stimuli, cognitive interviewing and usability testing should preferably present the survey questions in the same way as they will appear in the real survey (Gray, 2015). When surveys are conducted using multiple modes of data collection, the questionnaire should ideally be tested in all modes to ensure equivalence of measurement (Brancato, *et al.*, 2005).

When conducting cognitive interviews, competent staff are of paramount importance. The interviewers are responsible for ensuring that the collected qualitative data are of high quality (e.g. by correctly applying the method, making respondents feel comfortable, stimulating verbalisation) (Mohorko & Hlebec, 2015). Although some authors call for more standardisation in conducting cognitive interviews, flexibility in following up potential unanticipated problems is preferred by others but also requires more skilled staff (Beatty & Willis, 2007). In this more demanding role, the interviewer should be able to assess the collected information, identify any gaps and contradictions and follow them up to arrive at a full understanding of respondents' experiences (Willson & Miller, 2014). In establishment surveys, interviewers also have to be thoroughly familiar with the relevant business concepts and terminology, which are often unique to a particular survey (Gower & Nargundkar, 1991). Ideally the staff will have both cognitive interviewing skills and subject matter expertise (Nichols & Childs, 2009). In some cases it may be easier to train topic experts in cognitive interviewing than to train interviewers in the details of complex subject matter topics (Nichols & Childs, 2009), although equally the fact that cognitive interviewers are less knowledgeable about the specific subject matter can help uncover issues as they probe deeper, and this also means they are not tempted to help or coach respondents during cognitive

pretesting. In practice interview teams may be used with, for example, both a content matter specialist and a survey methodologist (Giesen, 2007). Such teams have to be briefed on the interviewing protocol to avoid introducing bias.

Cognitive interviewing can be implemented as think-aloud or verbal probing. There is almost no research addressing the comparative advantages of think-aloud and concurrent probing in self-administered business survey questionnaires. Self-administration and extensive retrieval of data from business information systems suggest that concurrent verbal probing might interfere with the visual content (Redline, Smiley, Lee, & DeMaio, 1998) while think-aloud might require an unacceptably long time. Giesen (2007) recommends using both visits where respondents are observed while filling out the questionnaire and visits in which the response process is reconstructed retrospectively, in order to combine the benefits of both methods. Retrospective probing might be reproached for missing the information on cognitive processes from the respondent's short-term memory. On the other hand, some authors argue that this should not be the purpose at all as respondents are generally neither good reporters nor evaluators of their cognitive processes (Miller, *et al.*, 2014). This concern is somewhat reduced in establishment surveys because respondents tend to keep notes of their response procedures – for example the location of data, and calculations or adjustments needed to make data from their business records meet the requirements of the question. Since many establishment surveys are repeated at regular intervals, business respondents retain these notes in order to replicate their responses for the next iteration of the survey. Respondents should be best at reporting their personal and organisational experience, which is in line with the ethnographic approach to cognitive interviewing (Gerber, 1999). Qualitative research interviews seem to be a promising method for reconstructing the meaning respondents attach to survey questions based on their business context. This method typically uses unstandardised and open data collection methods but relies heavily on the skills and expertise of qualitative interviewers. With an unstandardized approach the same probing questions are not asked in all interviews, which complicates the data analysis. This could be mitigated using some scripted probing questions, and by encouraging interviewers to continue probing until they have obtained all the necessary information.

With respect to data collection from cognitive interviewing in establishment surveys, we thus propose:

Recommendation 8: Test as realistically as possible:

- **Implement tests with real establishment respondents *in situ***
- **Use observation and reconstruction of the response process**
- **Test mixed mode questionnaires in each mode**

Recommendation 9: Secure interviewer competences

- **Qualitative interviewing skills**
- **Content matter knowledge about relevant business concepts and terminology**
- **Knowledge about the goal of the testing**

3.4 Data capture

Cognitive interviewing typically creates rich data with a narrative structure reflecting the semi-structured format of the interviewing procedure. Data collected as part of recruiting and/or cognitive interviewing may encompass a range of themes:

- information on the business, e.g. information on recruitment into the study, size, type of industry, location(s), organisational structure, information known about previous responses or complaints;
- information on the respondent(s), e.g. job title, educational background, years and type of working experience in this business and with reporting obligations;
- information about the conduct of the interview, e.g. the place where the interview took place, who was present or in hearing distance, relevant interactions with co-workers and management during

the interview, the general atmosphere of the interview and any changes in spirit noted during the interview;
- information on available business data;
- information about the response process, e.g. what the respondent said, how it was said, documentation and other resources that the respondent used for answering survey questions, observation of calculation and estimation methods used, the answers provided to the tested questions and an assessment of the quality of that answer.

The extent to which this information is captured affects both the information that is available for further analysis and the degree to which others can reconstruct the research process.

Typically, in qualitative testing of questionnaires we see three main approaches or combinations of these to capture the data: 1) note taking by the interviewer; 2) note taking by an observer; or 3) capturing information in as unfiltered a way as possible with e.g. audio recording, video recording, on-site transcription, eye-tracking, or screencapture.

The literature on cognitive testing implicitly (e.g. Willis, 2015) or explicitly (Gray, 2015; Miller*, et al.*, 2014) recommends making audio recordings of interviews (of course, with the respondents' consent). There are various views on how recordings of interviews should be used. Willis (2015) sees value in transcribing interviews or at least listening to recordings of the interviews again, in that this makes sure that analysis is based on the respondents' real words. However, this is much more time-consuming and expensive than just using notes and may often not be feasible. Willis suggests as a compromise reviewing only segments of the taped interview for which original notes are unclear or where the interaction between the respondent and interviewer was complex. Gray (2015) recommends explicitly audio recording every interview to allow the interviewer to focus on the interview (and not on note taking) and to provide a full record of everything that interviewers and respondents say that can be used for a written summary of the interview (which may be reviewed at a later point in time if for example the summaries prove inadequate). D'Ardenne & Collins (2015) mention as an additional benefit that listening to recordings (especially as soon as possible after the interview was conducted) helps to review how well the interview went and how techniques may be improved for the next interview. DeMaio & Landreth (2004) compared methods and results for three different teams in cognitive testing of a household survey. They conclude from their findings (p. 107) "The results suggest that the extra time and effort associated with listening to tapes of cognitive interviews have a big payoff in identifying respondent problems…., the added exposure to the thoughts and comments of respondents can supply further insight into or clarification of the response process. This provides some evidence that a more rigorous review of the data may result in a greater understanding of questionnaire problems".

If recordings are not an option, then having more than one interviewer is good practice, as this is an extra pair of eyes and ears and maybe also an additional type of expertise available for the interview.

With respect to data capture from cognitive interviewing in establishment surveys, we thus propose:

Recommendation 10: Capture data as naturally as possible, preferably by recordings or transcriptions, and paying attention to non-verbal actions (e.g. access to documentation and other resources)

3.5 Data analysis

Given the flexible and open nature of data collection in qualitative research, analysis actually already starts during the data collection. Also, during the capture of the data, decisions are made about what to record and/or transcribe (Davidsson, 2002). This section focuses on Willis's (2015, p. 56) definition of analyses, namely "the series of steps that occur between data collection and the communication of what we have found". He describes two contrasting objectives of cognitive interviewing that guide the focus of analyses: the reparative approach and the descriptive approach. The reparative approach focuses on detecting problems in measurement instruments and finding ways to fix them. The main goal of the descriptive approach is to get a broad understanding of how the measurement instrument works, including aspects that work well. Willis notes that in practice many studies contain elements of both approaches.

In the past, literature on cognitive interviewing did not provide much insight into how to analyse cognitive interviewing data (e.g. Boeije & Willis, 2013). Recent volumes on cognitive testing have addressed this topic extensively. We first examine these approaches, which are framed generally, and then consider how they may apply in the specific situation of establishment surveys.

Willis (2015) distinguishes five models, based on whether or not data are coded (text-summarisation versus coding), and the approach when they are coded (top-down – cognitive coding; top-down – question feature coding; bottom-up – theme coding and bottom-up – pattern coding). Each model has its strengths and limitations and, again, different ways of analysis can be combined. Miller *et al*. (2014) recommend using five incremental steps for the analysis: 1) conducting interviews, 2) producing summaries, 3) comparing across (all) respondents, 4) comparing across subgroups of respondents, and 5) reaching conclusions. D'Ardenne and Collins (2015) recommend a similar approach in four steps: 1) data collection, 2) data management to organise the data, to make navigation easier, 3) descriptive analysis to develop understanding of how questions were interpreted and answers formulated and, 4) explanatory analysis, to identify whether questions can be repaired and if so how.

All three recent cognitive interviewing handbooks (Collins, 2015b; Miller*, et al.*, 2014; Willis, 2015), and guidelines (Economic and Social Commission for Asia Pacific Region, 2010; Office of Management and Budget, 2016) recommend that data from each interview is summarised and structured to facilitate comparison across interviews. Willis (2015) stresses that analysis should also focus on what happens within an interview, across items, to detect context effects, for example. Bavdaž (2009) and Collins (2015a) make an additional point not to focus only on analysing data by survey question, arguing that this may prevent absorption of the big picture and detection of general patterns. They therefore recommend also analysing the data by sources of measurement errors (Bavdaž, 2010b) and by parts of the response process (e.g. comprehension issues, problems with retrieval of relevant data).

The specific context of business surveys suggests some additional analytical attention. Bavdaž (2010a) highlights that there are different processes happening at individual and organisational levels, operating together to form a response to a question, and these levels must be considered in analysing the data from cognitive interviews. For example, the institutional environment can have a large effect on the way that an individual approaches completing a questionnaire. Also, responses to business surveys typically rely very heavily on records, so the ease with which the respondent can interact with these is important (and in some cases multiple record systems may need to be accessed). *In extremis* the required data may not be available at all, or may be available only by calculation or approximation, and the impact of this also needs to be assessed.

In analysing these it is important to keep the skewed nature of the business population in mind – an issue for small businesses may affect many respondents but have a smaller impact than a competing problem for large businesses. Perhaps different approaches for different business sizes will be appropriate, although this increases the complexity.

Regardless of the method of analysis, interpreting the data will retain an element of subjectivity, with experts taking account of the strengths and limitations of different approaches. It is therefore recommended to build opportunities in the analytical process to discuss interpretations with others to reflect on alternatives and any biases that may occur (e.g. Shenton, 2004). Willis (2015) recommends an "ongoing, intensive communication and collaboration throughout the analysis and interpretation process" and good documentation; conclusions and recommendations should at least indicate their basis, including statements that "no problems were found".

With respect to analysis of data from cognitive testing of business survey questionnaires, we thus propose:

Recommendation 11: Summarise raw data in a structured and systematic way.

Recommendation 12: Analyse data in depth, preferably by immersion in raw interview data, coding all data and comparing data:

- **within each interview,**
- **across interviews about the same business (when more than one person is involved in the response process in connection with the same business),**

- **and across all interviews (taking into account potential impact of skewed population distributions).**

Recommendation 13: Analyse data not only by question but also by characteristics used in sampling, sources of measurement errors and parts of the response process (distinguishing individual and organisational levels).

Recommendation 14: Involve more than one researcher in the analysis process and recommendation generation (the minimum being one researcher with content matter knowledge of business concepts and terminology).

3.6 Data reporting and beyond

Reporting is a useful step in any research activity as it pushes all stakeholders to reconsider (once more) the whole research activity, at this point with all information about the implementation and newly collected data. Comprehensive reporting might be especially important in qualitative research studies because assessments of their objectivity and integrity rely on transparency. Reporting is usually tailored to the audience and depends on the purpose. Willis & Boeije (2013a, 2013b) call for reporting of cognitive interviewing (and other testing approaches) as such interviewing seems to be neither consistently implemented nor widely evaluated. They introduce a systematic, complete, and harmonised system of reporting, the Cognitive Interviewing Reporting Framework (CIRF), to start creating the evidence necessary for process evaluation and the comparison of the effectiveness of varied approaches (Boeije & Willis, 2013). The CIRF proposes a ten-category checklist, thus suggesting the minimum level of required information and an easier search for specific information.

Reporting in the case of establishment surveys is essentially an application of these procedures, taking account of the specifics of establishment surveys mentioned in preceding sections. There are some additional details to consider. In particular, disclosure control is usually more challenging for establishment surveys, so extra precautions (compared with social surveys) are needed to keep respondents' identities confidential (or to gain permission to relax this condition, where that is legally permissible). Documenting data availability, complexity of the response process, and the way in which reported data are eventually produced by respondents is also an important element.

With respect to data reporting and post-testing steps, we thus propose:

Recommendation 15: Document the study design and its strengths and limitations, and the results by all analysed aspects. Also document what was intentionally *not* examined

Recommendation 16: Disclose all methodological details that make the research process and outcome traceable and understandable, having regard for pledges made to study participants

Recommendation 17: Provide access to documentation, having regard for any limitations of disclosure control

Recommendation 18: Follow up and document whether and how recommendations were implemented

Recommendation 19: Evaluate with field-work data how well the questionnaire worked in practice (especially if performance of any part of the response process was systematically challenging, if any subgroup of businesses experienced problems etc.), and use this information to reflect on design and results of the pre-test

4. DATA

To understand whether NSIs use qualitative testing methods for question and questionnaire evaluation in business and establishment surveys and how much they follow the recommendations presented in section 3, we conducted the International Survey of Qualitative Testing Practice for Business and Establishment Surveys described below.

4.1 Survey and Questionnaire Design

The International Survey of Qualitative Testing Practice for Business and Establishment Surveys was a web survey of NSIs. Participation was invited by email. The list of NSIs (country, NSI name, the

director's name and email) was retrieved from the website of the International Statistical Institute (2016). The list excluded institutions that were not NSIs (e.g. societies and research centres) and those operating at a lower hierarchical level in the national statistical system so that only the main institution responsible for official statistics was kept per country. The only exception was the U.S.A. where 18 federal agencies were taken into account (because of the decentralised system for establishment surveys). The NSIs were then assigned to six geographical regions as defined by the United Nations (2016). The population initially consisted of 232 NSIs from 215 countries and they were all invited to the survey (see Table 1).

The email invitation was sent either to the director (general) of the NSI or to the person identified as knowledgeable about questionnaire testing. Specific people were mainly identified in European and North American NSIs through the authors' personal links and/or pre-contacts by email to NSIs; this is likely to have shortened the communication path, but unlikely to have had a significant impact on response. When no email was available or only a general one, an attempt was made to find the director's email. This exercise revealed some outdated information (e.g. changes of director) and use of private emails for work purposes in some regions.

The email invitation explained the purpose of the study and asked for help identifying the best respondent if the recipient did not have enough knowledge of questionnaire testing. The text included the web link to the web survey and an individualised access code. Respondents could also register for participation in the survey. Email and telephone contacts were provided in case of questions. The invitation was signed by the international team of five researchers (the authors of this paper).

The questionnaire was drafted, discussed and revised in several iterative steps to reach a consensus within the international team of five researchers. One of the authors completed the questionnaire on behalf of her NSI before the survey went into the field.

The questionnaire addressed five themes: (i) data to determine the eligibility of the NSI and the appropriateness of respondent selection; (ii) sampling and recruiting for qualitative testing; (iii) design and collection of qualitative interview data; (iv) analysis and reporting of qualitative testing; and (v) an important recent qualitative testing project. Most questions referred to the last five years. A copy of the questionnaire (as screenshots) is provided in the supplementary material.

4.2 Implementation and Response

The Social and Economic Sciences Research Center, an academic survey centre at Washington State University, hosted the survey. The survey was in the field between the end of April and the end of August 2016 though most responses came in by the end of June after three email reminders over six weeks. 49 emails had to be resent to new addresses, and even these emails did not reach 14 NSIs, so alternative email addresses were sought. Additional efforts were made as part of the non-response follow up: sending a personalised email request through a connection if available and asking just three questions in the email to better understand the situation in nonresponding NSIs. These three questions asked whether or not the NSI conducted establishment surveys, about how many establishment surveys they conducted annually and whether or not they interviewed or otherwise contacted people from businesses when preparing new or changing old survey questions and questionnaires.

Table 1: Response and Main Characteristics by Geographical Region

Geographical Region	*Invited NSIs*	*Contact Established*	*Establishment Surveys Conducted*	*Qualitative Testing Conducted*	*Response*
Africa	53	15	15	8	3
Asia	51	17	17	9	1
Europe	45	33	33	19	14

Latin America and the Caribbean	40	9	9	3	3
Northern America	20	15	12	11	9
Oceania	23	6	4	3	2
Total	*232*	*95*	*90*	*53*	*32*

Table 1 provides an overview of response by six geographical regions. We managed to establish a contact and get at least some data from 95 or 41% of 232 invited NSIs. From these contacts we learned that 5 NSIs did not conduct establishment surveys (e.g. they rely exclusively on administrative data) so they should not have been included in our population of NSIs. Out of the remaining 90 NSIs, 53 or 59% conducted (at least some sort of) qualitative testing of establishment surveys.

After careful examination of individual answers, some of the 53 responding NSIs were excluded from analyses because their answers had too many missing values or several answers suggesting miscomprehension of qualitative research vocabulary, thus questioning the conduct or even presence of qualitative testing (e.g. after three questions on coding of data from qualitative interviews, an open response referred to standard economic classifications such as ISIC that also contain codes; Stata listed as a way of documenting qualitative interviews; a sample of several thousand units used in qualitative testing). Some NSIs also started completing the questionnaire only to realise that they did not conduct this kind of testing. The final analysis data included 32 NSIs that conducted establishment survey qualitative testing, of which five were treated as partial respondents.

Respondents were assured that their data would not be used in a way that identified them, so the detailed responses from the survey are not available. For more information please contact the authors.

4.3 Respondents

Most respondents to the survey described themselves as knowledgeable of qualitative testing of establishment surveys across their organisation. 19 or 59% said they knew about qualitative testing for most or all of the establishment surveys conducted by the organisation, 11 or 34% of some establishment survey testing, and only two reported knowledge of testing in only one establishment survey. More than half, 19 or 59% of responding NSIs had a central team or unit responsible for carrying out qualitative research or testing of establishment questionnaires. The majority, 26 or 81%, had conducted this type of testing in 2015 and 2016.

5. RESULTS

In this section we present the practice/reality of conducting qualitative studies in the analysed NSIs and compare it with our recommendations. The presentations follow the same research stages as Section 3. All figure captions include the question number from the survey.

5.1 Overall design

Recommendation 1 suggests pre-testing new survey components. Our survey asked about the reasons why qualitative research methods were used, and gave a series of options, with respondents scoring each option. The type of qualitative research was not specified, as we wanted to capture information on any activity in this area. The response options covered different reasons why question testing might be required, including for new survey questions, as a result of issues identified through respondents, through measurement (for example in editing) and through item nonresponse rates. There was some redundancy among the categories, which should be borne in mind, but it was felt preferable to cover all the possibilities rather than have a complex coding scheme.

Developing new survey questions was the most common reason for conducting qualitative research but it was not used as a standard among analysed NSIs; only 22 out of 32 NSIs “Often/Always” used

qualitative methods when developing new questions. As we can see in Figure 1 qualitative research was also used with varying frequencies for a range of other purposes.

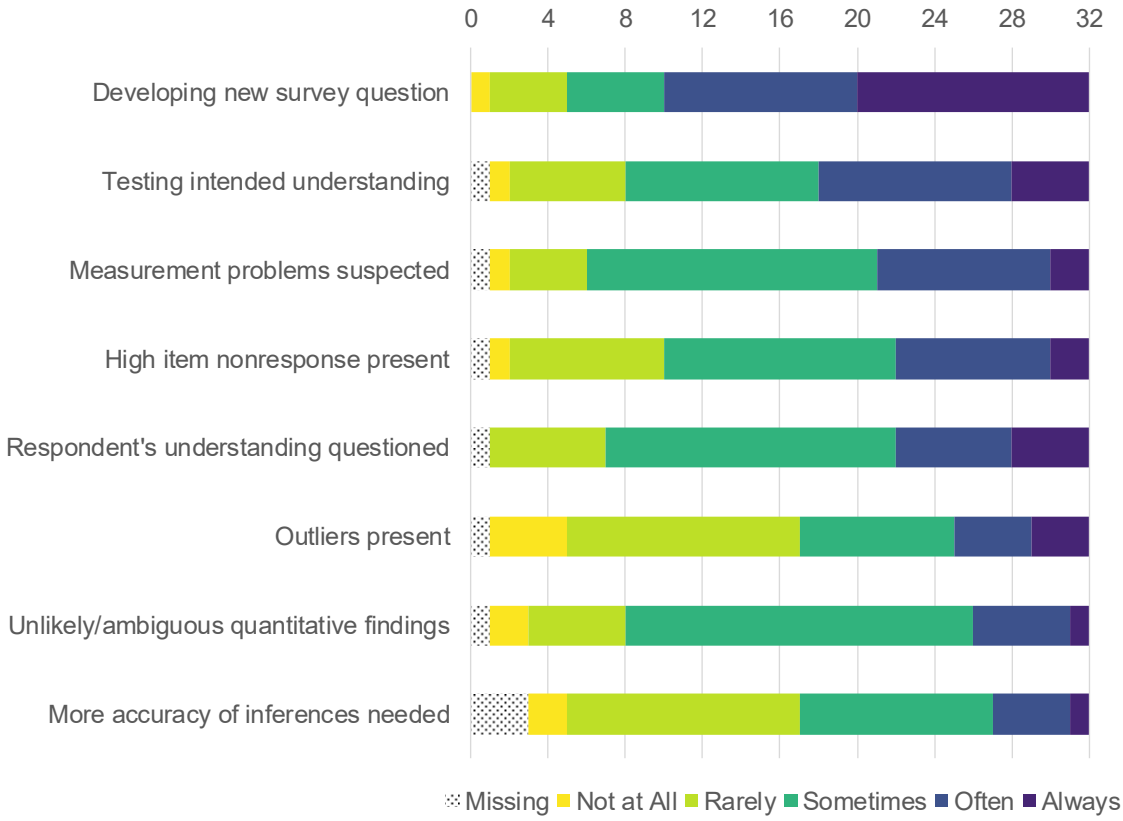

Figure 1: Reasons for Using Qualitative Research in Surveyed NSIs (q9)

Recommendation 2 suggests using more than one testing method. Our survey asked about focus groups, usability tests, observations, record checks and three types of interviews (in-depth, cognitive and pretest). We used these three types of interviews to be sure to accommodate different expressions used in NSIs around the world.

Pretest and cognitive interviews were most often part of the classic set of methods always used for testing (in eight and six NSIs respectively). Focus groups and in-depth interviews were most often completely absent (in eight and seven NSIs respectively). The method that was also most often absent was record checking (19 NSIs).

To do more justice to what happens in the field, we have to acknowledge that the three types of interviews we asked about might not be clearly differentiated in practice. The complexity of the response process often pushes cognitive interviewing to become a cognitive hybrid, exploring data availability and respondent roles along with cognitive response processes (Willimack, 2013). We therefore collapsed the three types of interviews into a single category to end up with five more distinct qualitative research methods (although some overlap is also possible here): focus groups, qualitative interviewing, usability tests, observations, and record checks. Figure 2 presents the answers from all surveyed NSIs about the use of qualitative research methods. Each column represents the responses of one of the 32 NSIs, and the columns are sorted (from right to left) by the number of methods used: Always, then Often, then Sometimes etc. Nearly a third of NSIs (around 10) intensively used a wide range of methods (indicated as columns of darker colours with no or few bright colours). No NSI indicated reliance on a single testing method; particularly worrisome is the low number of methods used and their rare application in a few NSIs.

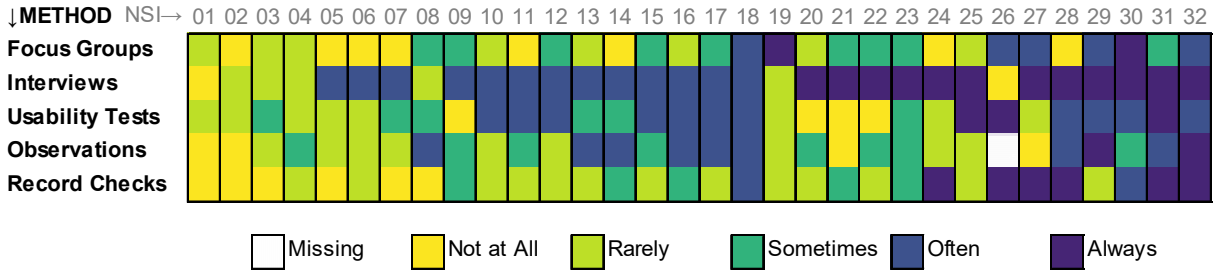

Figure 2: Use of selected qualitative testing methods across surveyed NSIs (n=32; a column = an NSI; columns are sorted (from right to left) by the number of methods used: Always, then Often, then Sometimes etc.; Interviews encompass in-depth, cognitive and pretest interviews; q10)

About half the surveyed NSIs had at least tried all five of these methods and a large majority had tried four out of five. At least two methods were used sometimes or more frequently by more than half of the NSIs. Although this was not direct proof that a combination of methods was used for every testing, it is an indication that this is possible.

We did not ask about iteration of testing as suggested in Recommendation 3, but some descriptions of sample selection made it clear that iterative testing was being used.

5.2 Sampling and Recruitment

Recommendation 4 suggests sampling for the greatest variety in as many relevant characteristics as possible, starting with business size, economic activity and organisational complexity. Our survey explicitly asked about eight characteristics that relate to either organisational characteristics (industry, size, geographical location, single vs. multiple locations) or survey behaviour (problematic, unproblematic, new (to the survey), previously surveyed). Figure 3 shows that covering a range of sizes was the most frequently used criterion among surveyed NSIs with 23 surveyed NSIs "Often/Always" using it, followed by covering a range of industries with 19 NSIs "Often/Always" using that. Previously surveyed establishments were also "Often/Always" considered in 13 NSIs. Only two NSIs "Often/Always" simultaneously targeted previously surveyed and new (to the survey) establishments and no NSI would always try to cover new establishments. When testing concerned new questions, differentiating between "old" and "new" establishments appeared less relevant, but when revising existing questions, "old" establishments may come with a baggage of experience which could be beneficial or not according to the changes proposed. Paying attention to establishments with multiple sites likely reflects their additional complexity and the need to ensure that questionnaires work for these businesses, which are often the most important for published estimates. The majority of surveyed NSIs (26 NSIs) "Often/Always" sought coverage of at least two of the listed characteristics.

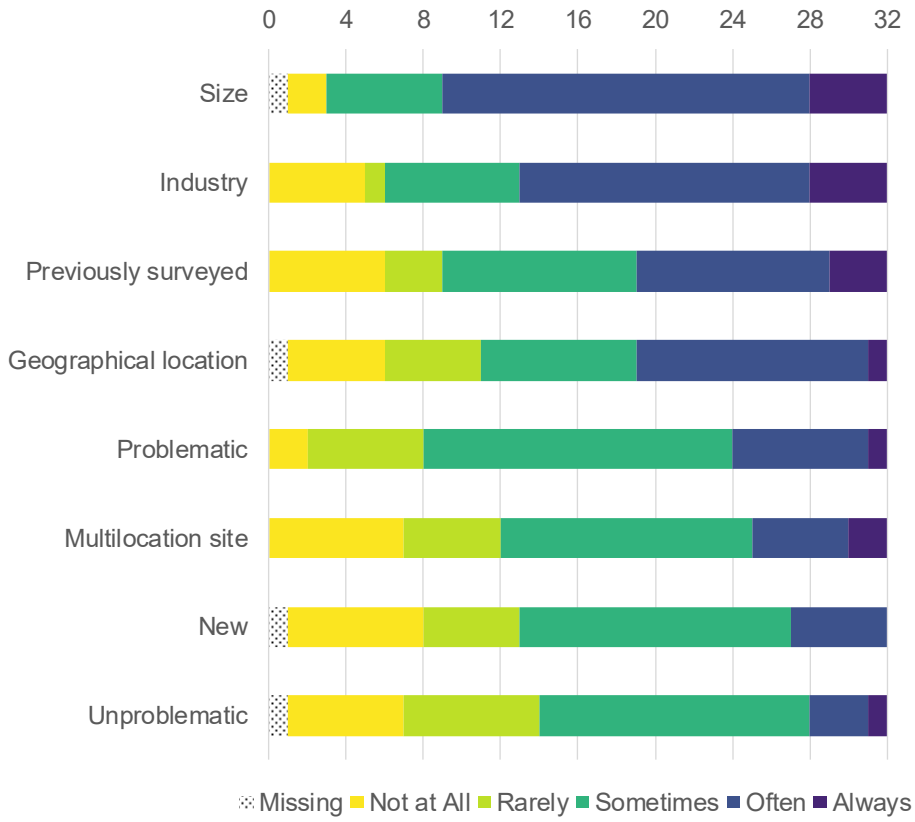

Figure 3: Range of Business Characteristics included within Qualitative Testing Samples (q22)

Participants were most frequently chosen for qualitative studies because they were already participating in surveys, and almost as frequently because they had previously participated in the survey (Figure 4). Fewer NSIs "Often/Always" used recruits drawn from businesses which had agreed to be recontacted. Contacts with businesses were most often with a named survey contact, and there was a very clear hierarchy of contact modes, with 21 NSIs "Often/Always" using telephone, 14 using email and only 4 "Often/Always" using post.

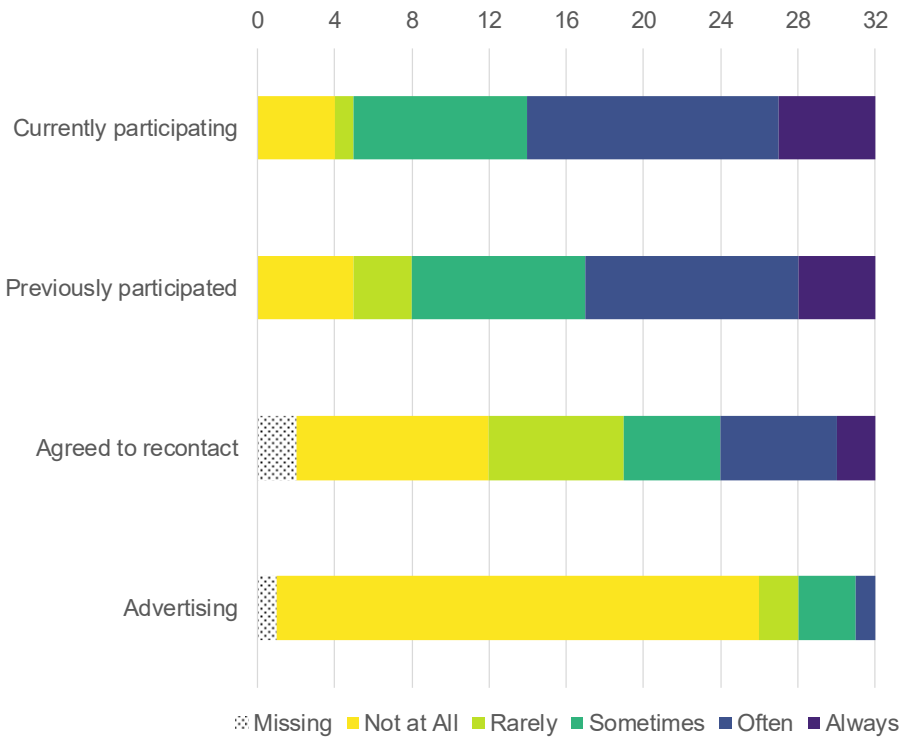

Figure 4: Criteria for Choosing Study Participants (q18).

Recommendation 5 suggests using sample sizes that allow as complete an evaluation as possible. Figure 5 shows boxplots of the distribution of qualitative testing sizes, specified by surveyed NSIs as the minimum, typical and maximum sizes. Most NSIs typically worked with tiny samples: 13 NSIs (out of 26 responding) typically used less than 11 units; 23 NSIs (out of 27 responding) had less than 11 units in the worst-case scenario; and in 13 NSIs even the largest samples were (only) up to 30 units. The sheer number of included units cannot tell us much about the completeness of the evaluation, except that the likelihood of attaining saturation does not seem to be very high for most NSIs because the range of sample sizes and reasons for sample size choices suggest they would like to do more (though Guest, *et al.*, 2006 document some situations where good results are obtained with small samples). Saturation is indicative that the range of variation has been covered (see section 3.2).

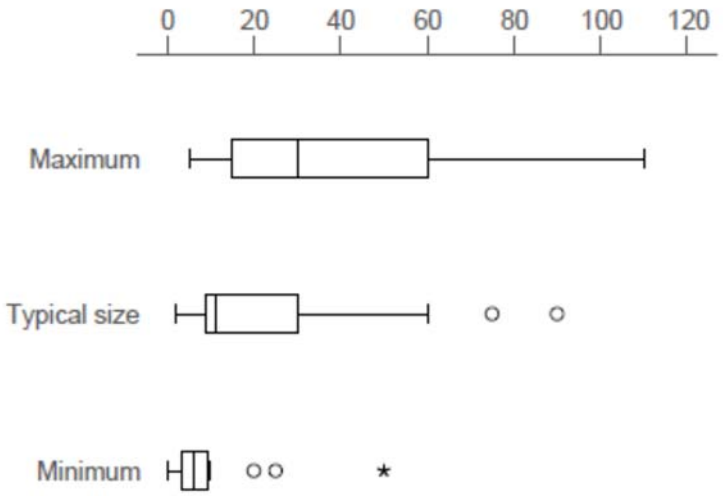

Figure 5: Boxplots of the Distributions of Minimum, Typical and Maximum Sample Sizes among Respondents (n=25-27; q15). (The circles and stars represent outliers and extreme outliers respectively.)

An insight into the factors influencing the sample size might help us understand the degree to which sample sizes are selected based on research goals and needs, rather than based primarily on constraints. Figure 6 shows that resource constraints (in terms of specialist staff time and budget) were the most common "Often/Always" important factors (in line with other qualitative research, Fusch, 2015; Tourangeau, 2004). The goal of qualitative research was the next determinant of size. An agency rule specified a size in some cases. Difficulty in obtaining recruits was "Often/Always" a factor in more NSIs than getting access to businesses and actual respondents, but the accuracy of the frame was mostly not an important constraint. The cost of cash incentives was likewise not a common constraint, and indeed 29 of the 32 surveyed NSIs "Rarely/Never" used incentives.

The resource constraints were also reflected in the sampling approach, where the procedures which were "Often/Always" used in most respondent NSIs were quota sampling (18 out of 30 NSIs), and sampling based on practical considerations (for example, a convenience sample choosing businesses near to the location of the office to reduce travel costs for specialist staff) (22 out of 32 NSIs).

On the positive side, the goal of the qualitative research was "Often/Always" guiding the choice of the sample size in 19 NSIs, which suggests that the extent of the evaluation might be sufficient to reach the research goals.

We did not ask directly about representativity, since it is difficult to define precisely, though "cover subgroups of businesses" relates to it. There is further discussion of representativity and recommendation 7 in section 5.5 below.

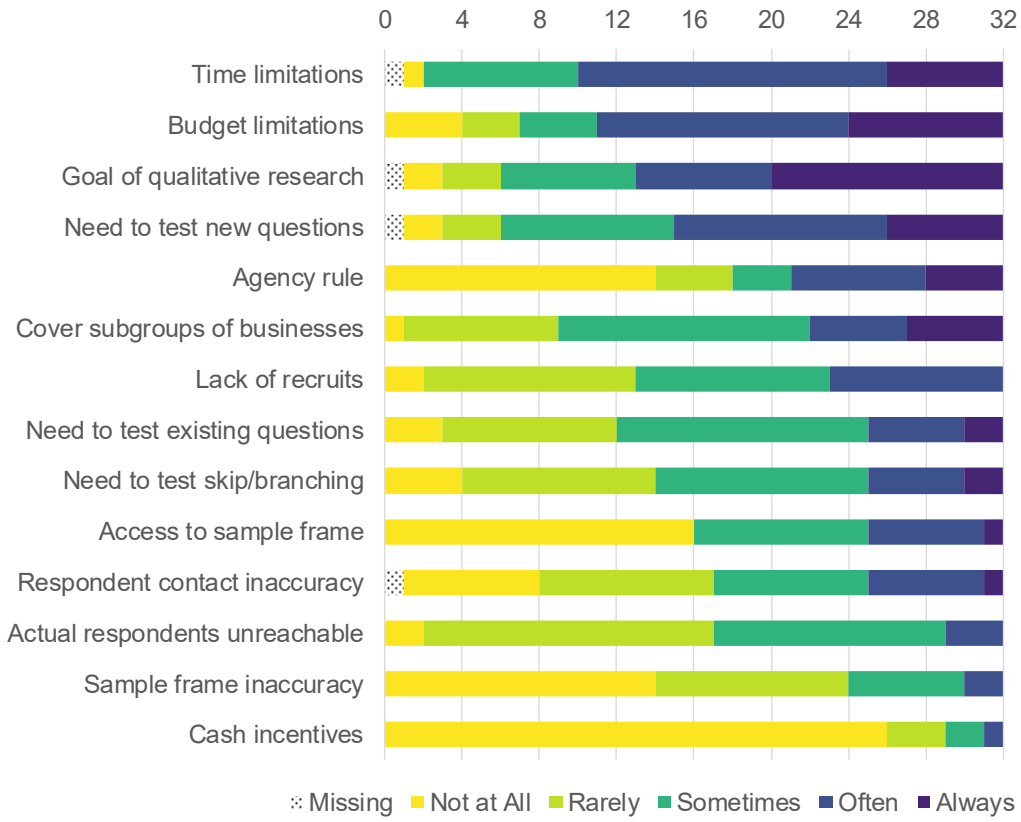

Figure 6: Reasons for Sample Size Choices in Qualitative Testing (q17).

Figure 6 shows that decision-making on sample size choice is multi-faceted. The open answers in our survey on how the number of business units was selected for qualitative testing also reflect how this decision was made based on different considerations. One of the NSIs explained

The major determin[ant] is the diversity of the survey sample. If the same questionnaire is to go to many different types of businesses (usually many different industries / sub industries and to many different sizes of businesses - micro through to very large) then the form will need to be tested with defined subgroups. […] A complex survey like this might include a sample of 45-60 units. We usually do at least two (iterative) rounds of observational testing as well as initial informational testing and have a policy of never returning to the same unit, so for a complex survey, we can expect that over one hundred units were in the combined test sample".

Another NSI wrote that sample size was determined by *"size of survey target populations; number and diversity of industries covered by the survey; the type of problem being investigated; potential impact/risk of error or how high profile the survey and data are; the location of the testing (e.g., local, distance); method being used for testing (e.g., cognitive testing vs usability testing vs exploratory "scoping" vs post-collection debriefings); the mode being used for testing (in-person interviews vs phone; amount of (sponsor) time and money available; staff availability and workload".*

Two NSIs commented explicitly on reaching saturation. One NSI stated *"We include a maximum of 12 business units, as we found out a point of saturation, where problems repeat themselves".* Another NSI that typically used a sample of five commented *"Beyond this point, the same things keep on coming up."*

Recommendation 6 suggests documenting sampling decisions, recruitment processes and outcomes. Figure 7 shows that many NSIs did not seem to systematically record their recruitment attempts and results, as for example only 10 out of 29 "Often/Always" recorded reasons for refusal.

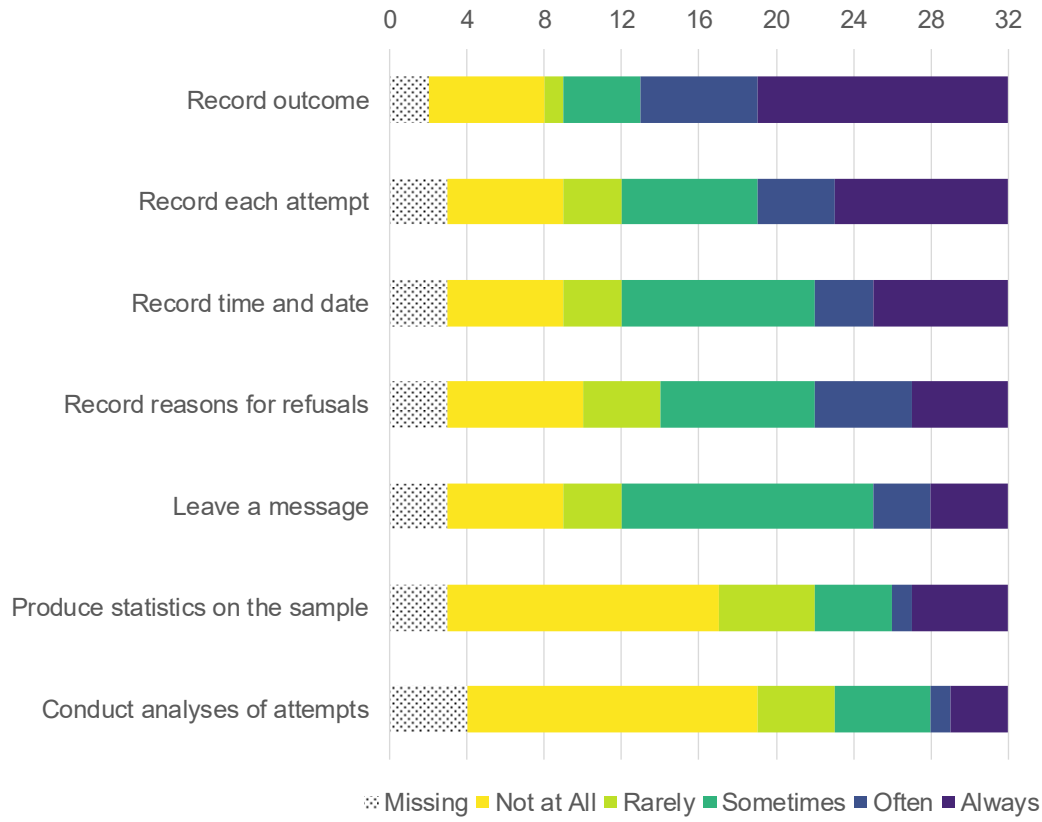

Figure 7: Practices and analyses for the outcomes of sampling and contact attempts (q27).

5.3 Data Collection

Recommendation 8 suggests testing in as realistic a setting as possible. Figure 8 shows that many responding NSIs indeed "Often/Always" used data collection methods that entail a direct contact with businesses: 22 NSIs conducted interviews face-to-face and 10 NSIs undertook observations; other methods (telephone, mail and web) were used less. 17 "Often/Always" conducted these interviews at the respondent's work place or desk at the business and 18 "Never/Rarely" conducted them at the NSI. An encouraging finding is that more than half of NSIs (17 out of 30) "Often/Always" combined qualitative interviewing with observations of the respondents when completing all or parts of the questionnaire using the actual data collection instrument, thus constructing a field setting that comes very close to the actual one. Somewhat fewer NSIs, 11 out of 25 that conducted establishment surveys in multiple modes, tested and evaluated all modes. How much impact this has depends on how many and which businesses use the different modes and to what extent mode effects can be expected. Practical constraints and the availability of IT tools inspire innovative ways to test surveys. One NSI conducted testing remotely over Skype with shared screens.

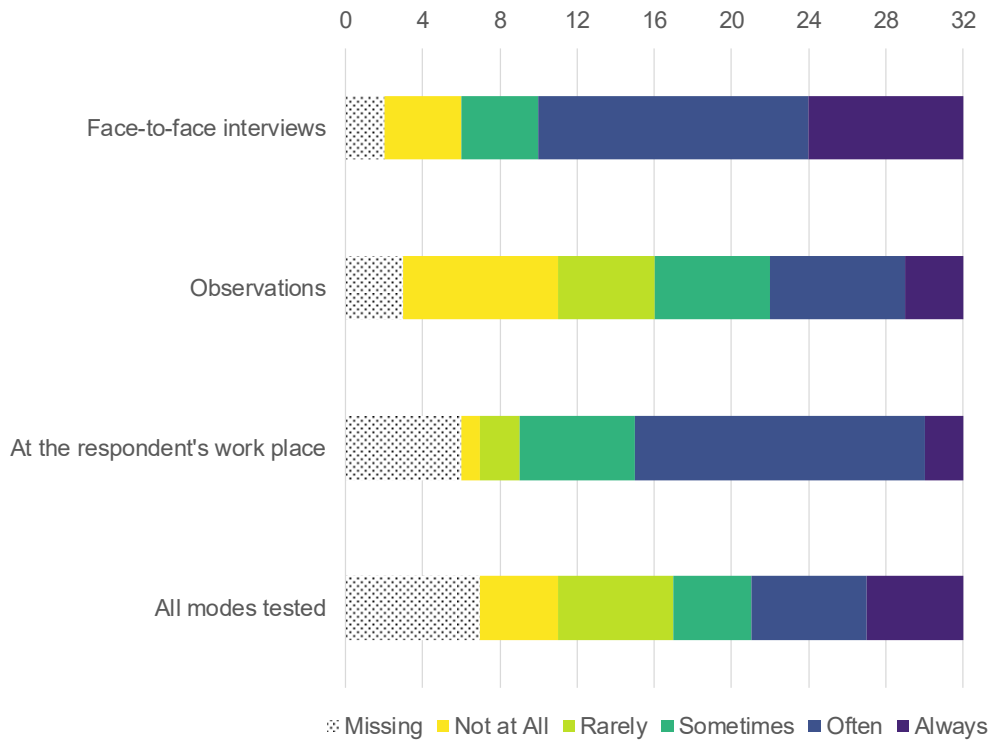

Figure 8: Data collection practices (q34, q35, q36).

Recommendation 9 suggests securing interviewer competences. As Figure 9 shows 23 NSIs "Often/Always" used staff trained or experienced in qualitative research, and 13 NSIs used staff with content matter knowledge. Involving interviewers or field staff (from the main survey collection) was less common. In 19 NSIs the same staff who drafted or developed questions "Often/Always" also tested them.

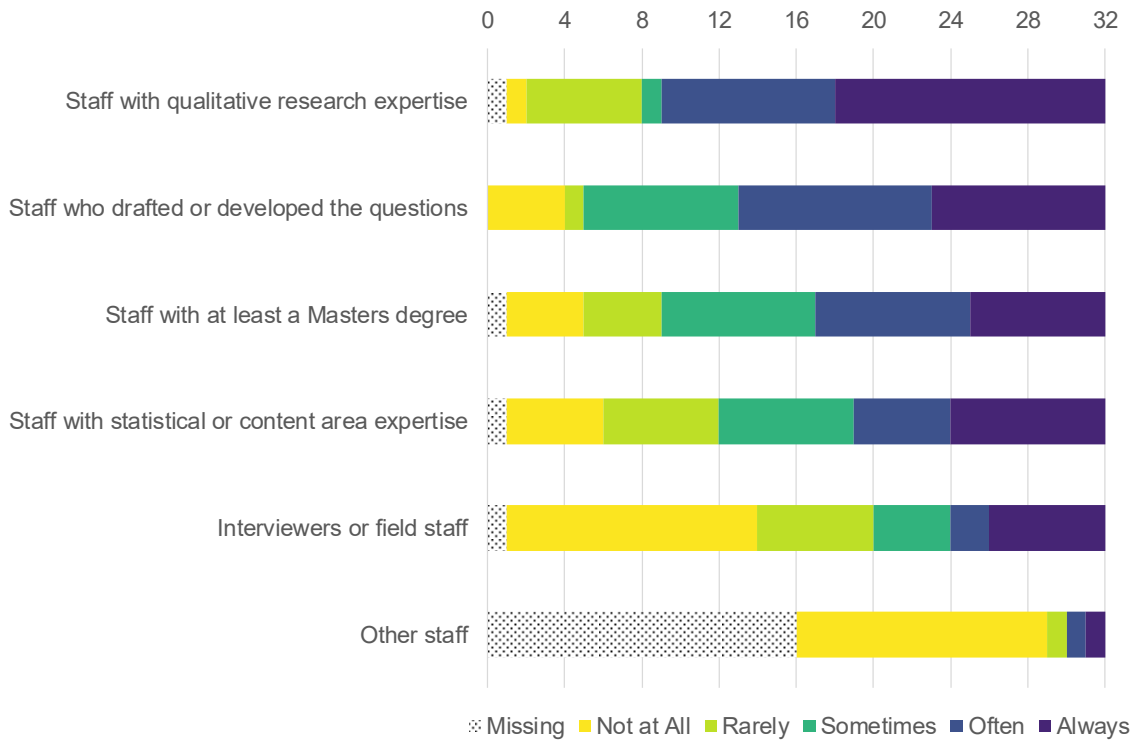

Figure 9: Staff conducting qualitative interviews (q28).

5.4 Data capture

Recommendation 10 suggests capturing data as naturally as possible and paying attention to non-verbal actions. In our survey 27 out of 30 NSIs indicated that they captured interviews "Often/Always" by taking notes (see Figure 10). Audio recording was "Often/Always" made by 11 NSIs, the same number was reported for the collection and entry of completed questionnaires. Less frequently "Often/Always" used were professional on-site transcription (4 NSIs), video recording (3 NSIs) and eye tracking (1 NSI). Respondents provided examples in responses to open-ended questions about other ways qualitative interviews were conducted and documented, including: (i) have an interviewer and a note taker conduct the interviews, (ii) in-situ (site of origin) noting on the paper questionnaire; or (iii) taking notes on enlarged screen shots or web pages.

One of the NSIs provided the following details about their data collection and capture practices, which shows a developed appreciation of the requirements for good documentation: "*We have always had the protocol of taking a copy of the paper form (if that is what we are testing) to the interview. Both the interviewer and the note-taker have a copy of the form and they will also have a sheet with roughly-scripted probes (for known/suspected issues and general 'how was X for you' topics) and both of these are used by the team to note in-situ what did or did not work, significant data, behaviour etc. We have a protocol of, after each interview - usually by end of day - the interview team writes up their notes and actively recalls what happened in each interview. This allows us to get the finer detail that otherwise might be lost, fills in gaps between/among the interviewer and note-taker and is a vital check on the quality of note-taking. We discourage teams from delaying this second stage as memory declines rapidly and test participants merge together. […] For usability testing, we have a protocol of ensuring we can see the respondents' screen - this is not easy in a business environment but we do not conduct the interview without a view of the screen. For test documentation, we screenshot every single page in an online test form and put those into power point docs. These images are large enough for interviewers to quickly circle and document usability and subject-matter issues. Each slide has lines for interviewers to make general notes. Sometimes we also use a standard checklist on each slide so that the test team*

can just tick an issue rather write notes. We then have the same protocol of test teams meeting after the interview to write up their notes."

Surveyed NSIs also reported that typically two staff members from an NSI attended the interviews. The description above shows that even without audio- or video recording efforts can be made to recall and record details of the interview as well as possible. However, overall there seems to be room for improvement in the method of data capture. A substantial group (12 out of 29 NSIs) report that they "Not at all/Rarely" use any kind of retraceable objective capture of what the respondent actually said in the interview (audio recording, video recording or on-site transcription). Any summary by the researcher of the exact words or behaviour of a respondent is a form of data reduction. Although this data reduction is essential for a meaningful interpretation of the data, it comes with the risk of making mistakes in understanding and interpreting the relative importance of different elements of the data. Capturing the data in as unfiltered a manner as possible (e.g. by audio recording) allows the researcher or others later on in the research process to go back to the raw data to consider judgments made.

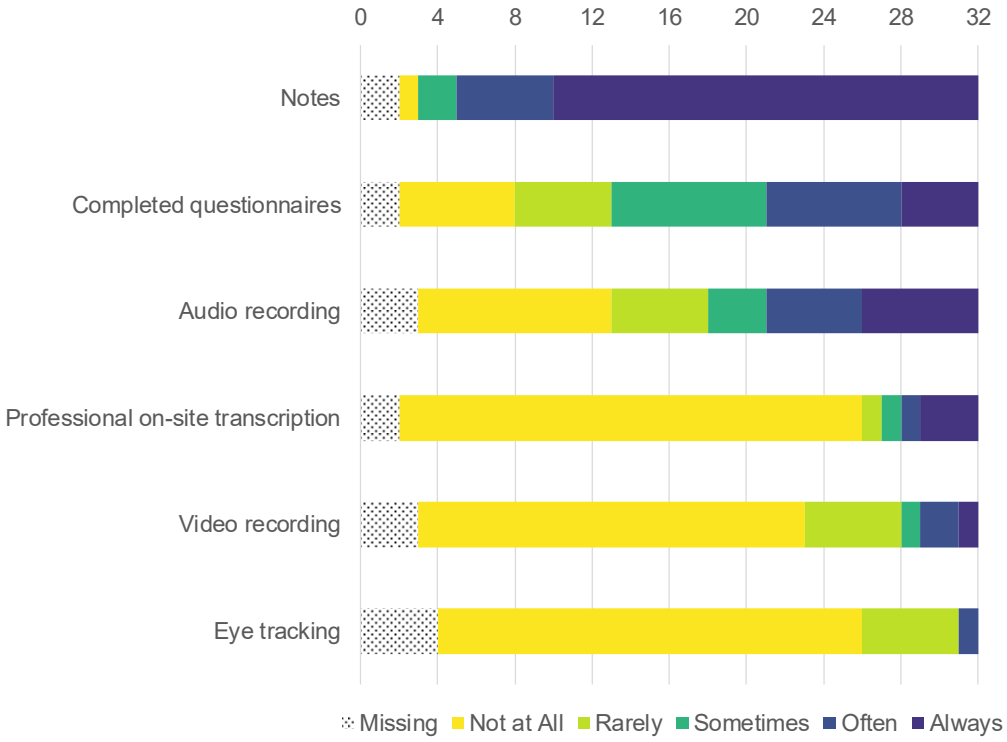

Figure 10: Ways of documenting qualitative interviews (q43).

5.5 Data analysis

In our survey, various aspects of the analysis process were assessed. Figure 11 provides an overview of how often specific practices were used in analysing qualitative interview data. 17 out of 30 NSIs summarise data in a standardised format, which is in line with Recommendation 11, although the relatively high proportion that do not gives some cause for concern.

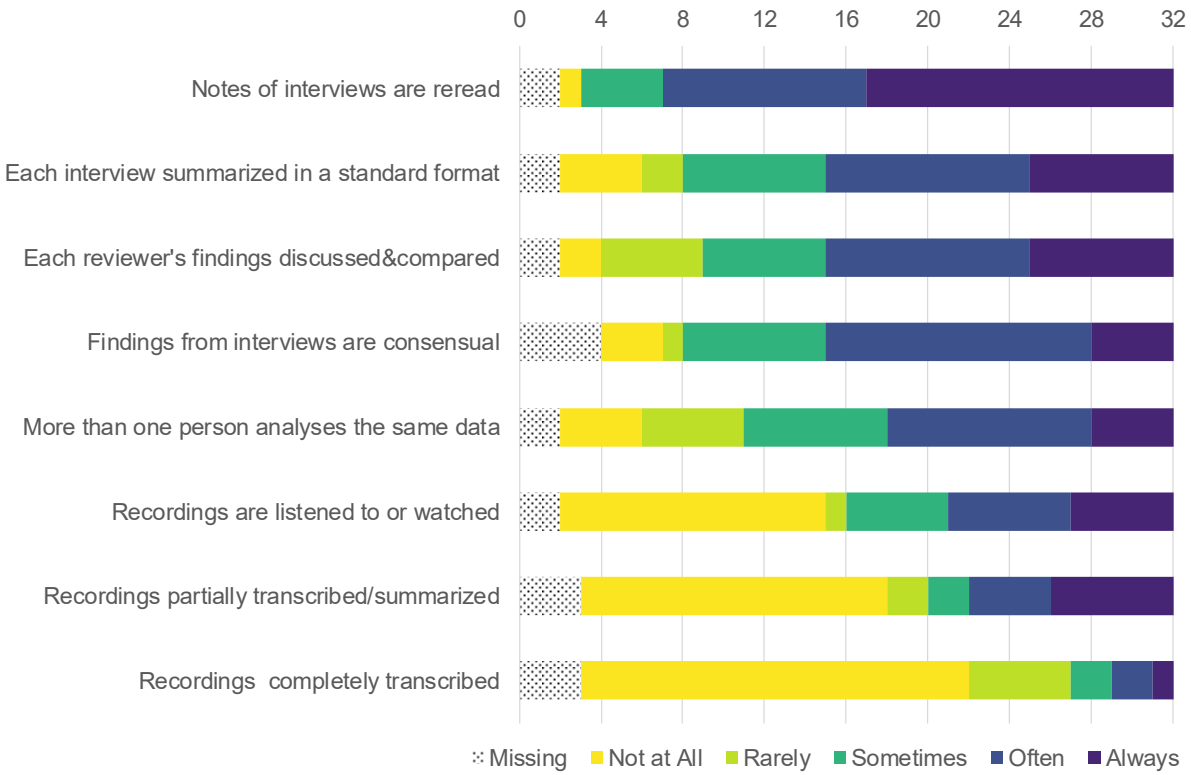

Figure 11: Data analysis practices (q49).

Recommendation 12 suggests in-depth analysis. The practices of listening to recordings, partially transcribing or summarising recordings and complete transcription of recordings all indicate that during analysis researchers had access to unedited versions of what respondents actually said. 15 NSIs reported that they "Often/Always" used at least one of these three practices. This means that about half of the surveyed NSIs did not use recordings or transcripts in their analysis. Out of 27 NSIs that "Often/Always" made notes, 14 always and 10 often reread notes. 10 out of 30 NSIs "Often/Always" coded the data from the qualitative interviews and seven of them "Often/Always" used standardised coding schemes. Nine NSIs provided in an open answer format more information about their coding system. Four of them noted that usability testing was more apt for coding and/or needed different codes. One NSI explained: *"Each interpretive note or observation we classify on the basis if it is related to usability or response process. Further we have two different coding schemes for usability and response process issues."*

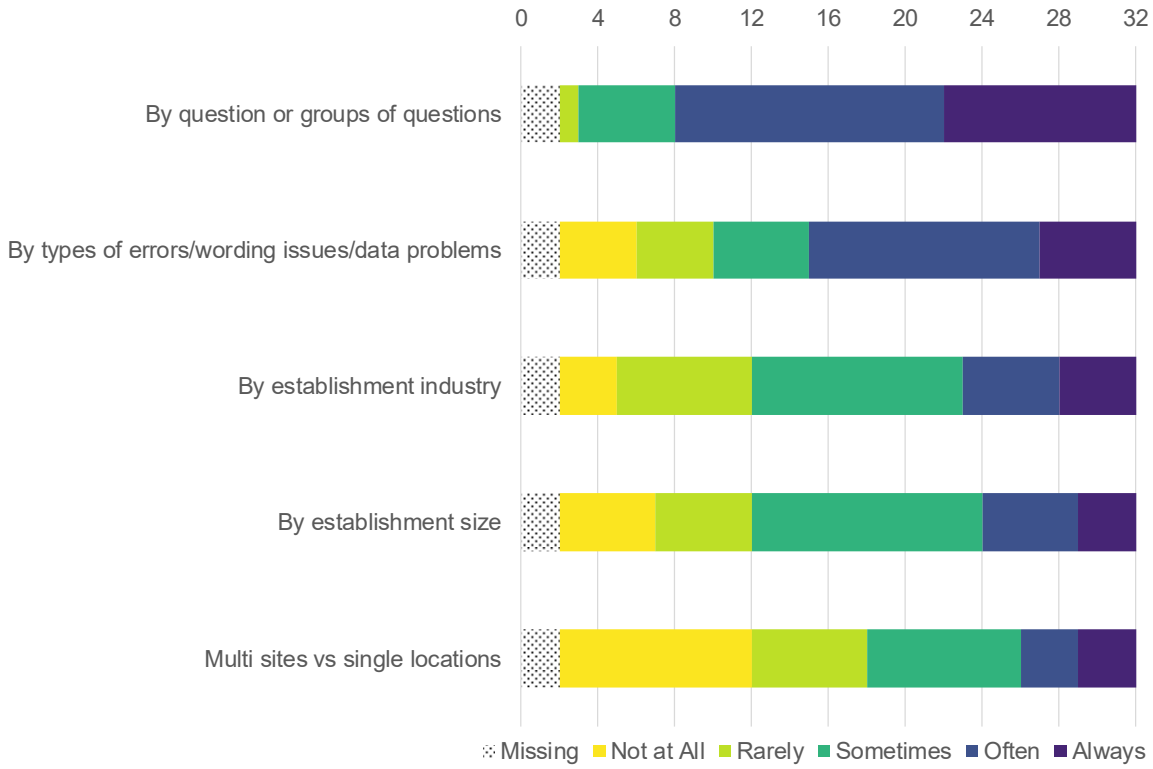

Figure 12: Ways of summarising qualitative data (q50).

Recommendation 13 suggests analysing data by question, characteristics used in sampling, sources of measurement errors and parts of the response process. Figure 12 shows that summarising at the question level (or meaningful part of the questionnaire) was "Often/Always" done in 24 out of 30 NSIs. This makes sense as this way the findings can be used directly to improve the questionnaire tested. Summaries by type of errors, wording issues or data problems were "Often/Always" done in 17 out of 30 NSIs. This suggests possibilities for gaining insights that go further than finding and fixing problems. Summaries by characteristics used in sampling (e.g. industry, size and number of locations) or in the qualitative design that could support Recommendation 7 (implications of sample (non-)representativity) were not prepared so often – "Often/Always" in 9 or fewer out of 30 NSIs.

Recommendation 14 suggests involving more than one researcher in data analysis. Three practices in Figure 11 reflect research approaches that allow for multiple persons to interpret the data: more than one person analysed the data, discussing and comparing to others each reviewer's findings, and seeking consensus of all people involved. 23 NSIs reported that they "Often/Always" used at least one of these practices. We asked who usually analysed the data from qualitative interviews. For 23 out of 31 NSIs qualitative interviews were "Often/Always" analysed by the same person who conducted the qualitative interview. Other people in the central qualitative testing team were "Often/Always" involved in the analysis in 12 out of 30 NSIs. Other people from the business area survey team were "Often/Always" involved in the analysis in 8 out of 30 NSI, and contracting out of the analysis was hardly done (23 out of 29 NSIs never did this, three "Often/Always" did it). For the eight NSIs who said data were only sometimes or less frequently analysed by the same person who conducted the qualitative interview we checked what they had said about their documentation practices. For these NSIs we found that only two "Often/Always" made some type of objective documentation of the interview (audio or video recording or on site transcription). This risks some misinterpretation between the different participants in the process, and is an area where some small changes in practice would reduce the risk of poor outcomes.

We also asked open questions about who was involved in proposing recommendations and how they were reached. The answers revealed a great variety of practices, even within a single organisation. One NSI said that with lots of variation this process might resemble "the sausage-making type" when nobody knew what was inside and some did not want to know. Some made it very transparent and documented initial recommendations from the testing and how they may be changed by discussing them with

stakeholders. Many NSIs mentioned involving several stakeholders and several types of expertise (often mentioning methodology, content matter and IT).

When deciding on recommendations, several strategies were mentioned, for instance, greatest impact on the quality of estimates (that means prioritising units contributing the most to the estimates rather than the number of units, which would also support Recommendation 7 on implications of sample (non-)representativity). Another consideration was the ease of fixing the problem, such as dealing first with issues that can be more easily corrected with some standard/conventional design solutions or guidelines, then focusing on finding new solutions to the issues that have been recognised as not working because this might involve some interface sketching or prototyping with the developers.

5.6 Data reporting and beyond

Our survey suggests that NSIs mainly have documentation practices in place, which is in line with Recommendation 15 to document the study design and results: 15 out of 30 NSIs always documented qualitative research studies in a report, and a further eight did it often. 20 NSIs also "Often/Always" presented results in an internal meeting while presenting results outside the organisation was less common with just three NSIs "Often/Always" doing it. In an open question, four NSIs mentioned that their reports were stored in a document repository for further consultation by interested parties, which is in line with Recommendation 17 to provide access to documentation. Apart from this, it is not known whether the documentation is accessible and to whom (as we did not ask explicitly).

To get an insight into the contents of reports and how much Recommendation 16 about disclosing all methodological details is applied, we asked how often specific methodological information was part of the report. We addressed four out of ten categories on the CIRF checklist:

- Participant selection (CIRF Category 4): the number and type of participating businesses.
- Data collection (CIRF Category 5): how the data were collected.
- Data analysis (CIRF Category 6): how the data were analysed.
- Report format (CIRF Category 7): tested survey questions or data collection instrument, and questions asked.

As Figure 13 shows, information on participating businesses and data collection was regularly part of the report: 18-20 NSIs out of 28 responding NSIs always included it, and a further 3-5 NSIs included it often. No responding NSI claimed to leave this information out. Other methodological information from our list appeared somewhat less frequently, though still quite regularly (in line with recommendations 16). A description of data analysis was "Often/Always" included in 20 NSIs. The tested survey questions or data collection instruments were "Often/Always" included by 23 NSIs. Questions asked about the tested survey questions were "Often/Always" included by 17 NSIs.

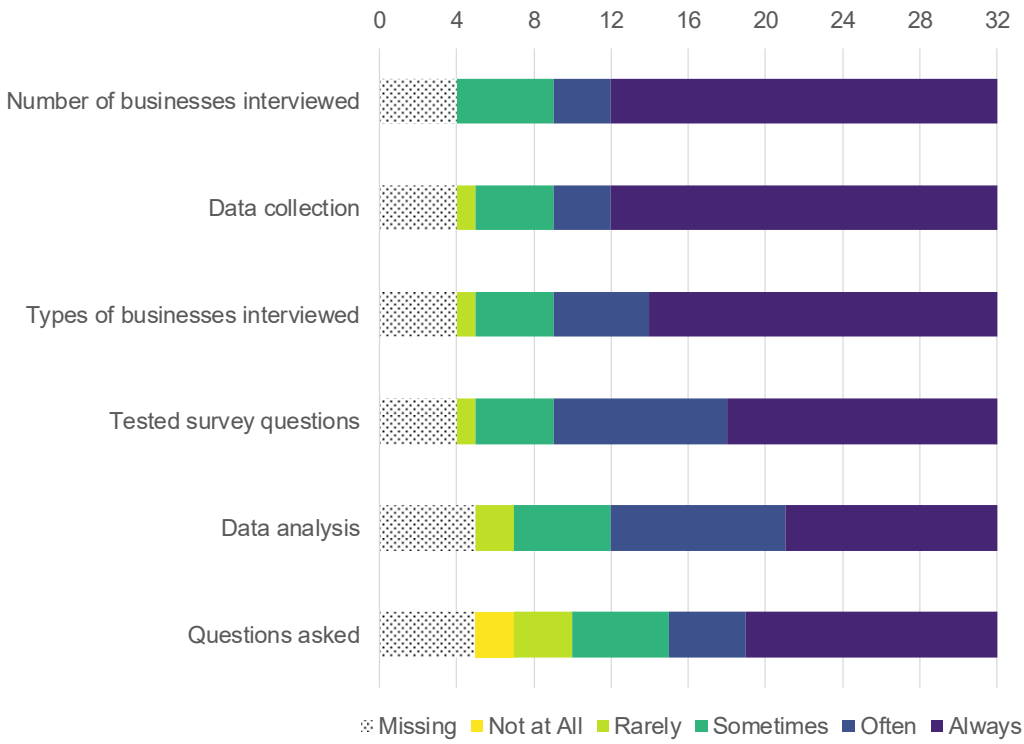

Figure 13: Information included in reports (q54).

Even if these figures are quite encouraging, three points have to be made. First, some NSIs still failed to report some key aspects of the study, such as which questions were asked. This might be sufficient for certain, probably short-term purposes but jeopardises the usefulness of such reports for reuse and contribution to general knowledge. Second, we addressed neither all categories of the CIRF checklist nor all the items within the included categories, so we do not have an overall picture of reporting completeness. Third, as Boeije & Willis (2013) note, blanket statements may purport to fulfil a category on the CIRF checklist but actually are not very informative – one potential area is the CIRF category of data analysis, because analytic procedures are rarely described clearly enough to allow replication or judge the reliability of the findings. Reports might also fail to incorporate a self-assessment of the methodological strengths and weaknesses (recommendation 15), which is important not only to guide immediate decisions but also to inform future users of the report. However, 20 out of 32 NSIs reported "Often/Always" including a description of the data analysis, and this is encouraging, although we did not gather sufficient detail to judge the comprehensiveness of descriptions.

Still, 27 out of 31 NSIs reported that recommendations from qualitative interviewing were "Often/Always" implemented (recommendation 18), though open answers suggested that final recommendations were already adjusted to known constraints, e.g. software limitations, a tight schedule etc. On the other hand, 17 of 28 NSIs reported that they never or rarely used other data to check evaluate the impact of the recommendations on data quality (recommendation 19).

6. DISCUSSION AND CONCLUSION

The literature review indicated many soft rules and not so much hard evidence, especially in the field of establishment surveys. Our recommendations provide some guidance on good practices but may be further elaborated as new evidence becomes available. There is a particular need to link quantitative evaluations of survey quality with the results of interventions derived from qualitative analyses, and we suggest this as a fruitful area for further research.

Our low contact and response rate showed that, at least for this study, it was difficult to get a broad overview of all NSIs worldwide. This may be due to limitations of our design and general challenges

related to collecting data from large establishments, but it may also be because the topic does not appear clear or relevant to many NSIs. Among the NSIs responding to the survey, many said that they did not use qualitative testing at all, and at least some responses indicated confusion about what qualitative methods for evaluating establishment questionnaires were. Most of the analysed NSIs came from Europe and Northern America, which confirms the presence of qualitative testing in these geographical regions; none of the excluded NSIs came from these regions. NSIs stating that they were not using qualitative testing came from all around the world.

In NSIs that declared some use of qualitative methods, many recommendations are already followed, but there is still plenty of room for improvement to increase the quality and accountability of the research process. The most critical areas, where our data showed that practices were least developed, concern documentation of recruitment practices (recommendation 6), capture of data collection (e.g. audiotaping) (recommendation 10), depth of analysis (recommendation 13) and post-testing evaluation (recommendation 19). These are also the areas that future research should address.

The fact is that design processes are done wherever there are business surveys. The question is whether these processes are somehow formalised or rather informal, and exactly what activities they embrace. Without the ambition of developing an exhaustive classification, we noticed a variety of approaches to qualitative questionnaire testing in NSIs when studying their eligibility for our survey (see section 4.2) and data on competent staff and other aspects of work organisation. Some categories to represent the use of the variety of qualitative approaches to questionnaire testing in NSIs are:

- No awareness of qualitative methods being useful in survey design processes.
- Qualitative methods are known but are practiced formally only in household surveys.
- Qualitative methods are known but are practiced informally, with untrained staff.
- Qualitative methods form part of established methods in development, testing and evaluation of establishment surveys.

The analysed sample in our study consists mainly of NSIs classified in the last category (e.g. 23 NSIs Often/Always used staff with specialized knowledge or experience in qualitative research and among them 17 NSIs “Often/Always” produced a report), which is to be expected – these are NSIs who are more likely to regard our survey as relevant (see e.g. Groves, Singer, & Corning, 2000). To the best of our knowledge, this sample includes all NSIs that are known for their use of qualitative testing in establishment surveys (because of their publications and presentations at scientific meetings). If NSIs are assumed to be at the leading edge of applying emerging methodologies, reported practices are likely descriptive of current best practices in establishment surveys and can serve as a benchmark. Nevertheless, we cannot completely rule out the possibility that some sort of qualitative testing is somewhat more widespread than our survey suggests.

Our study, even with this selective response, provides evidence on the prevalence and use of types of qualitative testing for establishment surveys. The results provide a snapshot of the state of practice for qualitative testing and may be an incentive for developing questionnaire testing plans. Developing a qualitative testing plan is recommended for NSIs, to be successful at meeting the changing needs for new data and to undertake complex data collections. Our study is valuable in identifying the prevalence of elements that make up qualitative questionnaire testing, and the recommendations serve as a checklist of the methods for NSIs to consider while preparing for upcoming data collections.

Given the availability of various handbooks and standards recommending the use of a range of qualitative methods in questionnaire design and testing, our major concern is with those NSIs that lack awareness of the benefits of these methods. We call for more promotion, and education about the benefits, of qualitative methods, particularly for establishment surveys. NSIs that are already aware of qualitative methods seem more likely to reach out for more information (e.g. available standards and handbooks). NSIs that have formalised the use of qualitative methods in questionnaire design processes are not at the end of their path either. They can work on fine tuning their own methods and, even more importantly, they are in the best position to conduct scientific research that would empirically test many soft rules and provide hard evidence as a basis for deciding whether there can be general recommendations, and developing them if appropriate.

ACKNOWLEDGMENT

We would like to take the opportunity to thank all the participants to our survey that has hopefully helped advance thinking on the use of qualitative testing methods for establishment surveys. We would also like to thank Ger Snijkers, Glenn White and three anonymous reviewers who provided many helpful comments. We also express sincere thanks to the staff of the SESRC at Washington State University for collaboration in implementing the web survey.

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of their employers.

REFERENCES

Baena, B. I., & Padilla, J. L. (2014). Cognitive Interviewing in Mixed Research. In K. Miller, S. Willson, V. Chepp & J. L. Padilla (Eds.), *Cognitive Interviewing Methodology* (pp. 133-152): Wiley.

Bavdaž, M. (2009). Conducting research on the response process in business surveys. *Statistical Journal of the IAOS, 26*(1-2), 1-14. doi: 10.3233/sji-2009-0692

Bavdaž, M. (2010a). The multidimensional integral business survey response model. *Survey Methodology, 36*(1), 81-93.

Bavdaž, M. (2010b). Sources of measurement errors in business surveys. *Journal of Official Statistics, 26*(1), 25-42.

Beatty, P., Schechter, S., & Whitaker, K. (1997). *Variation in cognitive interviewer behavior - extent and consequences*. Paper presented at the Proceedings of the Survey Research Methods Section, Alexandria, VA. http://www.amstat.org/sections/srms/Proceedings/papers/1997_183.pdf

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly, 71*(2), 287-311. doi: 10.1093/poq/nfm006

Blair, J., & Presser, S. (1993). *Survey procedures for conducting cognitive interviews to pretest questionnaires: A review of theory and practice*. Paper presented at the Proceedings of the Survey Research Methods Section, Alexandria, VA. http://www.amstat.org/sections/srms/Proceedings/papers/1993_059.pdf

Blake, M. (2015). Other pretesting methods. In D. Collins (Ed.), *Cognitive Interviewing Practice* (pp. 28-56). London: SAGE Publications Ltd.

Boeije, H., & Willis, G. (2013). The Cognitive Interviewing Reporting Framework (CIRF). *Methodology, 9*(3), 87-95. doi: 10.1027/1614-2241/a000075

Brancato, G., Macchia, S., Murgia, M., Signore, M., Simeoni, G., Blanke, K., . . . Hoffmeyer-Zlotnik, J. H. P. (2005). Handbook of Recommended Practices for Questionnaire Development and Testing in the European Statistical System Retrieved from http://ec.europa.eu/eurostat/documents/64157/4374310/13-Handbook-recommended-practices-questionnaire-development-and-testing-methods-2005.pdf

Bureau, M. (1991). *Experience with the use of cognitive methods in designing business survey questionnaires*. Paper presented at the Proceedings of the Surveys Research Methods Section, Alexandria, VA. http://www.amstat.org/sections/srms/Proceedings/papers/1991_123.pdf

Charmaz, K. (2005). Grounded theory in the 21st century: Applications for advancing social justice studies. In N. K. Denzin & Y. S. Lincoln (Eds.), *The SAGE Handbook of Qualitative Research* (3rd ed., pp. 507-535). Thousand Oaks, CA: Sage Publications.

Collins, D. (2015a). Analysis and Interpretation. In D. Collins (Ed.), *Cognitive Interviewing Practice* (pp. 162-174). London: SAGE Publications Ltd.

Collins, D. (2015b). *Cognitive Interviewing Practice*. London: SAGE Publications Ltd.

Collins, D., & Gray, M. (2015). Sampling and Recruitment. In D. Collins (Ed.), *Cognitive Interviewing Practice* (pp. 80-100). London: SAGE Publications Ltd.

Cox, B. G., & Chinnappa, B. N. (1995). Unique features of business surveys. In B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge & P. S. Kott (Eds.), *Business survey methods* (pp. 1-17). New York: Wiley-Interscience.

D'Ardenne, J., & Collins, D. (2015). Data Management. In D. Collins (Ed.), *Cognitive Interviewing Practice* (pp. 142-161). London: SAGE Publications Ltd.

Davidsson, G. (2002, 6.8.2006). Cognitive testing of mail surveys at Statistics Sweden. *International Conference on Questionnaire Development, Evaluation and Testing Methods (QDET), 2002, Charleston*, from http://www.jpsm.umd.edu/qdet/final_pdf_papers/round%20two/davidsson.pdf

Davis, W., DeMaio, T. J., & Zukerberg, A. (1995). *Can cognitive information be collected through the mail? Comparing cognitive data collected in written versus verbal format*. Paper presented at the 50th Annual Conference of the American Association for Public Opinion Research, May 18-21, 1995, Fort Lauderdale.

DeMaio, T. J., & Jenkins, C. R. (1991). *Questionnaire research in the Census of Construction Industries*. Paper presented at the Proceedings of the Survey Research Methods Section, Alexandria, VA. http://www.amstat.org/sections/srms/Proceedings/papers/1991_083.pdf

DeMaio, T. J., & Landreth, A. (2004). Do Different Cognitive Interviewer Techniques Produce Different Results? In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin & E. Singer (Eds.), *Methods for Testing and Evaluating Survey Questionnaires* (pp. 89-108). Hoboken, NJ: Wiley-Interscience.

Denzin, N. K., & Lincoln, Y. S. (2005). Introduction: The Discipline and Practice of Qualitative Research. In N. K. Denzin & Y. S. Lincoln (Eds.), *The SAGE Handbook of Qualitative Research* (3rd ed., pp. 1-41). Thousand Oaks, CA: Sage Publications.

Economic and Social Commission for Asia Pacific Region. (2010). Guidelines for cognitive and pilot testing of questions for use in surveys, from http://www.washingtongroup-disability.com/wp-content/uploads/2016/02/Disability-question-testing-guidelines.pdf

Edwards, W. S., & Cantor, D. (1991). Toward a response model in establishment surveys. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz & S. Sudman (Eds.), *Measurement Errors in Surveys* (pp. 211-233). New York: Wiley-Interscience.

Eldridge, J., Martin, J., & White, A. (2000). *The use of cognitive methods to improve establishment surveys in Britain*. Paper presented at the ICES-II Proceedings of the Second International Conference on Establishment Surveys, Buffalo, NY.

European Statistical System Committee. (2011). European Statistics Code of Practice Revised edition. Retrieved November 23, 2016, from http://ec.europa.eu/eurostat/web/quality/european-statistics-code-of-practice

Forsyth, B. H., & Lessler, J. T. (1991). Cognitive laboratory methods: a taxonomy. In P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz & S. Sudman (Eds.), *Measurement Errors in Surveys* (pp. 393-418). New York: Wiley-Interscience.

Forsyth, B. H., Weiss, E. S., & Miller Anderson, R. (2003). A comparison of appraisal and cognitive interview methods for pretesting organizational survey questionnaires *2003 Joint Statistical Meetings - Section on Government Statistics* (pp. 1492-1499). Alexandria, VA: American Statistical Association.

Fusch, P. I., & Ness, L. R. (2015). Are we there yet? Data saturation in qualitative research. *The Qualitative Report, 20*(9), 1408-1416.

Gerber, E. R. (1999). The view from anthropology: ethnography and the cognitive interview. In M. G. Sirken, D. J. Herrmann, S. Schechter, N. Schwarz, J. M. Tanur & R. Tourangeau (Eds.), *Cognition and survey research* (pp. 217-234). New York: Wiley-Interscience.

Giesen, D. (2007). *The Response Process Model as a Tool for Evaluating Business Surveys*. Paper presented at the Third International Conference on Establishment Surveys (ICES-3), 18-21 June, Montreal, Canada. https://ww2.amstat.org/meetings/ices/2007/proceedings/ICES2007-000056.PDF

Giesen, D., & Hak, T. (2005). *Revising the Structural Business Survey: From a Multi-Method Evaluation to Design*. Paper presented at the 2005 Federal Committee on Statistical Methodology Conference, 14-16 November, Arlington, Virginia.

Gower, A. R., & Nargundkar, M. S. (1991). *Cognitive aspects of questionnaire design: business surveys versus household surveys*. Paper presented at the 1991 Annual Research Conference, March 17-20, 1991, Arlington.

Gray, M. (2015). Survey Mode and its Implications for Cognitive Interviewing. In D. Collins (Ed.), *Cognitive Interviewing Practice* (pp. 197-219). London: SAGE Publications Ltd.

Groves, R. M. (1996). How do we know what we think they think is really what they think? In N. Schwarz & S. Sudman (Eds.), *Answering questions: methodology for determining cognitive and communicative processes in survey research* (pp. 389-402). San Francisco: Jossey-Bass Publishers.

Groves, R. M., Singer, E., & Corning, A. (2000). Leverage-Saliency Theory of Survey Participation. *Public Opinion Quarterly, 64*, 299-308.

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field Methods, 18*(1), 59-82.

International Statistical Institute. (2016). National Statistical Offices Retrieved March 11, 2016, from https://www.isi-web.org/index.php/resources/national-statistical-offices

Jones, J. (2003). A framework for reviewing data collection instruments in business surveys. *Survey Methodology Bulletin*(52), 4-9.

Kruskal, W., & Mosteller, F. (1979). Representative sampling, III: The current statistical literature. *International Statistical Review, 47*, 245-265.

Kvale, S. (1996). *InterViews: an introduction to qualitative research interviewing*. Thousand Oaks, CA: Sage Publications.

Lorenc, B. (2006). *Two topics in survey methodology: modelling the response process in establishment surveys; inference from nonprobability samples using the double samples setup (doctoral dissertation).* Stockholm University, Stockholm.

McCarthy, J., Ott, K., Ridolfo, H., McGovern, P., Sirkis, R., & Moore, D. (2018). Combining Multiple Methods in Establishment Questionnaire Testing: The 2017 Census of Agricultural Testing Bento Box. *Journal of Official Statistics, 34*(2), 341-364.

Miller, G., & Fox, K. J. (2004). Building bridges: the possibility of analytic dialogue between ethnography, conversation analysis and Foucault. In D. Silverman (Ed.), *Qualitative research: Theory, method and practice* (2nd ed., pp. 35-55). London ; Thousand Oaks, CA: Sage Publications.

Miller, K. (2011). Cognitive Interviewing. In J. Madans, K. Miller, A. Maitland & G. Willis (Eds.), *Question Evaluation Methods* (pp. 51-75): Wiley.

Miller, K., Willson, S., Chepp, V., & Padilla, J. L. (Eds.). (2014). *Cognitive Interviewing Methodology*: Wiley.

Mockovak, W., & Kaplan, R. (2015). Comparing Results from Telephone Reinterview with Unmoderated, Online Cognitive Interviewing *Proceedings of the American Association for Public Opinion Research, 2015* (pp. 4183-4193). Alexandria, VA: American Statistical Association.

Mohorko, A., & Hlebec, V. (2015). Effect of a first-time interviewer on cognitive interview quality.. *Quality & Quantity, 49*(5), 1897-1918. doi: 10.1007/s11135-014-0081-0

Mohorko, A., & Hlebec, V. (2016). Degree of cognitive interviewer involvement in questionnaire pretesting on trending survey modes. *Computers in Human Behavior, 62*, 79-89. doi: 10.1016/j.chb.2016.03.021

Neuert, C. E., & Lenzner, T. (2016). Incorporating eye tracking into cognitive interviewing to pretest survey questions. *International Journal of Social Research Methodology, 19*(5), 501-519. doi: 10.1080/13645579.2015.1049448

Nichols, E., & Childs, J. H. (2009). Respondent Debriefings Conducted by Experts: A Technique for Questionnaire Evaluation. *Field Methods, 21*(2), 115-132. doi: 10.1177/1525822x08330265

Nielsen, J., & Landauer, T. K. (1993). *A mathematical model of the finding of usability problems.* Paper presented at the Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems.

O'Brien, E., Fisher, S. K., Goldenberg, K., & Rosen, R. (2001). Application of cognitive methods to an establishment survey: a demonstration using the Current Employment Statistics survey *Proceedings of the Annual Meeting of the American Statistical Association*. Alexandria, VA: American Statistical Association.

O'Reilly, M., & Parker, N. (2013). 'Unsatisfactory Saturation': a critical exploration of the notion of saturated sample sizes in qualitative research. *Qualitative Research, 13*(2), 190-197. doi: 10.1177/1468794112446106

Office of Management and Budget. (2006). Standards and Guidelines for Statistical Surveys Retrieved November 23, 2016, from https://unstats.un.org/unsd/dnss/docs-nqaf/USA_standards_stat_surveys.pdf

Office of Management and Budget. (2016). Statistical Policy Directive No. 2 Addendum: Standards and Guidelines for Cognitive Interviews Retrieved November 8, 2018, from https://obamawhitehouse.archives.gov/sites/default/files/omb/inforeg/directive2/final_addendum_to_stat_policy_dir_2.pdf

Palmisano, M. (1988). *The application of cognitive survey methodology to an establishment survey field test*. Paper presented at the Proceedings of the Surveys Research Methods Section, Alexandria, VA. http://www.amstat.org/sections/srms/Proceedings/papers/1988_031.pdf

Patton, M. Q. (1999). Enhancing the Quality and Credibility of Qualitative Analysis. *Health Services Research, 34*(5 Part 2), 1189-1208.

Persson, A., Björnram, A., Elvers, E., & Erikson, J. (2015). A strategy to test questionnaires at a national statistical office. *Statistical Journal of the IAOS, 31*(2), 297-304. doi: 10.3233/sji-140863

Redline, C., Smiley, R., Lee, M., & DeMaio, T. J. (1998). *Beyond concurrent interviews: an evaluation of cognitive interviewing techniques for self-administered questionnaires*. Paper presented at the Proceedings of the Survey Research Methods Section, Alexandria, VA.

Rivière, P. (2002). What makes business statistics special? *International Statistical Review, 70*(1), 145-159.

Roller, M. R., & Lavrakas, P. J. (2015). *Applied Qualitative Research Design: A Total Quality Framework Approach*. New York: The Guilford Press.

Shenton, A. K. (2004). Strategies for ensuring trustworthiness in qualitative research projects. *Education for Information, 22*(2), 63-75.

Snijkers, G. (2002). *Cognitive laboratory experiences: on pre-testing computerised questionnaires and data quality (doctoral dissertation).* Universiteit Utrecht, Utrecht. Retrieved from http://www.jpsm.umd.edu/qdet/final_pdf_papers/Snijkers.pdf

Snijkers, G., & Bavdaž, M. (2011). Business Surveys. In M. Lovric (Ed.), *International Encyclopedia of Statistical Science* (pp. 191-194). Berlin, Heidelberg: Springer Berlin Heidelberg.

Stapleton Kudela, M., Forsyth, B. H., Levin, K., Lawrence, D., & Willis, G. (2006). Cognitive Interviewing versus Behavior Coding *The 61st Annual Conference of the American Association for Public Opinion Research, May 18-21, 2006, Montréal, Quebec* (pp. 4243-4249). Alexandria, VA: American Statistical Association.

Sudman, S., Willimack, D. K., Nichols, E., & Mesenbourg, T. L. (2000). *Exploratory research at the U.S. Census Bureau on the survey response process in large companies*. Paper presented at the ICES-II Proceedings of the Second International Conference on Establishment Surveys, Buffalo, NY.

Tourangeau, R. (2004). Experimental Design Considerations for Testing and Evaluating Questionnaires. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin & E. Singer (Eds.), *Methods for Testing and Evaluating Survey Questionnaires* (pp. 209-224). Hoboken, NJ: Wiley-Interscience.

Tuttle, A. D., Morrison, R. L., & Willimack, D. K. (2010). From Start to Pilot: A Multi-method Approach to the Comprehensive Redesign of an Economic Survey Questionnaire. *Journal of Official Statistics, 26*(1), 87-103.

United Nations. (2016). Composition of macro geographical (continental) regions, geographical sub-regions, and selected economic and other groupings Retrieved June 6, 2016, from http://unstats.un.org/unsd/methods/m49/m49regin.htm

Ursachi, K., & Jones, J. (2005). Recruiting respondents for a qualitative study of the availability of local unit level business data. *Survey Methodology Bulletin*(56), 66-73.

Willimack, D. K. (2013). Methods for the Development, Testing, and Evaluation of Data Collection Instruments. In G. Snijkers, G. Haraldsen, J. Jones & D. K. Willimack (Eds.), *Designing and Conducting Business Surveys*: Wiley.

Willimack, D. K., Lyberg, L. E., Martin, J., Japec, L., & Whitridge, P. (2004). Evolution and Adaptation of Questionnaire Development, Evaluation, and Testing Methods for Establishment Surveys. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin & E. Singer (Eds.), *Methods for Testing and Evaluating Survey Questionnaires* (pp. 385-407). Hoboken, NJ: Wiley-Interscience.

Willimack, D. K., & Nichols, E. (2010). A Hybrid Response Process Model for Business Surveys. *Journal of Official Statistics, 26*(1), 3-24.

Willis, G. B. (2005). *Cognitive interviewing : a tool for improving questionnaire design*. Thousand Oaks, CA: Sage Publications.

Willis, G. B. (2015). *Analysis of the Cognitive Interview in Questionnaire Design*: Oxford Universtiy Press.

Willis, G. B., & Boeije, H. (2013a). Reflections on the Cognitive Interviewing Reporting Framework. *Methodology, 9*(3), 123-128. doi: 10.1027/1614-2241/a000074

Willis, G. B., & Boeije, H. (2013b). The Survey Field Needs a Framework for the Systematic Reporting of Questionnaire Development and Pretesting. *Methodology, 9*(3), 85-86. doi: 10.1027/1614-2241/a000070

Willson, S., & Miller, K. (2014). Data Collection. In K. Miller, S. Willson, V. Chepp & J. L. Padilla (Eds.), *Cognitive Interviewing Methodology* (pp. 15-33): Wiley.

Yan, T., Kreuter, F., & Tourangeau, R. (2012). Evaluating Survey Questions: A Comparison of Methods. *Journal of Official Statistics, 28*(4), 503-529.