The importance of pretesting questionnaires: a field research example of cognitive pretesting the Exercise referral Quality of Life Scale (ER-QLS)

Hilton, CE

Original citation & hyperlink:
Hilton, CE 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*, vol 20, no. 1, pp. 21-34 https://dx.doi.org/10.1080/13645579.2015.1091640

DOI 10.1080/13645579.2015.1091640
ISSN 1364-5579
ESSN 1464-5300

Publisher: Taylor and Francis

This is an Accepted Manuscript of an article published by Taylor & Francis in International Journal of Social Research Methodology on 7th October 2015, available online: http://www.tandfonline.com/10.1080/13645579.2015.1091640

The Importance of Pretesting Questionnaires: a Field Research Example of Cognitive Pretesting the Exercise referral Quality of Life Scale (ER-QLS).

Abstract

The development of questionnaires, surveys and psychometric scales is an iterative research process that includes a number of carefully planned stages. Pretesting is a method of checking that questions work as intended and are understood by those individuals who are likely to respond to them. However, detailed reports of appropriate methods to undertake pretesting are currently underrepresented within the literature. This study presents a detailed protocol of a cognitive interview pretesting approach that informed the development of the Exercise Referral Quality of Life Scale (ER-QLS) - a measure of life-quality designed specifically for structured clinical exercise settings. This documented approach to pretesting was based upon Willis's (2005) recommendations and proved a vital stage in the scale development process, without which the item problems detected would have carried forward into the statistical analyses. The current protocol intends to contribute to reducing the current shortfall in pretesting guidance for practitioners and researchers.

Key words: Cognitive Interviewing, Pretesting, Questionnaire Design, Questionnaire Development, Item Generation, Quality of Life.

Introduction

A frequent difficulty with questionnaire design is that respondents commonly misinterpret questions and this difficulty has been consistently recognised within the literature (e.g., Belson, 1981; Hunt Sparkman & Wilcox, 1982; Nuckols, 1953). Pretesting is a method of checking that questions work as intended and are understood by those individuals who are likely to respond to them. It is also the case that pretesting has the capacity to reduce sampling error and increase questionnaire response rates (De Leeuw, 2001; Drennan, 2003) and may be a valuable method to evaluate whether a new measure performs in the field as planned (Greco & Walop, 1987).

Whilst the value of pretesting has been recognised as critical to the valid measurement of phenomena by survey methodology (Alaimo, Olson & Frongillo, 1999) few studies have been published that report specific pretesting protocols with an appropriate level of detail regarding the methods undertaken or guidance for others (Presser et al., 2004). There is generally no consensus regarding best practices (Beatty & Willis, 2007; Presser et al. 2004) and Collins (2003) has identified that an evaluation of the methods used is often lacking. For example, Subar et al. (1999) reported how cognitive interviewing methods assisted in the development of food frequency questionnaires. However, there was comparatively little detail reported on the specific methods employed compared to that which detailed the formulation of the food frequency measure itself. This typical example makes it difficult for scale developers to make well-informed decisions as to how pretesting could or should be undertaken and perhaps move towards a best practice approach to pretesting scales under construction. Nevertheless, cognitive pretesting is considered an important part of the

questionnaire design research process - the only way to determine in advance whether a questionnaire causes problems for interviewers or respondents (Presser et al., 2004) and also as a valuable addition to psychometric techniques when validating complex tools (De Silva, Harpham, Tuan, Bartolini, Penny, & Huttly, 2006). Indeed, the UK Census have published the procedures undertaken to test the questions used in the 2011 survey for England and Wales (see Census Programme, 2011 for an example of health question pretesting procedures). It is perhaps for these reasons that more recently, attempts have been made to publish the protocol of this vital stage of measure development (e.g., Vis-Visschers & Meertens, 2013) and with specific reference to checking for the influence of language differences on item interpretation (Berrigan, Forsyth, Helba, Levin, Norberg and Willis, 2010; Park, Sha & Pan, 2014). Although generally, understanding regarding the most effective approaches to pretesting, with published examples that researchers and practitioners may use to help inform the development of a suitable protocol, is currently lacking.

Foddy (1993) has offered a critical appraisal of pretesting methods. It is typical for the purpose of pretesting that: a) respondents will be asked to think out loud while completing the test questionnaire and/or b) the interviewer will introduce probe questions to check that the questions are understood and being interpreted as intended. Utilising only the think-aloud technique is difficult and probe questions tend to encourage think-aloud behaviour. Also, a combination of both methods usually removes the need to provide specific think-aloud instruction to participants that may find this difficult. Consequently, a number of researchers have deemed that cognitive interviewing is best characterised as a combination of think-aloud *and* probing procedures (Jobe, 1989; Sudman, Bradburn & Schwartz, 1996; Willis, 2005; Willis,

Royston & Bercini, 1991). Willis (2005) has offered one of the most comprehensive guides to pretesting. However, there are few published examples of the practical application of the methods proposed. As a consequence, whilst the approaches suggested may make intuitive and practical sense, empirical evidence of the application of the proposed techniques in practice is warranted.

The current paper reports the pretesting protocol used to test the performance of the items generated for a new quality of life scale designed for clinical exercise settings. The Exercise Referral Quality of Life Scale (ER-QLS) (see Hilton, Minniti & Trigg, 2015) is a 22-item measure with measurable domains of physical and mental well-being, injury pain and illness and physical activity facilitators. The scale may be scored globally or sub-dimensionally and responds directly to the requirement for exercise referral schemes to evaluate quality of life outcomes for those referred (NICE, 2014).

The purpose of the pretesting reported in the current paper was to utilise the pretesting recommendations of Willis (2005) to: assess how well the items were understood and interpreted, to provide insights into the general quality of the formatting, acceptability and face validity of the measure and to consider if the method of administration (i.e., self-complete, interview or telephone administered) would impact upon respondents interpretation of items. It was also anticipated that others seeking examples of pretesting protocols may utilise the approach employed here for use in their own research.

Methods

Participants

Ethical approval was granted from the National research Ethics Service (NRES) and also a UK university. Twelve females and three males (*N = 15*) were recruited from a local exercise referral scheme for the purposes of pretesting and this cohort of participants was exclusive to the cognitive pretesting phase of scale construction. It is typical that five to ten people are recruited for the purposes of pretesting (Willis, 2005) although recruitment continued until data saturation was reached whereby a concept was mentioned frequently, described in similar ways by different people or when the same ideas arose repeatedly (Holloway, 1997). Participants ranged in age from 36 – 76 years (*M = 60, SD = 10 years*). Table 1 indicates at what stage of their referral into exercise the participants were at when the pretesting was conducted. The employment status of participants included employed (*n = 4*), retired (*n = 10*), and unemployed (*n = 1*). The reasons for referral included weight loss, asthma, diabetes, hypertension, depression, mobility and joint difficulties, smoking cessation and post operative and cardiac rehabilitation (Table 1). These demographics are typical of those who are referred into 12-week exercise programmes (see Dugdill, Graham & McNair, 2005; James Johnston, Crone, Sidford, Gidlow, Morris, & Foster, 2008) and representative of the individuals for whom the ER-QLS was intended which is a recommended sampling approach to cognitive interviewing (Willis 1994, 2005).

[insert Table 1 here]

Materials and Procedure

The construction of the Exercise Referral Quality of Life Scale (ER-QLS) was undertaken in three distinct phases. Initially, focus groups comprising participants (*N = 23*) who had completed at least 12 weeks of an exercise referral programme were used to generate rich data. Phase two consisted of utilising this qualitative data to generate robust items for the ER-QLS by means of a systematic and iterative process with guidance from key texts (e.g., Brace, 2004; Foddy, 1993; Hague, 1993; Oppenheim, 1992; Streiner & Norman, 2008). This process was also employed to identify appropriate response options to each item and the development of appropriate response options was informed by the work of Skevington and Tucker (1999) who have published a comprehensive guide to designing response scales for cross-cultural use in health care. This complete process was subject to a process of iterative peer debriefing (Spall & Stephen, 1998) and this peer debriefing process also continued into the third phase of cognitive pretesting.

In each case, pretesting was facilitated by a researcher who is experienced in qualitative interviewing and counselling methods which helped to facilitate rapport, collaboration and engagement during the interviews. Participants for all 15 cognitive pre-tests were recruited through a local exercise referral scheme in the UK. All participants were provided with a detailed Participant Information Sheet and gave full written consent to their participation. Fifty questions and accompanying response options were subjected to cognitive pretesting protocols that were designed based upon the recommendations of Willis (2005) and utilised both think-aloud and probing question techniques with the purpose of assessing how well the questions were meeting their objectives (Beatty & Willis, 2007). The responses to questions that

comprised the test version of the ER-QLS were developed as a 5-point Likert scale and were carefully selected to match items (see Skevington and Tucker 1999).

As mentioned earlier, it is usual that interviews are conducted with five to ten people (Willis, 2005). However, a greater total number of participants were included in the current study because the questionnaire was administered in three different ways, namely: a) self-complete with think-aloud and question probes ($n = 5$), b) interview-administered with think-aloud and question probes ($n = 5$) and c) telephone administered with think-aloud and question probes ($n = 5$) and data saturation was reached at the same level (i.e., $n = 5$) in each condition. Testing in each of these conditions created a valuable opportunity to assess if the method of administration of the ER-QLS would affect the respondents understanding and interpretation of questions and subsequently inform recommendations for the administration of the final scale.

An initial testing protocol was developed prior to pretesting using the recommendations of Willis (2005) and each of the 50 questions that were subjected to pretesting were allocated corresponding probe questions that reflected areas of clarification as appropriate. For the questions where it was necessary to check the understanding of a particular element of the item wording, the probes were quite specific. For example, to determine if the terms 'physical activity' and 'exercise' in the same question would cause confusion, participants were asked: *"the question uses the words physical activity and exercise in the same question. Does that sound OK to you or would you use something different?"* Other probes were more general and included questions such as *"how did you arrive at that answer?", "was that easy or*

hard to answer?", "I noticed that you hesitated, tell me what you were thinking." The intention was that each interview conducted was collaborative in nature and therefore, the conduct of the interviewer reflected this. Participants were encouraged to generate the majority of the conversation, while the researcher introduced both the pre-formulated and any additional probes at key points throughout the interview in a similar manner to that of a qualitative semi-structured interview (e.g. Braun & Clarke, 2014).

As pretesting progressed and greater clarity was achieved regarding the interpretation of the questions by respondents, the initial probe protocol was amended slightly on two further occasions. For example, the question *"how much do you feel that you incorporate physical activity into your daily lifestyle?"* Had probes: *"what does the term physical activity mean to you?*" and *"In what way does this differ to the word exercise, if at all?*" These probes were included in the initial protocol. However an additional probe was introduced into a second version of the probe protocol: *"what sorts of things come to mind when you think of incorporating physical activity into your daily lifestyle?"* This allowed for an exploration of the term physical activity in addition to prompting for views regarding what is meant by daily lifestyle physical activity. Similarly, an example of an amendment that was made to the probe protocol on the third and final occasion was in response to the question *"how much do you feel that lifestyle factors (e.g., transport, time, childcare, poor health & the weather) affect your ability to be physically active?"* The initial and second probe protocol asked: *"how easy or hard was it to choose an answer?"* The third protocol also included an additional probe: *"you will notice that I provided you with some examples of lifestyle factors (transport, time, childcare, poor health and the weather) did these examples*

help you, or make it more difficult to answer the question?" This allowed for a greater exploration regarding whether the inclusion of differing examples within one question were viewed as problematic in addition to a more general understanding of item response difficulty. A summary table exemplifying some of the types of probes used for a sample of questions can be viewed in Table 2.

[insert Table 2 here]

For the purposes of face validity considerations, participants were encouraged to comment on the complete test measure including formatting, presentation and relevance of its intended use at the end of the interview. This is especially important because the person who designed the questionnaire can very often have a differing perspective to the people for whom it is intended (Greco & Walop, 1987). Each interview was digitally recorded and notes were taken throughout. In order to enhance familiarity with the data, the interviews were listened to on a minimum of two occasions (Hansen, 2006) and notes made during the interview process were combined with any additional notes made from retrospective reviews of the audio data.

Data Handling and Analysis

According to Willis' (2005) recommendations, the pre-test data set was subjected to procedures as follows:

1) Cognitive interviewing outcome reports that summarised the results of each of the three conditions under which the questionnaire was administered were produced.

2) Summary data records were given: a) qualitative consideration of what the problems were and whether they were similar across interviews, and b) quantitative consideration of the frequency with which problems emerged, to gain insights into the severity of the problem.

3) The complete pool of participant responses and recommendations underwent an iterative phase of peer debriefing (Spall & Stephen, 1998) whereby the data were critically reviewed by researchers experienced in scale construction and a consensus regarding question wording was reached.

The cognitive interviewing outcome reports were generated from carefully reviewing the audio data and the accompanying notes made by the researcher for each interview conducted under each of the three conditions (self complete, interview and telephone). For each question, the reports documented whether if any problems were experienced by participants in responding and if so, the nature of the difficulty. These summary reports revealed both the frequency and nature of item difficulties across all three administration methods and were used to generate an overview of item performance (Table 3). The resultant table and individual summary reports were reviewed by researchers, experienced in scale construction and consideration was given to each of the items that had been identified as problematic in terms of whether these items should be amended or removed from the test-item pool.

Results

Qualitative and quantitative consideration of the frequency with which problems emerged gave rise to the following amendments: the wording of six questions was changed to reflect the recommendations of respondents, one question was split into

two separate questions for clarity and accuracy of interpretation and one question was removed completely because respondents considered that is was too general (Table 3).

[insert Table 3 here]

What follows is a detailed account of the feedback provided by participants across the three cognitive pretesting conditions including those pretesting results that did not give rise to amendments.

The introductory instructions describing the purpose and method of completion of the questionnaire posed no difficulty with respect to understanding or interpretation. Comments provided by respondents indicated that the instructions were *"clear", "very clear"* and that there was *"no difficulty at all"* in understanding what was being requested.

Questioning probes were designed to test respondents' understanding of the term 'structured exercise'. Pretesting revealed that respondents interpreted the term as referring to exercise that was *"supervised"* or *"organised", "exercise with a leader"* or that was *"undertaken at a particular time".* These interpretations of 'structured exercise' reflected the interpretation that was intended.

One of the questions tested included examples of lifestyle factors that the focus group participants had identified as being potential barriers to exercise participation. These examples were transport, time, childcare, poor health and the weather. Probes were developed that aimed to clarify if including these examples in the question helped or

hindered a response. Generally, it was felt that the examples helped respondents to complete the question. Indeed for older people, 'childcare' allowed respondents to consider commitments to the care of grandchildren. When participants were asked if removing the examples altogether would make the question clearer, only one respondent agreed. Similarly when questioned as to whether providing a separate question for each example would add clarity to interpretation, a single respondent agreed, but acknowledged the potential increase in respondent burden due to the increase in questionnaire length by employing this amendment. For these reasons, the question remained unchanged.

Questioning probes were developed to establish what respondents understood by the terms 'physical activity' and 'exercise' and if including both terms in the same question was problematic. Results from cognitive interviewing across all three methods of administration revealed that including both terms within the same question posed no difficulties with the understanding, interpretation or the ability to respond to these questions.

When questioned if any injury prevented the respondent from being physically active, probes were introduced to determine if the response scale would account for respondents who did not consider themselves to have any injury. This was particularly important to explore as none of the response scales devised by Skevington and Tucker (1999) allow for a 'not relevant' option. The question asked *"how much does any injury you may have prevent you from being physically active?"* In this case the response options were *'not at all'*, *'not much'*, *'moderately'*, '*a great deal'* or

'completely'. Cognitive pretesting determined that the response *'not at all'* was suitable and selected by those who had no injuries to report.

The test questions *"how competitive are you"* and *"how determined are you"* were two of the most open questions included in the draft test measure and it was anticipated that these questions may subject to misinterpretation. General probes for these questions included: *"how did you arrive at that answer?", "was that easy or hard to answer?", "I noticed that you hesitated, tell me what you were thinking".* Pretesting revealed that some clarity regarding the context of competitiveness and determination would be required (i.e., generally or with respect to exercise). Two subsequent questioning probe protocols that were developed following initial testing included more specific probes that explored the level of specificity needed to respond to the question accurately. Two respondents who had completed the questionnaire under interview conditions and two respondents who had completed under telephone administration conditions reported that the question would require amending to reflect the specificity of general competitiveness and determination or with respect to exercise behaviour. Interestingly, those respondents who self-completed the questionnaire did not report any such difficulties. A close inspection of the qualitative data used to generate the items indicated that amending these questions to refer to exercise behaviour was more appropriate.

Respondents were asked how confident they were to exercise in a leisure centre with minimum support. One respondent who self-completed the questionnaire reported that greater clarity may be required regarding the source of support; for example, from

friends and family or from an exercise instructor. Because the frequency of the difficulty was so low (i.e., a single report) and because another test question that targeted perceived support from *"others"* was amended and split into two to identify: a) support from family and friends, and also b) an exercise instructor to be physically active, this 'leisure centre support' question remained unchanged.

Pretesting indicated that the word 'adhere' included in the question *"how well do you feel you adhere to eating habits that are beneficial to your health and any illness you may have?"* may prove problematic. During an iterative phase of peer debriefing (Spall & Stephen, 1998) researchers experienced in methods of scale development proposed alternative words and phrases such as *"stick-to", "sustain" "maintain" and "uphold"*. It was decided that the question focus was regarding the maintenance of healthy eating habits and for this reason a consensus was reached for the question to be re-worded to *"how well do you feel you maintain eating habits that are beneficial to your health and any illness you may have?"*

With respect to face and content validity considerations, participant responses to ending probes that encouraged feedback on the ease or difficulty with which they completed the scale and if they had any suggestions for further development/amendments indicated that the measure was easy to complete and relevant to them. Participants' reports included that they had *"no difficulty at all"* with completing the questionnaire and that it was *"easy"* to understand and complete. A 64 year old female participant stated that the questions were *"particularly relevant to older people,"* and explained that from her own experience, some of the questions she had been asked throughout the course of her contact with medical professionals

had been less relevant to her daily lifestyle than those contained within the ER-QLS. No suggestions were made for the future development or amendments to the measure.

Discussion

The purpose of the current research was to cognitively pre-test the performance of items and corresponding response options of the ER-QLS to ensure that they were understood and interpreted as intended. The overall presentation and appropriateness of the scale was also assessed to ensure adequate face validity and it was intended that the protocol utilised would not only assess the performance of the ER-QLS but that others may benefit from a deeper understanding of the necessity of pretesting and use the current methods to inform the development of cognitive interviewing protocols for a similar purpose.

The construction of the scale was guided by a number of key texts (e.g., Brace, 2004; Foddy, 1993; Hague, 1993; Oppenheim, 1992; Streiner & Norman, 2008) and overseen by the principles of brevity, simplicity and concreteness (Foddy, 1993). The instructions as to how to complete the test questionnaire, the response scales developed and general format were all based upon an existing validated measure of general life-quality (WHOQOL-BREF; Skevington, Lofty, & O'Connell, 2004). Such attention to the detail of item construction in this manner aimed to ensured adequate interpretability of the measure (Streiner & Norman, 2008). However, cognitive pretesting revealed problems with eight questions in total. This number of problems may have been minimised by the careful attention given to item construction but the detection of these errors further highlights the value of undertaking cognitive

pretesting during the initial phases of scale construction at the item level. Had cognitive pretesting not been undertaken, these problems would have been carried forward into the scale level and psychometric phases of research that followed which is important because no amount of statistical manipulation can account for poorly chosen questions (Streiner & Norman, 2008). In similar terms, a qualitative interview guide should be considered as a data collection too or perhaps collaborative encounter between interviewer and interviewee (Qu & Dumay, 2011) l that is potentially subject to the same participant interpretation difficulties. Consequently, it is reasonable to suggest that the findings of the current research may have helpful implications for pretesting qualitative interview guides. The cognitive pretesting approach documented here has the capacity to check that interview participants: understand the interview questions as they were intended, have the capacity to answer, are not burdened by the quantity and focus of questions asked and that memory recall does not inhibit their ability to respond, for example.

The requirement of eight item amendments from a pool of 50 is comparatively less than identified in examples of previous research. For example, early work by Nuckols (1953) reported that one in six participants incorrectly re-defined a test question presented to them when asked to explain the question in their own words. Two items in the current study adopted this re-wording approach to item testing. These items were: *"how confident are you in your ability to participate in regular physical activity and exercise?* And *"how would you rate your current knowledge of the benefits of physical activity and exercise for health?"* Neither item posed misinterpretation difficulties. More recently than Nuckols' (1953) findings, Belson (1981) reported that only 29% of respondents offered the intended interpretation of a question and,

moreover, that the highest score of accuracy for any of the questions tested was only 58%. However, it is important to recognise that Belson (1981) chose to test those questions deemed to be particularly problematic from a review of existing measures and that are typical of problematic items. For example, questions that required more than one answer. Such problems were avoided during the construction of the ER-QLS by ensuring that items of this nature were not included in the test pool. This may account for the particularly frequent incidences of misinterpretation reported by Belson (1981) and also reinforces the value of employing a rigorous approach to item construction before commencing pretesting.

The combination of both think-aloud and probe techniques alongside evolved probe protocols that responded to the changes in the depth of understanding regarding question performance was particularly effective. Specifically, this approach to pretesting allowed for flexibility within an otherwise structured design which complemented the overall rigour of the research in terms of how the data were collected and reported. Previous studies have explored the use of think-aloud technique only (e.g., French, Cook, Mclean, Williams, & Sutton, 2007). However, French et al. (2007) discuss the results within the context of the performance of a Theory of Planned Behaviour questionnaire in much more detail than a critique of the think-aloud pretesting techniques applied. In turn, this limits the learning to be gained from the experiences of applying the think-aloud technique in such a manner. A further limitation, but that is recognised by French et al. (2007) was that utilising the think-aloud approach in isolation required participants to verbalise their thoughts and if this is not done effectively, problems remain undetected. The current study employed both think-aloud and probe techniques and as a consequence this limitation

was minimised. Furthermore, the results reported in the current study support what has previously been identified as face validity criterion (e.g., Murphy & Davidshofer 2001; Rust & Golombok, 2009) which should be considered an asset to the validity procedures often undertaken during scale development.

No problems were reported regarding the response options either in terms of understanding the wording or the appropriateness of the question to which they were allocated. The iterative phases of peer debriefing undertaken during the item and response construction phase of the current research undoubtedly contributed to the process of ensuring that the most appropriate response scales were matched with each item. However, the careful selection of response scale options appropriate to the design of the scale under construction, particularly those that have been grounded in scientific research and field tested (Skevington & Tucker, 1999) also reduced the likelihood of respondent error and/or misinterpretation. In addition, the absence of reported difficulty regarding the interpretation or appropriateness of the response scales used in the current study offers support for the use of these response options in the development of new population specific quality of life (QoL) scales. Furthermore, a particular strength of the current study was that the constructed items were tested under three distinct conditions: self-complete, interview and telephone administered. It is more usual that if scales under construction are pre-tested, only a single condition (typically interview-administered) is employed (e.g., Wildy & Clarke, 2009). Pretesting results would suggest that the ER-QLS is suitable to be administered in any of the three conditions as no single method generated distinct difficulty with understanding or interpretation.

Questionnaire data comprise an important part of data collection across a broad range of the science and social science disciplines both for research purposes and in practice. A critical consideration is that the data obtained from such measures are only as valid as the items used to measure them Alaimo et al. (1999). Cognitive pretesting can only serve to identify, not resolve or amend problem items as this is the role of the researcher (Willis, 2005) and therefore, it is critical that items are rigorously constructed and honour good practice of question formulation before reaching the pretesting stage and that they are subject to review and amendment where appropriate thereafter. Typical examples of good practice approaches to question formulation include avoiding ambiguous language or the use of jargon for example (Foddy, 1993; Hague, 1993). Incorporating structured pretesting protocols that identified potential problems with items prior to pretesting combined with both think-aloud and probe questioning techniques facilitated the clarity of data reported. Such clarity of data a) increased the likelihood that recommendations were interpreted appropriately by the researcher, b) facilitated the production of reports that documented the item modifications recommended by respondents and c) supported the clarity of data communicated to the research team for an amendment consensus (Spall & Stephen, 1998; Willis, 2005).

The probing protocols ensured that an adequate amount of attention was given to assessing the performance of test items and that potential item failures where not overlooked. Additionally, in cases where respondents struggled to think-aloud or the frequency with which they undertook this task reduced, introducing additional probes helped to generate on-going feedback from respondents. In this respect, the think-aloud and probing techniques tended to complement one another and with respect to

the aim of item problem detection, neither approach could be deemed any more effective than the other which has been found previously (Priede & Farrall, 2011). In addition, whilst cognitive pretesting is not a qualitative approach as such, it was considered that the interviewer's expertise in qualitative interviewing contributed to the quality and depth of the feedback provided from participants in each case and in this respect, it is recommended that cognitive interviews are conducted by those with similar expertise.

In summary, the recommended approaches to cognitive pretesting recommended by Willis (2005) proved effective in testing the item performance and initial acceptability and face validity of the ER-QLS. This approach also established that the ER-QLS may be administered either in self-complete, interview or telephone format. Despite careful planning, cognitive pretesting highlighted problems with eight of 50 questions included in the original item pool (the final validated measure comprises 22 items) which further emphasises the critical necessity to cognitively pre-test scales under construction. In broader terms, the results of the current study support recent findings that both think-aloud and probing approaches to cognitive pretesting that are guided by the recommendations of Willis (2005) are effective at detecting item problems (Buers, Triemstra, Bloemendal, Zwijnenberg, Hendricks & Delnoj, 2014). Although it is recognised that this approach may be best suited to pretesting scales that are designed to measure similar multi-faceted constructs to that of life-quality and as such this is worthy of further exploration. Whilst it has been recognised that cognitive interviewing may be as much an art as it is science (Beatty & Willis, 2007) - and arguably therefore make attempts at developing a consistent approach problematic. It is intended that the current findings will add to the growing evidence-base for the

value of pretesting such that it is not omitted from scale construction. Whilst guidance on pretesting is available in the literature (one of the most comprehensive being Willis's 2005 proposals), mobilising such guidance into examples of empirical field testing is in its infancy. This lack of specific guidance form the field may account for cognitive pretesting seemingly not comprising a standard part of the development process of survey instruments despite this recommendation previously (Collins, 2003). There is a critical and exciting opportunity to develop the evidence-base for examples of rigorous, detailed and effective approaches to the pretesting of survey and scale items. The findings from the current paper contribute to support for the use of Willis' (2005) recommendations; perhaps especially for multi-faceted scales as a robust practical and effective method for researchers and practitioners to follow.

References

Alaimo, K., Olson, C. M., & Frongillo, E.A. (1999). Importance of cognitive testing for survey items: an example from food security questions. *Journal of Nutrition Education,* 31, 5, 269-275.

Beatty, P. C. & Willis, G.B. (2002). Research synthesis: the practice of cognitive interviewing. *Public Opinion Quarterly,* 71, 2, 287-311.

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

Belson, W.A. (1981). *The design and understanding of survey questions.* Aldershot: Gower.

Berrigan, D., Forsyth, B., Helba, C., Levin, K., Norberg, A., and Willis, G.B. (2010). Cognitive testing of physical activity and acculturation questions in recent long-term Latino immigrants. *BMC Public Health*, 10, 2-14.

Blair, J. & Presser, S. (1993). Survey procedures for conducting cognitive interviews to pretest questionnaires: a review of theory and practice. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 370-375.

Braun, V. & Clarke, (2014). *Successful qualitative research: a practical guide for beginners.* London: Sage.

Buers, C., Triemstra, M., Bloemendal, E., Zwijnenberg, N. C., Hendricks M. & Delnoj, D. M. J. (2014). The value of cognitive interviewing for optimizing a patient experience survey. *International Journal of Social Research Methodology* 17(4), 325-340.

Census Programme (2011). *Final recommended questions for the 2011 census in England and Wales: health* Retrieved from http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/2011-census-questionnaire-content/question-and-content-recommendations-for-2011/final-recommended-questions-2011---health.pdf.

Collins, D. (2003). Pretesting survey instruments: an overview of cognitive methods. *Quality of Life Research,* 12, 229-238.

De Leeuw, E.D. (2001). Reducing missing data in surveys: an overview of methods. *Quality and Quantity*, 35, 147-160.

De Silva, M. J., Harpham, T., Tuan, T., Bartolini, R., Penny, M. E., & Huttly, S. R. (2006). Psychometric and cognitive validation of a social capital measurement tool in Peru and Vietnam. *Social Science and Medicine,* 62, 941-953.

Drennan, J. (2003). Cognitive interviewing: verbal data in the design and pretesting of questionnaires. *Journal of Advanced Nursing,* 42, 57-63.

Dugdill. L., Graham, R. C., & McNair, F. (2005). Exercise referral: the public health panacea for physical activity promotion? A critical perspective of exercise referral schemes; their development and evaluation. *Ergonomics,* 48, 1390-1410.

Foddy, W. (1993). *Constructing questions for interviews and questionnaires. Theory and practice in social research.* London: Cambridge University Press.

French, R.C., McLean, M., Williams, M., & Sutton, S. (2007). What do people think about when they answer theory of planned behaviour questionnaires? A 'think-aloud' study. Journal of Health Psychology, 12(4), 672-687.

Greco, L. D. & Walop, W. (1987). Questionnaire development: the pretest. *Canadian Medical Association Journal,* 136, 1025-1026.

Hague, P. (1993). *Questionnaire design*. London: Kogan Page Limited.

Hansen, E. C. (2006). *Successful qualitative health research. A practical introduction.* London: Open University Press.

Hilton, C. E., Minniti, A., & Trigg, R. (2014). Psychometric properties of the exercise referral quality of life scale (ER-QLS): a psychological measurement tool specifically for exercise referral. *Health Psychology Open*. Retrieved from http://hpo.sagepub.com/content/2/2/2055102915590317.full.pdf+html

Holloway, I. (1997). *Basic Concepts of qualitative research.* London: Blackwell Science.

Hunt, S.D., Sparkman, R.D., & Wilcox, J.B. (1982). The pretest in survey research: issues and preliminary findings. Journal of Marketing Research XIX, 269-273.

James, D.V.B., Johnston, L.H., Crone, D., Sidford, A.H., Gidlow, C. Morris, C & Foster, C. (2008). Factors associated with physical activity referral uptake and participation. *Journal of Sports Sciences*, 26(2), 217-224.

Jobe, J. B. & Mingay, D. J. (1989). Cognitive research improves questionnaires. *American Journal of Public Health* 79, 1053-5.

National Institute for Health and Clinical Excellence (NICE) (2014) *Exercise Referral Schemes to Promote Physical Activity.* London: The Stationery Office.

Nuckols, R.C. (1953). A note on pretesting public opinion questions. Journal of Applied Psychology, 37, 119-120.

Oppenheim, A. N. (1992). *Questionnaire design, interviewing and attitude measurement new edition.* Continuum: London.

Park, H., Sha, M., M. & Pan, Y. (2014). Investigating validity and effectiveness of cognitive interviewing as a pretesting method for non-English questionnaires: Findings from Korean cognitive interviews. *International Journal of Social Research Methodology* 17(6), 643-658.

Presser, S., Couper, M.P, Lessler, J.T., Martin, E., Martin, J., Rothgeb, J. M., & Singer, E. (2004). Methods for testing and evaluating survey questions. Public Opinion Quarterly 68(1), 109–30.

Priede, C. & Farrall, S. (2011). Comparing results from different styles of cognitive interviewing: 'verbal probing' vs. 'thinking aloud.' *International Journal of Social Research Methodology* 14(4), 271-287.

Sandy Q. Qu, John Dumay, (2011). The qualitative research interview. *Qualitative Research in Accounting & Management*, 8(3), 238 – 264.

Skevingon, S. M., Lofty, M., & O'Connell, K.A. (1994). The World Health Organization's WHOQOL-BREF quality of life assessment: Psychometric properties and results of the international field trial. A report from the WHOQOL Group. *Quality of Life Research,* 13(2), 299-310.

Skevington, S. M. & Tucker, C. (1999). Designing response scales for cross-cultural use in health care: data from the development of the UK WHOQOL. British Journal of Medical Psychology, 72, 51-61.

Spall, S., & Stephen, F. (1998). Peer debriefing in qualitative research: Emerging operational models, *Qualitative Enquiry*, 4(2), 280-292.

Streiner, D. L., & Norman, G. R. (2008). *Health measurement scales: A practical guide to their development and use.* (4th ed.). New York: Oxford University Press.

Subar, A. F., Thompson, F. E., Smith, A. F., Jobe, J. B., Ziegler, R. G., Potischman, N., Schatzkin, A.,…Harlan, L. C. (1999). Improving food frequency questionnaires: a qualitative approach using cognitive interviewing. *Journal of the American Dietetic Association,* 95(7), 781-788.

Sudnman, S., Bradburn, N. M., & Schwartz, N. (1996). *Thinking about answers: the application of cognitive processes to survey methodology.* San Francisco: Jossey-Bass.

Vis-Visschers, R., & Meertens, V. (2013). Evaluating the cognitive interviewing reporting framework (CIRF) by rewriting a Dutch pretesting report of a European health survey questionnaire. *Methodology,* 9(3), 104-112.

Willis, G. (1994). *Cognitive interviewing and questionnaire design: a training manual.* Cognitive methods staff working paper series, No. 7. Hyattsville, MD: National Center for Health Statistics.

Willis, G.B. (2005). *Cognitive interviewing. A tool for improving questionnaire design.* London: Sage.

Willis, G. B., Royston, P., & Bercini D. (1991). The use of verbal report methods in the development and testing of survey questionnaires. *Applied Cognitive Psychology,* 5, 251-67.

Acknowledgements

The researcher wishes to acknowledge Dr Antoinette Minniti and Dr Richard Trigg for their roles in peer debriefing throughout the development of the ER-QLS.