Asian Journal of Social Science Studies; Vol. 4, No. 2; 2019
ISSN 2424-8517 E-ISSN 2424-9041
Published by July Press

Survey Questionnaire Survey Pretesting Method: An Evaluation of Survey Questionnaire via Expert Reviews Technique

Emmanuel M. Ikart[1]

[1] Corrective Services NSW, Department of Justice, Australia

Correspondence: Emmanuel M. Ikart, Corrective Services NSW, Department of Justice, Australia.

Received: February 27, 2019 Accepted: March 25, 2019 Online Published: April 23, 2019

doi:10.20849/ajsss.v4i2.565 URL: https://doi.org/10.20849/ajsss.v4i2.565

Abstract

Whereas the literature on questionnaire pretesting has revealed a paradox, questionnaire pretesting is a simple technique to measure in advance whether a questionnaire causes problems for respondents or interviewers. Consequently, experienced researchers and survey methodologists have declared questionnaire pretesting indispensable. All the same, published survey reports provide no information about whether a questionnaire was pretested and, if so, how and with what results. Moreover, until recently, there has been limited methodological research on questionnaire pretesting. The universally acknowledged importance of questionnaire pretesting has been honoured more in theory than in practice. As a result, we know very little about pretesting and the extent to which a pretest serves its intended purpose and leads to value-added on questionnaires. An expert review is a traditional method of questionnaire pretesting. Expert reviews can be conducted with varying levels of organisation and rigor. On the lower end of the spectrum, an experienced subject matter expert, or survey methodologist reviews a draft questionnaire to identify issues with question wording or administration that may lead to measurement error. On the more rigorous end of the spectrum, as employed in this study is the Questionnaire Appraisal Scheme method, a standardized instrument review containing 28 problem types that allow experienced researchers and/or coders to code, analyse and compare the results of questionnaire problems reported by the independent expert reviewers for consistency and agreement across the expert reviewers. However, in spite of the wider use of the expert review as a pretest method, fewer empirical evaluations of this method exist. Specifically, there is little evidence as to whether different expert reviews consistently identified similar questionnaire problems. Similarly, there has been no reasonable level of agreements across the expert reviewers in their evaluation of questionnaire problems. This paper addresses these shortcomings. The protocols employed in the paper would contribute to reducing the shortfall in pretesting guidelines and encourage roundtable discussions in academia and management practice.

Keywords: expert reviews, methodologists, questionnaire pretesting, respondents, QAS

1. Introduction

Expert reviews are frequently used as a questionnaire evaluation method. Whereas prior evaluative studies have attempted to assess the effectiveness of expert reviews in improving questionnaire problems (e.g., DeMaio & Landreth, 2003; Yan et al. 2012), they have received limited empirical attention. Up to now, researchers have often asked subject matter experts(e.g., methodologists, sociologists and psychologists) who have theoretical questionnaire knowledge and/or practical experience, to review draft questionnaires and provide a critique of the questions as a technique of spotting questionnaire problems, potential measurement errors or a breakdown in question answering process (Olson, 2010). The expert review provides a fresh set of eyes to critically look at the questions because the developer routinely gets too close to the subject matter to be able to see all the problems (Grealish, 2003). Obviously, expert reviews have become common practice in questionnaire development (Yan et al. 2012). However, expert reviews can be done by having individuals review the questionnaire (Willis et al. 1999), or, by convening a group of experts (called an expert panel), to review questionnaires (DeMaio & Landreth, 2003). Expert reviewers can rely exclusively on their own judgements, making informal assessments that typically yield open ended comments about the survey items to be evaluated (Olson, 2010). They can also be guided by formal appraisal systems that provide a detailed set of potential problem codes (DeMaio & Landreth, 2003; Rothgeb et al. 2001; Yan et al. 2012). Nevertheless, as no field costs are involved, expert reviews are a relatively cheap pretesting method. The number of expert reviewers tends to range from two, three to over twenty reviewers (Presser & Blair, 1994; Willis et al. 1999; Grealish, 2003; DeMaio & Landreth, 2003; Oslon, 2010). Presser and Blair (1994) included expert reviewers in their study on the effectiveness and reliability of different methods of pretesting questionnaires. Moreover, as part of an experiment on alternative cognitive interviewing methods, DeMaio and Landreth (2003), utilized the results from individual experts (three experts from different organisations) to gauge the breadth of results produced by three different teams of cognitive interviewers. Furthermore, Olson (2010) employed expert reviews in their study to measure the variations across experts in their reviews of questionnaire problems. They also assess whether questions identified by experts as problematic are systematically related to data quality problems – item nonresponse and report inaccuracy on those questions. An evaluation of the quality of expert reviews compared to other methods containing proxy indicators of data quality e.g., interviewer and respondent behaviour during an interview, suggested that expert reviewers identify more problems in survey questionnaires than other methods (Presser & Blair, 1994). Tourangeau (2004) asserted that the questions identified by expert reviewers as being "problematic" actually yield higher levels of measurement errors than other questions. Further, Willis et al. (1999) and his colleagues compared expert reviews with cognitive interviewing and behaviour coding to evaluate the number of problems found by each method. They concluded that the expert review found the most problems. However, in spite of the wider use of the expert review, fewer empirical evaluations of this method exist. Specifically, there is little evidence as to whether different expert reviews consistently identified similar questionnaire problems (Oslon, 2010; Tourangeau, 2004; DeMaio & Landreth, 2003). Similarly, there has been no reasonable level of agreements among expert reviewers in their evaluation of questionnaire problems (DeMaio & Landreth, 2003). This study addresses these shortfalls.

The remainder of this paper is organised as followed: Section 2 reviews the related literature beginning with the presentation of a survey design best practices and how to write a good questionnaire. This is followed by the classifications of expert reviewers and main goals of expert reviews. Section 3 presents the research objectives. Section 4 presents the methodology of the study; Section 5 presents the analysis and results and finally, Section 6 presents the discussion and conclusion and implications of the study for academia and management practice.

2. Literature Search

2.1 Survey Design Best Practices: How to Write a Good Questionnaire

When it comes to survey questionnaire design best practices, there are a number of key considerations. These include establishing relevant information, research aims and objectives, data collection methods, clarity and writing style of questionnaire, question structure, look and feel, the flow and questionnaire pretesting (Market Research Guy, 2017; Ikart, 2018). Together, these components, as shown in Figure 1, help construct effective survey questionnaires.

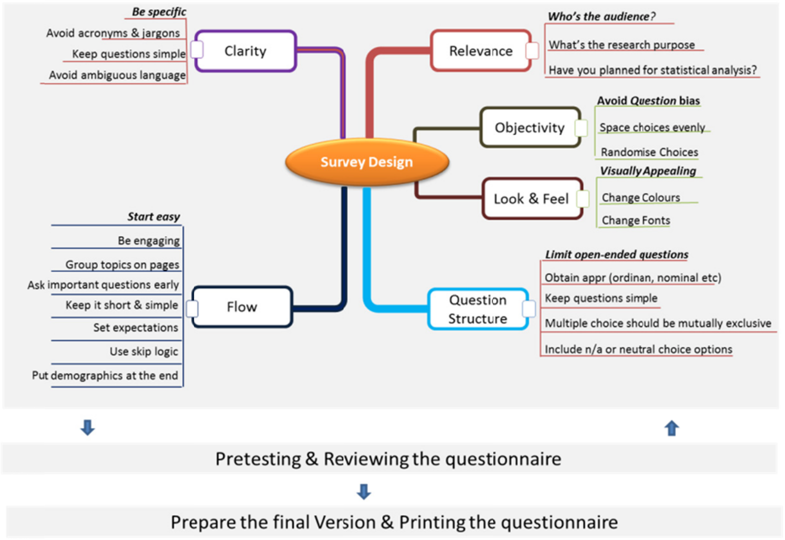


Figure 1. Survey design best practices: how to write a good questionnaire

Source: Adapted from Market Research Guy, 2017

Firstly, the questionnaire should focus on the research aims and objectives. This should include asking and collecting the right types of information and making sure that each question is specific, objective and understandable. Further question options, including use of mutually exclusive multiple-choice questions, rating or ranking scale and closed-ended questions, will produce different types of responses. These options are useful for gathering information about preferences, attitudes, opinions and behaviours. For instance, closed-ended questions help in gathering demographic and other fact-based information which can then be used to classify people or situations. Also, your survey needs to be free of errors so that it assists the researchers to make critical decisions. Although errors may still occur, it is important to be aware that these happen in order to reduce them. However, researchers need to be aware of the types of errors that might impact on the research and ensure that the research objective is appropriate and that the questionnaire survey is trialled before use. Such action will minimise the impact of errors. Moreover, it is a good idea to randomise question choice options so that the same option is not listed first. More often respondents will select choices near the top of the list. But by randomly rotating the order of multiple choice matrix questions for each respondent, you can feel more confident that your results are genuine and not simply the result of being the ones listed first. This helps correct the impact of errors in surveys (Market research Guy, 2017; Ikart, 2018).

Secondly, when writing your questions for a survey, it is best to structure your questionnaire using what is called the “funnel” technique. Start your questionnaire with broad general interest questions that are easy for the respondent to answer. These questions warm up the respondents and get them involved in the survey. The most difficult questions are placed in the middle – those that take time to think about, and those that are of less general interest. At the end, we can again place general questions that are easier to answer and of broad interest and application. Typically, these last questions can include demographic and other classification questions. Consideration should also be given about where to place the most sensitive questions (e.g., in the middle or towards the end of the questionnaire). Furthermore, instructions and definitions of key words for the participants have to be provided to ensure a smooth flow of questions from one topic area to others in the questionnaire. Nevertheless, limit your use of open-ended questions because these can be more taxing on respondents than structured questions and will lead to survey fatigue if overused. Although it is ideal to use appropriate scales for the survey questions, do not ask unnecessary questions that would cause people to freak-out. For example, asking the respondents their date of birth and their ages in your survey questionnaire (Synodinos, 2003; Market research Guy, 2017; Ikart, 2018).

Thirdly, one of the critical steps in effective survey questionnaire design is to have a plan based on the information gathered from the literature review of secondary data, where publications and journals are searched to gather preliminary knowledge of the topic under consideration. Before you start drafting the questionnaire, it might be important you ask yourself the following questions: (i), what business decision am I trying to inform? (ii), if I knew them--? – I would be more prepared to make these important decisions; (iii), what am I trying to measure? Perceptions? Attitudes? Intentions or behaviours of the population sample? (iv), who is my audience? Is my audience familiar with the subject of my research? Is the level of knowledge/understanding in the audience of the questionnaire widely varied or broadly equivalent? (v), what kind of statistics do I want to come out of this project (e.g., descriptive vs. inferential statistics)? And what will my analysis look like? Your answers to each of these questions will help you to craft the survey questions that really matter, and will yield actionable data (Market research Guy, 2017; Ikart, 2018).

Fourthly, it is best to keep your survey questionnaire short, to the point and engaging. Most long surveys are not completed. A quick look at surveys containing too many pages of boring questions produce a response like, “there is no way I’m going to complete this thing”. The response rate for long surveys will drop off dramatically unless the respondents must either be very interested in the topic, an employee, or paid for their time. How long is too long? Although the general rule of thumb is to keep the survey short, the average respondent is able to complete about 3 multiple choice questions per minute. An open-ended text response question counts for about three multiple choice questions, depending on the difficulty of the question. Whilst only a rule of thumb, these formulas will assist you to predict accurately the limits of your survey (Market research Guy, 2017; Ikart, 2018).

Fifthly, as a subject matter expert, you will naturally include acronyms, slang and jargon into your survey without thinking about it. However, it is important to be specific and avoid using big words, jargon, acronyms and/or ambiguous language in the survey. Instead, it is best to use simple sentences and simple choices for the answers with a good and clear layout. You should also evaluate and verify the contents and styles of the questions to ensure the objectives of the study are covered in the questionnaire. Again, simplicity with clarity of your writing style is the best teacher (Synodinos, 2003). Moreover, the way your survey looks and feels can determine how well it performs. To make your survey perform better in the field we recommend you use a visually appealing survey questionnaire with eye catching colours and appropriate font sizes in order to enhance the readability and the look and feel. Adding visual separators between each question, and a Progress bar that shows the respondents how far along they are in the survey, would boost the look and feel of the survey questionnaire (Market research Guy, 2017; Ikart, 2018).

Additionally, pretesting your questionnaire is a key consideration of the survey questionnaire construction process. It is a stage of undisputed importance, without which even the most experienced researchers may come to administer uncertain instruments that will lead to an accumulation of doubts about the research results. Although a more careful examination of the literature on pretesting survey questions have revealed a paradox, pretesting is the only technique to evaluate in advance whether a questionnaire poses problems for the interviewers or respondents. Consequently, experienced specialists declare pretesting indispensable (Babonea & Voicu, 2011). Therefore, we recommend that you find some people including a subset of the population to take your survey without any coaching from you, and then gather their feedback. You can ask them about their general impression. You should also ask them specific questions such as:

1. Was the survey engaging?
2. How long did it take you to complete?
3. Did the question flow logically?
4. Were there any confusing questions?
5. And, were there any areas of frustration?

Designing a perfect questionnaire is impossible. However you can conduct good research by developing an efficient questionnaire based around the above critical information and key considerations of Figure1. In order to design such a questionnaire, we reemphasis you pretest your questionnaire to ascertain its effectiveness. When you test the questionnaire with people, look at the data that comes back, including the comments and feedback. You can also get this data into a data analysis program and start the process of analysing the data. If there are any issues with data structure, improper question type or scale types, they will likely present themselves here. You can fix them before you launch the production survey (Market Research Guy, 2017). We believe this activity helps determine the strengths and weaknesses of the questionnaire. Questionnaire pretesting helps in identifying inappropriate terms in questions wording, inappropriate order, errors in questions, layout, instructions and other problems which might result in respondents' inability to answer certain questions (Synodinos, 2003; Babonea & Voicu, 2011; Ikart, 2018).

To summarise (see Figure 1), the pretesting process aims to evaluate whether:

- Respondents understand all the terms and concepts in the questionnaire;
- Closed questions provide at least one answer choice that would apply to every respondent;
- Questions were interpreted in the same manner by all the respondents;
- Answer choices to be selected correct;
- Every survey question measures what it should measure;
- Questionnaire creates a positive impression, thus motivating people to respond to the question;
- And finally, whether any aspect of the questionnaire suggests any bias from the researcher.

2.2 Classifications of Expert Reviews and Goals of Expert Reviews

In his seminal work, Fricker (2012) classified expert reviews into two main groups. He called the first group*, 'Survey and Questionnaire Experts*' and, the second group, '*Substantive or Subject Matter Experts'.* As asserted by him, whilst the goal of the *Survey and Questionnaire Experts* are to ensure instruments and questions are up to best practices, the *Substantive or Subject Matter Experts* are tasked to making sure that the facts are right and that the questionnaires meet the research objectives (Fricker, 2012). However, DeMaio & Landreth (2003) referred both groups of expert reviews as, '*Survey Methodologists'* who have 10 or more years of experience in questionnaire design, cognitive interviews and survey interview process research with the ability and skills in reviewing survey questionnaire. People who have theoretical questionnaire knowledge or practical experience are asked to review draft questionnaires with an eye to identify questionnaire problems (p 1). Going further, Willis and his colleagues referred expert reviews as, a group of survey design 'experts' who review a questionnaire to identify potential sources of non-sample error by understanding the respondent's task and providing suggestions for ways to minimise potential errors. In other words, experts are individuals who are considered to be experts in the critical appraisal of survey questionnaires (Willis et al. 1999). In practice, however, they are people who can apply their theoretical understanding of, and extensive experience in, survey development in critiquing questionnaires. This technique can also incorporate subject matter 'experts' and interviewers as well (Fricker, 2012; Australia Bureau of Statistics, 2001).

Expert reviewers provide guidance to survey designers in the development of an effective questionnaire. They reveal problem questions, including questions with linguistic and structural issues, so that they can be improved prior to their inclusion into the questionnaire for a field test (Yan et al. 2012; Presser & Blair, 1994). Additionally, expert reviewers sort questionnaire items into groups that are more or less likely to exhibit measurement errors (Willis et al. 1999; Presser & Blair, 1994). Expert reviewers are tasked to ensuring that all questions in the questionnaire are understood in the same way by all respondents, and also that the respondents' understanding of the questionnaire matches what the survey designers intended (DeMaio & Landreth, 2003; Yan et al. 2012). As the subject matter experts, their aim is to make certain that the wordings of the questionnaire are technically correct, appropriate and that the questions are logically presented and response sets reasonable. Characteristically, experts aid in distinguishing questions in survey instruments that are prone to item non-response and/or inaccurate reporting, and as a result, provide appropriate adjustments where applicable (Olson, 2010). Whilst evaluating the questionnaire, experts also make sure that the questionnaires meet: the research objectives, best practices and are easy to administer and respondent and interviewer friendly (Grealish, 2003). When conducting the reviews, experts would systematically analyse the response task for each question in terms of, comprehension, information retrieval, judgement and response generation (Australian Bureau of Statistics, 2001). Any questions they identified as potentially posing difficult retrieval problems, or burdensome, may receive special attention for amendment before their inclusion in production survey (Olson, 2010).

In light of the above considerations, expert reviewers can identify a broad range of errors, including problems with the questionnaire layout, question wording, respondent burden and interviewer considerations. Moreover, they can provide information about interviewer, respondent and mode effects and limited information about interaction effects (Australian Bureau of Statistics, 2001). Presser & Blair (1994) found that expert reviews identified the largest and most consistent number of problems. Furthermore, Willis et al. (1999) conducted a study to compare the number of problems identified by cognitive interviewing, behaviour coding and expert review, evaluating the consistency of the pretesting methods, both externally, (across different techniques) and internally (across different researchers and research organisations) and assessing the types of problems identified. They found that overall the different pretesting techniques including expert review appeared to exhibit a "reasonable degree of consistency". Expert reviews can provide solutions and recommendations for minimising identified sources of errors in questionnaires (Australian Bureau of Statistics, 2001). Table 1, grouped research studies on questionnaire pretesting and expert review into five groups and summarised the related literature. A discussion of these studies follows the table.

Table 1. Classifications of research studies on questionnaire pretesting & expert reviews

Research Areas:	**References**
i. Survey design best practices: How to write a good questionnaire	Ikart, 2018; Market Research Guy, 2017; Babonea & Voice, 2011; Synodinos, 2003
ii. Improving survey quality through pretesting	Ikart, 2018; Haeger et al. 2012; Willis, 2005; Hughes, 2004; DeMaio, et al.1998; ABS, 2001
iii. Methods for pretesting & evaluating survey questions	Presser et al, 2004; Presser & Blair, 1994; Grealish, 2003; ABS, 2001; Rothgeb et al. 2001
iv. Comparing pretesting methods	Ikart, 2018; Yan et al. 2012; Presser at al. 2004; Hughes, 2004; Willis, et al. 1999; Demaio et al. 1998; DeMaio & Rothgeb, 1996; Presser & Blair, 1994; ABS, 2001; Rothgeb et al. 2001
v. Examining questionnaire pretesting by expert reviews	Olson, 2010; DeMaio, & Landreth, 2003; Fricker, 2012; Grealish, 2003; Rothgeb et al. 2001; Willis et al. 1999

In summary:

i. Survey Design Best Practices: How to write a good Questionnaire - research studies (e.g., Ikart, 2018; Market research Guy, 2017; Babonea & Voicu, 2011; Synodinos, 2003) in this group focus on survey design best practices and how to write a good questionnaire. The Market research Guy and Ikart underscored seven components e.g., clarity, relevance, objectivity, look and feel, flow, question structure and survey testing that should be considered when designing survey questionnaires. Similarly, Babonea & Voicu (2011) Synodinos (2003) focus on the art of questionnaire construction and pretesting e.g., establishing the research aims and objectives, data collection methods, questionnaire design, pretesting and reviewing the questionnaire for production surveys. Findings from the group of studies suggest that these considerations are critical in developing a high quality survey questionnaire.

ii. Improving survey quality through pretesting - Research studies (e.g., Haeger, 2012; Willis, 2005; DeMaio et al. 1998) focus on improving questionnaire quality through cognitive interview pretesting. Whilst Ikart (2018) utilised cognitive interviewing (entailing administering a draft survey questionnaire while collecting additional verbal information about the survey responses, which is used to evaluate the quality of the response, or to help determine whether the question is generating the sort of information that its author intends) and respondent debriefing, (involving incorporating follow-up questions in a standardized interview) in improving questionnaire development, Hughes (2004) employed multiple methods including cognitive interviews, respondent debriefing and behaviour coding (which is a system of coding the interactions between an interviewer and a respondent) to improve the quality of the questionnaire through pretesting. Likewise Hughes, DeMaio et al.(1998) & ABS (2001) utilised multiple methods including cognitive interviews, expert reviews, respondent debriefing and behaviour coding in their field test to improve the quality of questionnaire. Findings suggest that cognitive interviews behaviour coding and respondent debriefing and expert reviews are critical pretesting methods for improving survey questionnaires.

iii. Methods for pretesting & evaluating survey questions - research studies (e.g., Presser et al. 2004; DeMaio & Rothgeb, 1996; Presser & Blair, 1994) in this group discussed and compared various methods of pretesting surveys. This included cognitive interviews, behaviour coding, formal respondent debriefings and vignettes to determine whether different methods actually produce different results. Presser & Blair (1994) compared the results of cognitive interview, expert review and behaviour coding along with the conventional pretest which involved telephone interviews conducted by four interviewers to determine whether they varied in terms of reliability, validity and cost. Findings revealed that although the results varied, each method produced relevant results towards improving the quality of questionnaire. Presser et al. (2004) also discussed other pretesting methods e.g., computer assisted telephone interview, computer-assisted personal interviews and computer assisted self-interviews. These have expanded the researchers' ability to measure a range of phenomena more effectively and with improved data quality.

iv. Comparing pretesting methods - Various techniques have been developed over the years to pretest new survey questions or to evaluate the effectiveness of pre-existing questions. The research studies (e.g., Hughes, 2004; DeMaio et al. 1998 & Ikart, 2018) in this group utilised multiple pretesting techniques to pretest questionnaires. This included: (a) the coding of interviewer and respondent behaviour, (b)debriefing respondents to obtain additional information about their views of the questions/concepts, and (c)debriefing interviewers on how respondents react to and understand the question/concepts, and (d) comparing item non response rates and response distributions for different versions of the same survey questions. For example, as part of an effort to collect quality data for the study of decision support systems by Knowledge Workers, Ikart (2018) employed cognitive interviewing and respondent debriefing methods to evaluate questionnaire problems and the quality of interviewer – respondent interactions.

v. Examining questionnaire pretesting by expert reviews - research studies (e.g., Fricker, 2012; Olson, 2010; DeMaio & Landreth, 2003) focus on improving draft questionnaires via expert reviews. Fricker (2012) underscored the importance of developing clear, robust questions through careful review by people (e.g., survey and questionnaire design experts and substantive (subject matter) experts) with alternative viewpoints because good survey data starts with good questions. DeMaio & Landreth (2003) & Olson (2010) reemphasised the importance of having individuals with theoretical questionnaire knowledge or practical experience to evaluate a draft questionnaire either alone or in a group also known as "expert Panel" for consistency and agreement towards improving the quality of questionnaire. To conclude, research studies (e.g., Willis, 1999; Rothgeb et al, 2001) in this group suggest that expert reviews exhibit a reasonable degree of consistency. They contended that expert reviews provide insights into the nature of problems encountered, and in some situations, recommendations for lessening sources of errors in questions.

3. Research Objectives and Research Question

Whereas the literature on questionnaire pretesting has revealed a paradox, questionnaire pretesting is a simple technique to measure in advance whether a questionnaire causes problems for respondents or interviewers (Babonea & Voicu, 2011; Presser et al. 2004). Consequently, experienced researchers and survey methodologists have declared questionnaire pretesting indispensable. All the same, published survey reports regularly provide no information about whether a questionnaire was pretested and, if so, how and with what results. Moreover, until recently, there has been fewer methodological research on questionnaire pretesting. The universally acknowledged importance of questionnaire pretesting has been honoured more in theory than in the practice. As a result, we know very little about an aspect of pretesting and the extent to which a pretest serves its intended purpose and leads to value-added on questionnaires (Babonea & Voicu, 2011; Ikart, 2018).

Throughout the past three decades there has been increased emphasis on building quality into the survey questionnaire through pretesting. Whilst this has been approached from an operational perspective (e.g., Dippo & Norwood, 1992), it was informed by theoretical work in the fields of cognitive psychology and social psychology (e.g.,Turner & Martin, 1984; Ericsson & Simon, 1984). Whereas prior to this time, the main contributors to diagnosing questionnaire problems were; the questionnaire designers (through their expertise in the subject) and the interviewers (through their experience in administering the questionnaire). In recent periods, the emphasis has shifted to learning about questionnaire problems from the respondents themselves (DeMaio et al. 1998). This has been grounded on the development of a model of survey response (Haeger et al. 2012; DeMaio & Lanreth, 2003; Ikart, 2018) that divides the response process into four major groups: comprehension, retrieval, judgement and response formulation – which occur within the respondent. The understanding of which allows researchers to get a grasp on issues that impact the quality of the data collected in the survey.

Theoretically, there are four actions that the respondent enacts when answering a survey question. They must comprehend the question, be able to retrieve information, make a judgement as to its relevance and accuracy as an answer to the question, and respond to the question (Hughes, 2004; Willis, 2005; Haeger et al. 2012). Nevertheless, detailed reports of an appropriate method to undertake questionnaire pretesting have been underrepresented within the literature (Hilton, 2015),

An expert review is one of the traditional methods of questionnaire pretesting other than cognitive interviewing, respondent behaviour and behaviour coding. Expert reviews can be conducted with varying levels of organisation and rigor. On the lower end of the spectrum, an experienced subject matter expert, or survey methodologist, review a draft questionnaire to identify issues with question wording or administration that may lead to measurement error. On the more rigorous end of the spectrum, as we utilised in this study, is the Questionnaire Appraisal Scheme (QAS) method, a coding scheme developed by Lesser & Forsyth (1996) and adopted by prior studies (e.g., Rothgeb et al. 2001; DeMaio & Landreth, 2003). The QAS is a structured, standardized instrument review containing 28 problem types that allow experienced researchers and coders to review and code questionnaire problems that have been reported by the independent expert reviewers based on the 28 problem types of the QAS. Using a database system, the coders coded and processed the questionnaire problems by applying the 28 problem types of the QAS. They then conducted an analysis and compared the results for consistency and agreement and improvement of the questionnaire for the production survey.

Regardless of the widespread use of the expert reviews and the QAS method in researches (e.g. Presser and Blair, 1994; Olson, 2010; Yan et al. 2012), fewer empirical evaluations exist of the expert review itself. In particular, there is little evidence as to whether the results produced by individual expert reviewers are consistent. Furthermore, there is a limited reasonable level of agreement among expert reviewers in their evaluations of questionnaire (Rothgeb et al. 2001; DeMaio & Landreth, 2003; Olson, 2010). Building on previous studies (e.g., Presser & Blair, 2004; DeMaio & Landreth, 2003; Olson, 2010), this study aims to answer two specific research questions:

1. Clarify whether the results produced by individual expert reviewers are consistent.
2. If any, to what extend is the level of the agreement among experts in their evaluations of the questionnaire?

4. Research Methodology

Three survey methodologists and one management expert consisting of: two academics, from two different academic institutions, one Senior Education Officer from the Government Agency and, a CEO from the Financial Management Industry in Australia, were enlisted to conduct expert reviews of the draft questionnaire. Each reviewer had more than ten years of experience in questionnaire design, research in survey interview, cognitive interviews and other methods of pretesting. They were selected based on their ability to review draft questionnaires and willingness to complete the reviews in a timely manner (Rothgeb et al. 2001; DeMaio & Landreth, 2003; Olson, 2010).

The draft questionnaire for the review consisted of 17 questions on vocational education and training programs and prisoners' work readiness outside gaol, at post release employment. A subset of questions was based on prisoners' experience in vocational education and training program, whilst in custody. The second subset was based on the perceived usefulness of vocational and training programs. The third subset of the questionnaire was about the prisoners' satisfaction with vocational education and training. The final subset of the questionnaire, the facilitating conditions and social environments of correctional education centre was extracted from the questionnaire developed for the study. We aimed at selecting important topics of the questions which could be administered face-to-face to the population sample of prisoners. Also, the question topics contained limited skip patterns, which helped maximize the number of sample cases receiving each question (DeMaio & Landreth, 2003).

Adopted from prior study, e.g., DeMaio & Landreth, each expert reviewer was provided with independent evaluator record sheets consisting of three parts – namely; Part1, Part 11 & Part 111. In Part 1: *Question by Question Problem Identification*. For each survey question, the expert reviewer was asked to identify and briefly explain each specific problem associated with the question in a 17- question survey on vocational education and training, treatment programs and prisoners' work readiness outside gaol. They were also asked to record each problem separately on a pre-numbered form according to the numbers of questions in the questionnaire provided. Further, they were asked to classify each problem identified either as; 'High Priority' (a problem that should be addressed before the instrument is fielded because it will likely adversely affect the response process in unacceptable ways) or 'Low Priority'(a problem that could be addressed before instrument is fielded, but may not adversely affect the response process in unacceptable ways,) and to record whether the problem will be a problem with administering the question, response problem or both. In Part 11*: Five (5) Most Important Problems*, each expert reviewer was asked to briefly state the five (5) most important problems they found with the questionnaire. And for each of the problems identified, they were asked to list the question numbers that were likely to be affected. Finally, in Part 111: *Five (5) Worst Questions*, each expert reviewer was asked to identify the five (5) worst questions. Also, for each question identified, they were asked to provide their comments or a short explanation for their selection. No other specific instructions were provided to the expert reviewers, except, a short description of the goals of the questionnaire.

5. Analysis and Results

The expert reviewers reported their review on the paper forms provided by the researcher (see *ATTACHMENT A1-3* for a sample of the report forms). We (including the researcher and one other experienced coder) coded the completed forms with a significant level of inter-coder agreement (76.95%) by reviewing the open-ended evaluator notes concerning the problems that had been recorded in the independent evaluator sheets by the four expert reviewers and by applying the QAS containing 28 problem types (see *ATTACHMENT B*). Each item received as many codes as we agreed were found to apply to the item based on expert reviewers' written comments of the questionnaire problems. We recorded the problem types and their locations in the database and compared this across the four expert reviewers. We also grouped the 28 problem types of the QAS at the highest level under the familiar headings of the four stage cognitive response model: a) comprehension and communication, b) retrieval, c) judgement and evaluation, and d) response selection. In addition, we considered the mid-level and lowest level categories of the problem types, though we collapsed them into the major groups of the highest level at some point in the analysis. Furthermore, we generated agreement statistics by dividing the total occurrences of cases (problem type) where the researcher and the other coder agreed on problem types and locations (that is, question number) by the total number of mutually exclusive problem types across all expert reviewers (273).

Table 2 presents the results of problem types identified by expert reviewers, and the extent of agreement among expert reviewers in identifying the types of problems and number of problems in the questionnaire. As can be seen from the top row, out of a total of 76 *Interview Difficulties problem type* (which was 20% of all problem types) identified across the four expert reviewers; expert A identified the majority, 30.26%, followed by expert C, 26.32%. Next, was expert D, 25% though expert B identified the least, 18.42%. Further, of a total 198 *Comprehension problem type* (53.23% of all problem types) identified across the four expert reviewers, expert A identified the highest, 41.41%; followed by expert B, 23.23% and expert D, 21.21%. Expert C identified the least, 14.14%. For the *Retrieval problem type* (which was 6.18% of all problem types), whilst expert A and expert C identified 30.43% respectively, expert D identified 21.74% and expert B identified just 17.39%. Regarding the *Judgement problem type* which was 6.1% of all problem types, expert A identified 30.43%, followed by expert C

26.1%. Both Expert B and expert D identified equally, 21.74%. Finally, for the *Response problem type* which was 13.98% of all problem types, expert C identified the highest, 30.77%; followed by expert A, 26.92%; then expert D 23.1% and finally, expert B 19.23%. Table 2 clearly illustrates the degree to which expert reviewers agreed among themselves to identify the number and types of problems in the questionnaire. Also, as can be seen from the table, there are enormous differences among expert reviewers regarding the number of problems each of them identified. For example out of a total 133 problems identified by expert A, expert B and expert C identified almost half, 59.4% and 59.94% respectively though expert D identified 62.41% of the amount.

Table 2. Problem types identified by expert reviewers

	Expert A	**Expert B**	**Expert C**	**Expert D**	**Total**	**Accumulative frequency %**
Problems Types						
interview difficulties	23 (30.26 %)	14 (18.42%)	20 (26.32%)	19 (25%)	76	20.43%
Comprehension	82 (41.41%)	46 (23.23%)	28 (14.14%)	42 (21.21%)	198	53.23%
Retrieval	7 (30.43%)	4 (17.39%)	7 (30.43%)	5 (21.74%)	23	6.18%
Judgement	7 (30.43%)	5 (21.74%)	6 (26.1%	5 (21.74%)	23	6.18%
Response	14 (26.92%)	10 (19.23%)	16 (30.77%)	12 (23.1%)	52	13.98%
Total	**133**	**79**	**77**	**83**	372	100

Further, we evaluated the level of agreement across expert reviewers who identified specific problems in a particular question. The result was as low as 20.01%. Nevertheless, as the data in Table 2 suggests, it was reasonable to say that expert reviewers found similar types of problems (*vague term/unclear question such as, how many years have you been sentenced for*? Or, *how many year have you personally been participating in vocational education and training programs in your incarceration*) though chose to document them at varying points in the questionnaire evaluation form. Also, when we collapsed the percentage of problem types of the lowest level and middle level into the highest level of *interview difficulties, comprehension, retrieval, judgement and response*; the percentage of problem types across the four expert reviewers were comparable. As can be seen from Table 2, the *comprehensio*n category ranks highest with regard to the percentage of the problem types identified by each expert reviewer; between 41.41% and 21.21% with, accumulative frequency of 53.23%. This was followed by the *interview difficulties* category problem types which were 30.26% and 18.42% with accumulative frequency of 20%. The *response* category ranked third highest, between 30.77% and 19.23% respectively with accumulative frequency of 13.98%. The *retrieval and judgement* categories ranked fourth highest, 30% and 21% with accumulative frequency of 6.18%, though one expert reviewer had 17.39% for retrieval problem types.

The majority of the problem types identified by expert reviewers were from the comprehension, interview difficulties and response categories. There was also consistent agreement across the expert reviewers for *retrieval* and *judgement*, which were the lowest ranked problem types.

Within each of the four stage problem types were the mid-level, and the lowest level categories, the most detail description of the problems e.g., *complex or awkward syntax and erroneous assumption*. It was critical the QAS codes be independent of one another and that rules applied on the use of any codes which may be ambiguous because the QAS was designed in order to attempt to maximize inter-coder agreement with respect to the assignment of individual codes. Table 3 presents the frequency results with each of the 28 QAS codes assigned, overall to the questionnaire items. The problem types were coded on the basis of the 28 QAS coding system in order to maintain consistency. Collectively, the expert reviews identified a total of 372 problems across the four expert reviewers for an average of 2.75% codes per question.

Table 3. Frequency results & QAS codes assigned to questionnaire problematic items

Code problem label	**Expert Reviewers**				**Frequency**	**Percent**
Problem types:	**A**	**B**	**C**	**D**		
Interviewer Difficulties						
1. Inaccurate instruction	15	5	6	7	33	8.87
2. Complicated instruction	6	9	14	10	39	10.48
3. Difficult for interviewer to administer	2	0	0	2	4	1.08
Question Content						
4. Vague/unclear question	20	14	4	8	46	12.37
5. Complex topic	3	4	4	4	15	4.03
6. Topic carried over from earlier Q	4	4	1	2	11	2.96
7. Undefined/vague term	18	8	8	8	42	11.29
Question structure						
8. Transition needed	2	2	3	4	11	2.96
9. Unclear respond instruction	5	2	4	6	17	4.57
10. Question too long	0	0	0	0	0	0,0
11. Complex/awkward syntax	20	10	4	6	40	10.75
12. Erroneous assumption	2	0	0	0	2	0.54
13. Several questions	0	0	0	0	0	0.0
Referenced period						
14. Period carried over from earlier Q	2	0	0	0	2	0.54
15. Undefined period	4	0	0	2	6	1.61
16. Unanchoring/rolling period	4	0	0	2	6	1.61
Retrieval from Memory						
17. Shortage of memory cues	1	2	3	3	9	2.42
18. High detail require or info unavailable	6	2	4	2	14	3.76
19. Long recall or reference period	0	0	0	0	0	0.0
Judgement & Evaluation						
20. Complex estimation	6	4	3	5	18	4.84
21. Potentially sensitive/bias	1	1	3	0	5	1.34
Respond terminology						
22. Undefined term	3	2	4	3	12	3.23
23. Vague terms	3	4	4	4	15	4.03
Response Units						
24. Responses use wrong or mismatching units	2	0	6	3	11	2.96
25. Unclear to respondent what the response options are	4	2	0	2	8	2.15
26. Multi-dimensional response set	0	0	0	0	0	0.0
Response structure						
27. Overlapping response categories	2	2	0	0	4	1.08
28. Missing response categories	0	0	2	0	2	0.54
Grand Total					372	100.0

As can be seen from the Table 3, a small number of codes accounted for a high proportion of problems identified by the expert reviewers. For example the five codes namely, *inaccurate instruction* (8.87%), *complicated instruction* (10.48%), *vague/unclear question* (12.37%), *undefined/vague term* (11.29%) *and complex/awkward syntax* (10.75%) accounted for 53.76% of all identified problems. Note that all the codes were classed on the

QAS system as interview difficulties, comprehensive, retrieval, judgement and response problems. Also, two codes namely: *vague/unclear question* (12.37%) and *undefined/vague terms* (11.29%) account for approximately 23.66%. This result is consistent with the previous findings (e.g., Rothgeb et al. 2001; Willis et al. 1999), who found that, vague/unclear questions and undefined/vague terms dominated the results of the coding system.

Furthermore, we looked at the results from another technique concentrating on the number of questions that expert reviewers identified as problematic questions. The second row of Table 4, showed the number of questions with problems across the expert reviewers. As illustrated, there were some similarities between Expert C and expert D except for expert A and expert B which were 73.33% and 53.33% respectively. Hence, the lowest number of problem questions was 53.33% of the highest number (8/15). In contrast the lowest number of problems found from row one of Table 4, was 57.89% of the highest number (77/133) for expert reviewer C and expert reviewer A.

To summarise, there were many inconsistencies when we compared the number of problems found across the expert reviewers, though these inconsistencies were fewer when we compared the number of questions expert reviewers identified as problematic questions. These results were consistent with prior research findings (e.g., DeMaio & Landreth, 2003).

Based on the experiences in questionnaire evaluations, each expert reviewer was asked to briefly state the five most important problems they found with the questionnaire. For each of the problems they identified, they were also asked to list the question numbers that are likely to be affected by the problem. The results of the most important problems of the questionnaire as identified by the expert reviewers varied significantly. For instance, one asserted that ‘*the use of modifier in sentences was unnecessary and that it caused redundancy to sentences*’. Another expert reviewer claimed that ‘*some terminologies may not be understood by the respondents’*. One other expert reviewer commented that ‘*the questions are leading though lack objectivity*’. Another expert reviewer stated that ‘*terms use in survey are imprecise or open to subjective interpretation*’ With regards to agreement on a particular problem, there was no agreement by the four expert reviewers on a single issue.

Table 4. Number of questionnaire problems & flawed questions found

Experts **Problems & flawed Questions:**	**A**	**B**	**C**	**D**
Number of problems found	133	97	77	83
Number of questions with problems	11	8	15	14
Number of Questions affected by major problems	13	6	9	15
Number of worst questions	5	5	5	5

As mentioned previously, the questionnaire problems that expert reviewers identified as problematic greatly varied, except for three expert reviewers who identified a comparable problem. Nonetheless, one major problem that was specifically identified by an expert reviewer was stretched into three problems by one other expert reviewer, and also into two problems by another expert reviewer. Ostensibly, expert reviewers inconsistently agreed on major problems of the questionnaire far less than the expectation of any research community.

Additionally, there were many differences in the numbers of questions affected by major problems. Although two or three expert reviewers identified a similar problem, the affected questions they reported were significantly different in numbers and locations. As shown in Table 4, the third row, expert reviewer D reported the highest number (15) of affected questions by major problems, followed by expert reviewer A, thirteen (13). But expert reviewers C and B reported fewer affected questions, which were nine (9) and six (6) affected questions respectively.

Furthermore, each expert reviewer was asked to identify the five worst questions of the questionnaire. For each question they identified, they were also asked to provide a comment or a short explanation for its selection. The results of their worst questions and comments significantly varied. As shown in the fourth row of Table 4, although the four expert reviewers reported five questions each as the worst questions, the question numbers and comments were mixed. Only two questions (Q5 and Q6), were identified by three expert reviewers as the worst questions, though their comments regarding the question numbers were somewhat different.

Moreover, two other expert reviewers identified one other question as the worst question but their comments regarding this question differed significantly. Additionally, one other question was comparable in terms of worst question between two expert reviewers though with dissimilar comments. All other worst question numbers identified by the expert reviewers were not comparable. That is, expert reviewers did not agree with one another in terms of the question numbers they identified as their worst questions. About eight other questions that expert reviewers identified between themselves were not comparable in terms of question numbers and explanations provided.

To summarise, there was little agreement among expert reviewers. Although two or three expert reviewers identified same questions as worst questions, their explanations and comments regarding these questions were mixed.

6. Discussion and Conclusion

This paper employed expert reviews to pre-test the draft questionnaire developed for the study of the impact of vocational education and training programs at post release employment and prisoners' work readiness outside gaol. The method employed in the present study has both qualitative and quantitative approaches. An expert review is frequently used earlier in research in refining the survey questionnaire before it is administered in the field to the respondents (Ikart, 2018; DeMaio & Landreth, 2003).

The overarching objectives of this paper were; first, we wanted to evaluate whether the results produced by individual expert reviewers are consistent. Second, if any of those results are in fact consistent, to what extend is the level of agreement among the expert reviewers in their evaluations of the questionnaire.

The results of this study were mixed. First, there were vast differences among the expert reviewers vis-à-vis the number of problems found across expert reviewers, though these differences were not too distinct across the expert reviewers (except for expert A, 73.33% and expert B, 53.33%) when we looked at the results from another technique, focusing on the number of questions that the expert reviewers found to be problematic. Second, the level of agreement across expert reviewers in identifying a specific problem in a particular question was as low as 20.01%. Noticeably, the expert reviewers found similar problems though, chose to document them at various points in the evaluations. Nevertheless, when we ranked the problem types of those problems, we observed that the majority of problem types were from *comprehension* and was followed *by interview difficulties* problem type. The *response category* ranked third highest. This agreement was consistent across the expert reviewers for *retrieval* and *judgement* problem types respectively. These findings were supported by previous studies (e.g., DeMaio & Landreth, 2003; Rothgeb et al. 2001; Willis et al. 1999).

Third, when we factored-in the most important problems of the questionnaire together with the affected questions that each expert reviewer was asked to identify and list, the results also varied across the expert reviewers. The results illustrated that the expert reviewers inconsistently agreed on major problems much lesser than the expectation (DeMaio & Landreth, 2003).

Finally, when we compared the five worst questions selected by each expert reviewer, as well as their comments and explanations for the selection across the expert reviewers, we also found that expert reviewers inconsistently agreed substantially smaller than the expectation (DeMaio & Landreth, 2003).

The results of the study suggest that different review styles from the evaluators were at play (Robbins et al. 2003; DeMaio & Landreth, 2003). We felt that Expert A for instance, utilised a country club review style, because s/he presented a comprehensive and detailed review of the questionnaire to help improve its quality to the needs of the respondents. Expert A also found almost twice the problems found by expert C. Moreover, Expert A found 27% more of the problems found by Expert B and, 38% more of the problems found by Expert D. Further, when we compared the five most important problems and the affected questions across the expert reviewers, expert A presented thorough explanations and comments of the problems. The five major problems that expert A identified affected almost fourteen questions out of the seventeen questions of the questionnaire. This was approximately 82.35% of the questions. When we compared the comments and explanations of the five worst questions identified by each expert reviewer, it was clear that expert A provided extensive and exhaustive comments and explanations. This was enough to conclude that a focussed review work such as the work of expert A was necessary in order to get a detailed understanding of the importance of pretesting the questionnaire and quality improvement.

Expert B on the other hand, can be thought of applying impoverish review style and exerting minimum effort in evaluating the questionnaire (Robbins et al. 2003). When we compared the number of questions with problems across the expert reviewers, expert B identified relative few questions, approximately 53.33% of the highest number (8/15) of questions with problems of expert C. Also, when we compared the number of problems found across the expert reviewers, expert B identified relatively fewer problems, approximately 73% of the highest of expert A. However, this was comparable to expert C and expert D. But in terms of the number of questions affected by the major problem, expert B identified the lowest, about 40% of highest number (6/15) of expert D. Further, when we looked at the five worst questions identified by expert B, we noticed that these questions have been repeated all through, though were sensible. On the basis of comparing the information provided between expert A and expert B, it was difficult to say that both experts reviewed the same questionnaire.

Regarding expert C and expert D, we thought that both expert reviewers utilised middle –of-road and/or balanced review style (Robbins et al. 2003). For instance when we compared the number of problems found across the expert reviewers, we noticed that those found by expert C and expert D were in the middle of the problems found by expert A, though were similar to the number of problems found by expert B. Also, when we looked at the number of questions with problems across the expert reviewers, the number of questions with problems found by both experts was similar. Additionally, when we compared the number of questions affected by major problems as identified between these two experts, although they were dissimilar because Expert C identified 60% of the highest (9/15) of expert D, two of the problems identified by expert D were unnecessarily stretched to affect some questions. However, the strengths of their explanations regarding the most important problems resulting in the affected questions were comparable and adequately explained and enumerated for the improvement of the questionnaire.

The question is why are the results inconsistent? Few explanations can be tapped into to justify the inconsistencies of the results. First, may be the expert reviewers in this study usually work in an expert panel as a team of questionnaire evaluators. Therefore working individually to review the study questionnaire may have been an uphill task to some of them, particularly expert reviewer B. Second, the amount of time and effort that the expert reviewers have dedicated for the evaluation may have had an influence on the outcome. In the case of this study, we employed four expert reviewers. Let’s say, on one hand, one or two expert reviewers devoted limited time and effort in the review. The results would be minimal and substandard. On the other hand, if one or two other expert reviewers put in enough time and effort in providing a comprehensive and detailed review of the questionnaire, the results are of high quality and standard. It is therefore very clear that the results between the two groups of expert reviewers would be unequalled, even though the task itself was standardized across the expert reviewers and designed to minimise variation in problem identification. Perhaps, this was the case about the present study. Third is the diverse level of details of the comments and explanations across the expert reviewers. For example when we compared expert reviewer A and expert reviewer B works of this study, expert reviewer A spent time to highlight every small piece of problems and also comment on those problems, even though s/he had previously highlighted similar problems. Expert reviewer B on the other hand, rarely highlighted and/or commented on repeated problems of the questionnaire. Hence, the differing level of details and perspectives by the evaluators may have contributed to the degree of inconsistencies and disagreements among expert reviewers. The variation across expert reviewers may have also been intensified because evaluators with more dissimilar backgrounds were used. Fourth, the perspective of one expert reviewer may not always match those of other expert reviewers, particularly where expert reviewers employed their informal judgement and/or instinct in the review. In other words, it may be difficult for an expert review to put him/her-self in the shoes of other expert reviewers in terms of, views and opinions of each problem of the questionnaire and anticipate all other expert reviewers would come up with similar views and opinions. Fifth, in practice, however, expert reviewers are survey methodologists who can apply their theoretical understanding of, and extensive experience in, survey development in critiquing questionnaire (Australian Bureau of Statistics, 2001). We felt that in this present study this claim may have left them better off or worse off.

Notwithstanding the inconsistency and differences across the expert reviewers, the expert ratings successfully identified questionnaire problems that were more likely to have high levels of item nonresponse or inaccurate reporting, although success in this review differs across the expert reviewers. Although expert reviewers are survey methodologists with ten or more years of experience in questionnaire design, cognitive interviews and survey interview process research with ability and skills in reviewing questionnaire (Olson, 2010; DeMaio & Landreth, 2003; ABS, 2001), the high degree of inconsistencies and low level of agreement of the results of the present study are enough to cause concern concerning expert reviews itself and/or to extrapolate the results.

The second question that is of concern is; how is the expert review organised? And what procedures are in place to guide the evaluators? In this study, four expert reviewers from well-established academic institutions, financial industry and a Government agency were enlisted to review the questionnaire based on their many years of experience in reviewing questionnaires for quality and improvement. Furthermore, the review process was structured. For example, each expert reviewer was provided with independent evaluation record sheets and specific instructions to follow, based on Questionnaire Appraisal Scheme QAS, coding system as a control mechanism for the review. However, it was difficult to assess whether each expert reviewer utilised the procedures that were in place or relied exclusively on their own judgements, making informal assessments that typically yield open ended comments of the survey items evaluated. Our contention is that an adequate control mechanism should be established and monitored in expert reviews against the established standards and procedures if expert review is to serve as a critical traditional method of questionnaire pretesting. We suggest computer aided application expert review can serve as a check and balance on individual evaluators review work for consistency and agreement among reviewers. Future research may focus on this suggestion. Of vital importance, we are of the view that collaborative questionnaire reviews by two or more expert reviewers would make a significant contribution for consistency and agreement among expert reviewers and should be embraced (Presser & Blair, 1994; ABS, 2001). Otherwise, it should be noted that individual experts used in researches are not from a uniform group, and the variation in the review outcomes should be anticipated when pretesting and/or review questionnaires (Olson, 2010).

One final thought, though the QAS coding scheme has been depicted as a significant tool for identifying the questionnaire problems, given the extremely high rate of detection, it is very possible that such an appraisal technique as applied, may encourage a low threshold for problem identification and in so doing producing a large number of differences in the results across the expert reviewers. The QAS method as used may have high sensitivity but poor specificity. Probably, the QAS is useful for the identification of questionnaire problems, and that the questionnaire used for this study does in fact possess the various problems that it revealed from the problem types. Conversely, it is possible that a system like the QAS that was designed primarily as an aid to questionnaire designer, rather than pretesting, requires a fair degree of additional expert judgement to be considered practically effective for questionnaire pretesting. Nevertheless, the results strongly support the conclusion of previous studies (e.g., Willis et al. 1999; Rothgeb et al. 2001) and that any evaluation designs depending on the notion that “finding more problems is better” is suspect, because of the exclusive focus on technique sensitivity.

It should be noted that this study was an observational study, but not a control experiment. Although the paper is based on a research project in progress, no part of the pretested questionnaire was administered to prisoners for primary data collection. The views expressed in this study are those of the authors though supported by the views from previous publications. However, we published this paper to inform persons in academia and management practice and to encourage roundtable discussions. Finally, we believe that research in expert reviews would continue to evolve and change as research into its validity and practice continues.

References

Australian Bureau of Statistics. (2001). *Pre-Testing in Survey Development: An Australian Bureau of Statistics Perspective, Population Survey Development.* Australian Bureau of Statistics.

Babonea, A., & Voicu, M. (2011). Questionnaires pretesting in marketing research. *Challenges of the Knowledge Society, 1*, 1323-1330.

DeMaio, T., & Landreth, A. (2003, October 21-23). Examining Expert Reviews as a Pretest Method. In P. Prufer, M. Rexroth & F. J. Fowler Jr. (Eds.), *QUEST 2003: Proceedings of the 4th Conference on Questionnaire Evaluation Standards* (pp. 60-66). Nachrichten: ZUMA

DeMaio, T.J., Rothgeb, J., & Hess, J. (1998). *Improving Survey Quality Through Pretesting.* U.S. Bureau of the Census, Washington, DC. Retrieved from https://www.census.gov/srd/papers/pdf/sm98-03.pdf

Dippo, C. S., & Norwood, J. L. (1992). A review of research at the Bureau of Labor Statistics. In J. M. Tanur (Ed.), *Question about Questions.* New York: Russell Sage Foundation.

Fricker, R. (2012, June). Evaluating Survey Questions. *Paper presented at the meeting of Naval Postgraduate School*. Monterey, California.

Grealish, D. (2003, October 21 -23). Pre-testing questionnaires: the New Zealand experience. In P. Prufer, M. Rexroth & F. J. Foeler Jr. (Eds.), *QUEST 2003: Proceedings of the 4th Conference on Questionnaire Evaluation Standards* (pp. 37-42). Nachrichten: ZUMA

Haeger, H., Lambert, A. D., Kinzie, J., & Gieser, J. (2012). Using cognitive interviews to improve survey instruments. *Proceedings of the Annual Forum of the Association for Institutional Research*. New Orleans, Louisiana.

Hilton, C. E. (2015). The importance of pretesting questionnaires: A field research example of cognitive pretesting the exercise referral quality of life scale (ER-QLS). *International Journal of Social Research Methodology*, 1-14. https://doi.org/10.1080/13645579.2015.1091640

Hughes, K. A. (2004). Comparing pretesting methods: Cognitive interviews, respondent debriefing, and behavior coding. *Proceedings of the Annual Meeting of the Federal Committee on Statistical Division Methodology*. Arlington, VA. Retrieved from https//www.census.gov/srd/papers/pdf/rsm2004-02.pdf

Ikart, E. M. (2018). Questionnaire Pretesting Methods: A Comparison of Cognitive Interviewing and Respondent Debriefing Vis-à-vis the Study of the Adoption of Decision Support Systems by Knowledge Workers. *International Journal of Business and Information, 13*(1), 119-154.

Lesser, J., & Forsyth, B. (1996). A Coding Systems for Appraising Questionnaires. In N. Schwarz & S. Sudman (Eds.), *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research* (pp.259 -291). San Francisco: Jossey-Bass.

Market Research Guy. (2017). *Survey Design Best Practices: How To Write a Good Questionnaire.* Retrieved from http://www.mymarketresearchmethods.com/survey-design-best-practices/

Olson, K. (2010). An Examination of Questionnaire Evaluation by Experts. *Field Methods, 22*(4), 295-318. https://doi.org/10.1177/1525822X10379795

Presser, S., & Blair, J. (1994). Survey pretesting: Do different methods produce different results?. In P.V. Marsden (Ed.), *Sociological Methodology, 24,* 73-104. https://doi.org/10.2307/270979

Presser, S., Couper, M. P., Lessler, J. T., Martin, J., Rothgeb, J. M., & Singer, E. V. (2004). Methods for testing and evaluating survey questions. *Public Opinion Quarterly, 68*(1), 109-130. https://doi.org/10.1093/poq/nfh008

Robbins, S. P., Bergman, R., Stagg, I. & Coulter, M. (2003). *Foundation of management.* Prentice Hall.

Rothgeb, J., Willis, G., & Forsyth, B. (2001). Questionnaire Pretesting Methods: Do Different Techniques and Different Organisations Produce Similar Results?. *Annual Conference of American Association for Public Opinion Research Montreal.* National Cancer Institute, Westat Inc.

Synodinos, N. (2003). The 'art' of questionnaire construction: Some important considerations for manufacturing studies. *Integrated Manufacturing Systems, 14*(3), 221-237. https://doi.org/10.1108/09576060310463172

Tourangeau, R. (2005). Survey research and societal change. *Annual Review of Psychology, 55,* 775-801. https://doi.org/10.1146/annurev.psych.55.090902.142040

Turner, C., & Martin, E. (1984). *Survey Subjective Phenomena.* New York: Russell Sage Publishers.

Willis, G. B. (2005). *Cognitive Interviewing. A Tool for Improving Questionnaire Design.* London: Sage Publishing.

Willis, G. B., Schechter, S., & Whitaker, K. (1999). A comparison of cognitive interviewing, expert review and behaviour coding: What do they tell us?. *Proceedings of the Section on Survey Research Methods* (pp. 28-37). Washington, D.C.: American Statistical Association. https://doi.org/10.4135/9781412983655

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

Appendix: Examples on Record Sheets Selected at Random

Attachment A1: PART 1: Question-by-Question Problem Identification

Q2	Priority:	Problem for:
Problem 1:	□ H □ L	□ A □ R □ B

It is not clear whether the question is referring to this sentence or accumulation of all sentences.

Q 12	Priority:	Problem for:
Problem 1:	□ H □ L	□ A □ R □ B

I think respondents will find it more difficult to answer questions that call for self-evaluation or forecast of benefits than those calling for more simple opinion on past events.

Q15 & 16	Priority:	Problem for:
Problem 3:	□ H □ L	□ A □ R □ B

The CMT is common to staff, but offenders call it “Classo” writing it out in full or using the slang term will help.

Q4	Priority:	Problem for:
Problem 1:	□ H □ L	□ A □ R □ B

This is a self-serving question. I doubt respondents will provide an honest answer to the question. All will answer ‘very likely’. They may also be fearful of the consequence of providing any answer which is anything less than 5. To sum up, respondents will not believe in the confidence nature of the survey. Unless, the expected outcome is measured indirectly, the question will most certainly induce bias.

Q3	Priority:	Problem for:
Problem 1:	□ H □ L	□ A □ R □ B

Good question but I think the responders need to be able to determine whether they are compared to what they can do… can they read newspapers or safety signs?

Attachment A2: PART II: Five (5) Most Important Problems

Briefly state the problem and list affected question numbers:

Problem 1:

Most questions are vague, unclear and/or too broad. Honestly, I doubt respondents will answer these questions truthfully. In my opinion, unless the expected outcomes are clear and/or measured indirectly, these questions will certainly induce biases.

A*ffected Numbers*: Q1, Q2 & Q4

Problem 2:

Use of English could be plainer

Affected question Numbers: Q8, Q12, Q13 & Q14

Problem 3

You use terms that are imprecise or open to subjective interpretation. I have identified for each question. *Affected questions*: Q9 & Q12

Problem 4

Some questions are almost similar. I doubt respondents will give their honest answers.

Affected questions: Q4, 5 & 6

Problem 5

Most questions and tables are poorly formatted. This is critical for professionalism of your work.

Affected questions: Q1, Q2, Q4, Q5, Q6, Q6, Q8, Q9, Q10, Q12, Q13, Q14 & Q15

Attachment A3: PART III: Five (5) Worst Questions

Identify question number and provide comments or a short explanation for why it was selected:

Problem 1: Question #1: It is not clear whether you are referring to multiple sentences or this sentence only. You need to be specific (Q1&Q2).

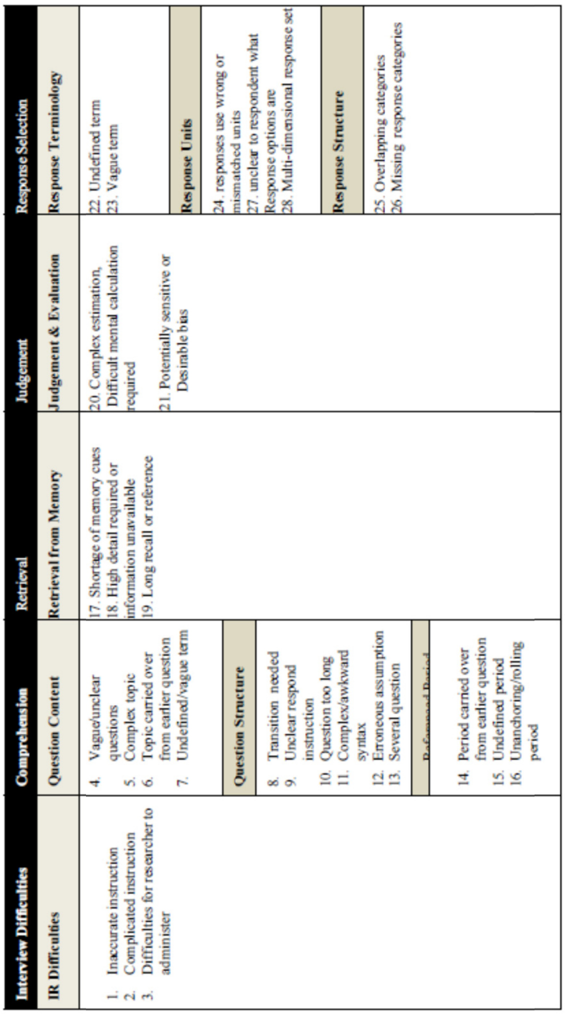
Problem 2: Questions #Q5 & #Q6: These questions compound ‘improvement in skills’ with ‘absolute level of skills attainment’. If you want to know about the resulting level of capacity then this would be better asked in a separate question. Q6 almost duplicates this. I would ask about what key competencies the respondent thinks will be useful for employment.

Problem 3: Questions #Q15, & #16: The CMT is common to staff, but offenders call it “Classo” writing it out in full or using the slang term will help.

Problem 4: Question #Q12. "Participation in treatment programs directly related to my offence to address offending behaviour"… this is overly complicated for offender, just say the programs they have to do is probably clearer

Problem 5: Question #Q17. Asking about the social relationships is likely to confuse many offenders. Better to just ask who encourage them to attend.

Attachment B: Questionnaire Appraisal Coding Scheme, QAS



Interview Difficulties	Comprehension	Retrieval	Judgement	Response Selection
IR Difficulties	**Question Content**	**Retrieval from Memory**	**Judgement & Evaluation**	**Response Terminology**
1. Inaccurate instruction 2. Complicated instruction 3. Difficulties for researcher to administer	4. Vague/unclear questions 5. Complex topic 6. Topic carried over from earlier question 7. Undefined/vague term	17. Shortage of memory cues 18. High detail required or information unavailable 19. Long recall or reference	20. Complex estimation, Difficult mental calculation required 21. Potentially sensitive or Desirable bias	22. Undefined term 23. Vague term
	Question Structure			**Response Units**
	8. Transition needed 9. Unclear respond instruction 10. Question too long 11. Complex/awkward syntax 12. Erroneous assumption 13. Several question			24. responses use wrong or mismatched units 27. unclear to respondent what Response options are 28. Multi-dimensional response set
	Referenced Period			**Response Structure**
	14. Period carried over from earlier question 15. Undefined period 16. Unanchoring/rolling period			25. Overlapping categories 26. Missing response categories

Copyrights