African Journal of Hospitality, Tourism and Leisure, Volume 6 (2) - (2017) ISSN: 2223-814X

How many items are too many? An analysis of respondent disengagement when completing questionnaires

Professor Renier Steyn
Graduate School of Business Leadership
University of South Africa
PO Box 392, Unisa, 0003
South Africa
steynr@unisa.ac.za

Abstract

Researchers are often confronted with the dilemma of how many items to include in a questionnaire. On the one hand the researcher requires accurate and comprehensive information but, on the other, textbooks alert us to the danger of respondent fatigue and respondents disengaging as a result of completing lengthy questionnaires. This paper aims to empirically address the question of how many items a questionnaire can contain before additional items can be considered to be superfluous, as well as the question of at what stage fatigue and disengagement set in. The responses of 3 180 individuals to a questionnaire of 153 items were analysed, focusing on the number of missing cases as well as on zero-variance and flat response patterns. The results reveal that neither the number of missing values, nor zero-variance or flat response patterns, are a function of the number of items (already) completed. The results further reveal that the response pattern of respondents, on a battery of 153 items, can be attributed to the types of items included in the questionnaire rather than to the number of items completed. Recommendations are presented on strategies to design questionnaires that include a low number of items, but still fulfil the requirements of reliability and validity.

Keywords: Questionnaires; tourism; items; measurement.

Background

Most people are accustomed to receiving questionnaires[1], often via the internet, on a diverse range of topics. This frequently happens when the recipient has returned from a holiday and when service providers wish to enquire about client experiences of, among other things, the quality of the trip to the destination, the features of the destination, satisfaction with accommodation and meals, and other services rendered Some recipients complete these questionnaires, some do so up to a point and some do not complete them at all. Although annoying from a consumer perspective, those in the field of tourism, tourism marketing, and market research, are very interested in finding out how consumers evaluate their products and services.

The aim of this paper is to add to the present literature on questionnaire design by determining the point at which respondents disengage from completing a questionnaire mindfully. The focus will be specifically on that that moment when respondents lose interest in the survey due to its length. The principal assumption is that, after completing a critical number of items in a lengthy questionnaire, most respondents will reach a point of fatigue and start disengaging from answering the questions in a considered manner. This aim of this research is to identify that number of items – but by using a different strategy than other researchers, such as Wagner-Menghin and Masters (2013), have used to investigate the same phenomenon.

[1] The term "questionnaire" is used in the article as a synonym for survey. Also, subsections in the questionnaire are referred to as tests. The term "item" will be used to refer to the different questions included in the questionnaire or test.

The information arising from this investigation may be valuable to those in the fields of tourism, tourism marketing, and market research, as proper questionnaire design may result in higher response rates, a higher number of completed surveys, and more reliable and valid results.

Literature review

From what follows below it will become clear that the determinants of a quality questionnaire are multiple and complex, and that the designers of such questionnaires are required to make several subjective decisions when considering questionnaire length. Focusing on the length of questionnaires, Wright (1992) argues that matters related to questionnaire length are more often related to folklore and accident than to intention. A summary of some of the information that designers should consider when deciding on the length of a questionnaire is presented below.

The concepts of reliability and validity are central to all measurement and also influence the length of a questionnaire. These concepts will first be introduced here before they are discussed in relation to questionnaire length:

- In general, reliability refers to the constancy of the scores generated through responding to a questionnaire (Shaughnessy, Zechmeister, & Zechmeister, 2009; Tredoux, & Durrheim, 2013). Many types of reliability measures are reported in literature, including test-retest reliability, half-split reliability, parallel-forms reliability and internal consistency (Cohen, Swerdlik, & Sturman, 2013; Moerdyk, 2015). The measure of reliability most often used is internal consistency (Cronbach, & Shavelson, 2004; Novick, & Lewis, 1967; Kaiser, & Michael, 1975; Lord, & Novick, 1968), and it is expressed as coefficient alpha. Coefficient alpha shows stability of scores (similar to those achieved through the test-retest approach) and can, as such, be a useful estimate of reliability (Gregory, 2011).
- Validity refers to the degree to which evidence and theory support the interpretations of scores generated through completing questionnaires (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Cohen, Swerdlik, & Sturman, 2013; Gregory, 2011). The classic trinitarian view of validity is still common among contemporary authors on psychometrics (Cohen, Swerdlik, & Sturman, 2013; Gregory, 2011; Moerdyk, 2015). According to this view, content validity, criterion-related validity, and construct validity are identified. These elements are discussed in more detail below.

When considering reliability within the context of the length of the questionnaire, using more items almost always seems to be beneficial. As Anastasi and Urbina (1997) stated, "other things being equal, the longer a test, the more reliable it will be" (p. 96). Wells and Wollack (2003) agree and state that, in general, longer questionnaires produce higher reliabilities. Adding items to a survey generally enhances the reliability of the questionnaire and, using the Spearman-Brown prophecy formula (Nunnally, 1978; Stanley, 1971), it is possible to calculate the effect of adding further similar items to a questionnaire, given that the reliability of the existing questionnaire is already known. At some stage, however, additional items may yield very little value. Unfortunately, the formula does not explain how to determine the questionnaire length to obtain an acceptable level of reliability (Wright, 1992). Indeed, the answer to determining the correct length of a questionnaire may not be embedded in the reliability of the questionnaire at all.

The three major types of validity may influence the length of a survey. The influence of each type of validity on the length of a questionnaire is discussed below, following a short definition of the types of validity:

- Content validity is reflective of the judgement, of often a panel of experts, of the degree in which items in a questionnaire are adequately representative of the universe of what is being assessed (Cohen, Swerdlik, & Sturman, 2013; Gregory, 2011). Therefore, enough items should be included to assess every aspect of the domain under

investigation. Satisfaction with a holiday experience can, in other words, not be assessed with a unidimensional item such as "To what extent did the hotel provide a value-for-money service?"

- Construct validity relates to the extent to which scores on the assessment relate to other constructs in the way they in which they are expected to relate (DeVellis, 2012). Moerdyk (2015) uses theoretical validity as a synonym for construct validity and states that the basic question of construct validity is whether the results are in line with what is theorised. In many respects, this relates to the thinking behind the survey and has little to do with the number of items. For example, if the intention of a general holiday hotel satisfaction questionnaire was to gauge the possibility of a return visit to the hotel, the questionnaire may be inadequate in terms of construct validity, but if it is used in market research on what types of hotels may be preferred, it could be adequate. Construct validity therefore relates to the use of gathered data, and does not directly influence the number of items which need to be included in a questionnaire.
- Criterion-related validity is demonstrated when a measure is effective in estimating the respondents' behaviour on some outcome measure (Gregory, 2011), with the outcome measure being the criterion. Stated differently, it is "a judgement of how adequately a test score can be used to infer an individual's most probable standing on some (other) measure of interest" (Cohen, Swerdlik, & Sturman, 2013: 190). This may influence the number of items, should the questionnaire administrators want to statistically test the probability of consequent behaviour using items in the same questionnaire to do so. In such a case, the designer of the questionnaire may furnish additional items regarding the outcome behaviour and for inclusion in the questionnaire. They may, for example, ask whether the respondent would recommend the hotel to friends or even enquire about the likelihood that he or she might return to the same hotel. Adding criterion items to the questionnaire will also allow the questionnaire administrators to pinpoint the antecedents to the desired outcome behaviour.

Content, as well as criterion validity concerns, may thus result in adding more items to a questionnaire.

Item quality has a large impact on the length of a questionnaire, as including items which do not discriminate does not add real value (Wells & Wollack, 2003). Unusable items are thus included, which may result in unnecessary fatigue. Lunz (2009) reports, along the same lines, that the quality of items is as important as, or more important than, the absolute number of items when achieving satisfactory reliability is the aim.

Wright (1992) suggests four quality factors which could determine the length of a questionnaire. The first has to do with the number of items needed to ensure that those who complete the questionnaire do not misinterpret the central question posed in the questionnaire. The question must therefore be asked as to whether there are enough items to accurately reflect the intention of the questionnaire.

The second factor relates to the calibration of the answer to the posed central question. By including more items, it is possible to comprehensively qualify the variable under investigation.

Thirdly, when more items on the same variable are included, the effects of responses, which may be due to normal or random error, will decrease. Where respondents complete more than one item, the responses will include fewer normal or random errors.

Fourthly, enough items to enable sufficiently precise inferences to be made in the process of decision-making by those who administered the questionnaire are required. In this regard, it is valuable to note that doubling precision, thus halving the standard error, requires four times the original number of items (Wright, 1992). Wright's (1992) suggestions contribute to ensuring the reliability as well as the validity of the questionnaire.

Item design, particularly with reference to the number of points on the scale, and linked to Wright's (1992) notion of calibration, may play an important role. Fitzpatrick and Yen (2010) report that, to obtain acceptable reliabilities and accurately equated scores, questionnaires should have at least eight 6-point items or at least twelve 4-point item. The aforementioned research was conducted within a specific context, and should therefore not be regarded as generalisable. It does, however, add the important matter of the width of the scale to the length of the questionnaire debate.

Wells and Wollack (2003) suggest that, before lengthening a questionnaire (to improve reliability and validly), it is important to consider practical constraints such as the time limit and fatigue level of the respondents. As respondent fatigue is central to this article, research regarding both this aspect and questionnaire length is relevant:

- Tulsky and Zhu (2000) found, in groups that were matched on key demographic variables, that fatigue did not play a role in the scores on a cognitive test, irrespective of whether the test was administered before or after fatigue set in.
- Ackerman and Kanfer (2009: 163) reported that "subjective fatigue increased with increasing time-on-task", but found that "mean performance increased in the longer test length conditions, compared with the shorter test length conditions". They also found that individual differences in personality, interest and motivation have greater predictive power than the questionnaire length when predicting subjective cognitive fatigue associated with completing questionnaires.

Kanfer (2011), following a review of literature on fatigue, proposes that, apart from the most extremely fatiguing conditions, the impact of subjective cognitive fatigue on future task-efforts depends on the perceived utility of achieving high levels of task-performance, as well as on the instigation of self-regulatory activities to sustain or increase task efforts in the face of subjective mental fatigue.

It is within this context of effort being related to the utility of the outcomes and individual trait differences that this study was conducted. When would a person, who does not have any particular interest in the outcome of his or her responses to a questionnaire, lose interest in the task and disengage?

Method

Research design

A cross-sectional survey design was used to collect the data. Only quantitative data was collected and the analysis of the data focused on the response patterns of the respondents, rather than on the contextual meaning of the responses.

Procedure

Archival data, collected by 53 research assistants from 53 South African organisations, were used for the analysis. The author of this paper gained access to the data based on being the guardian of the data. The dataset consisted of data representative of 10 individual tests.

It is important to note that these questionnaires were completed by respondents who had very little interest in the outcome of the assessments. All personal identifiers were removed from the answer sheets, which created anonymity. Also, no interventions followed these assessments. As such it would be fair to state that the respondents' external motivation to complete the assessment in a diligent manner would have been low, similar to most individuals who complete questionnaires distributed in the tourism industry.

The data linked to each of the 10 individual tests, which made up the questionnaire data, were analysed separately and the number of missing items per test was calculated. Furthermore, the number of respondents per test who responded in an unengaged manner was also recorded. An unengaged respondent was defined as a respondent whose response pattern was without noticeable fluctuation. An example of an unengaged respondent would be one

who answers 3,3,3,3,3,3,3,3 and 3 or 5,5,5,5,5,5,5,5 and 5 on a nine-item test. These were called zero-variance respondents. Low variance respondents were also identified. The rule-of-thumb used here was to label as low variance respondents those respondents whose responses had a variance of less than .5 of a standard deviation of the group's response style.

The questionnaire

The questionnaire used consisted of a battery of 10 tests, of which the length varied between 8 and 21 items. In total, the number of items was 154. The tests were standardised psychometric instruments used regularly within the context of work psychology. The respondents needed to respond to a statement, indicating the extent to which they agreed or disagreed with the statement. For the sake of this research, the content of the tests was not deemed important as the focus was rather on the chronological placement of the test in the questionnaire; in other words, was the test presented earlier or later in the questionnaire?

Analysis

Two measures were proposed to assess whether respondents lost interest in the questionnaire:

- The first related to the number of items responded to, suggesting that respondents who lost interest in the test would complete fewer items. The respondent may not have read a particular item and thus may not have responded to that item.
- The second measure considered that the respondent may not have read the item and may simply have answered all the items in the same manner, normally selecting a middle option. Using this strategy, the respondent may have completed the test or even the whole questionnaire but provided very little information regarding his or her perceptions.

Although it will be acknowledged that missing answers or answers with zero or lower variance may be the result of factors other than test fatigue, fatigue is also deemed as a possible explanation for such behaviour.

Based on the aforementioned, two hypotheses were tested:

- The first hypothesis tested aimed to determine, through inspection, when the respondents stopped answering the questions. After what number of items did the count of missing value start to increase? This was done through inspecting a table drawn up for this purpose, as well as by visual inspection of a chart depicting the same data.
- The second hypothesis tested aimed to determine, through inspection, when the respondents stopped answering the questions mindfully. After what number of items did the count of zero and low variance responses start to increase? This was also done through inspecting a table drawn up for this purpose, as well as by visual inspection of a chart created for this purpose.

No statistical test to detect differences was performed. However, trend lines were generated to integrate the different points on the charts.

Findings

Demographics of the sample

A total of 3 180 respondents completed the questionnaire, 57.1% men and 42.5% women. Percentage wise, 8.3% identified as Asian, 58.4% as Black, 8.4% as Coloured, and 24.6% as White. The average age of the respondents was 37.80 years (standard deviation = 9.11). A small percentage (5.0%) reported that they had received less than 12 years of formal schooling; 25.5% said that they had completed 12 years of formal schooling, whilst 40.2% reported that they had completed a degree or diploma. 28.9% Indicated that they had a higher

degree or higher diploma. The respondents therefore represented a wide spectrum of the South African population.

Results pertaining to the first hypothesis

The first hypothesis could read as follows: After what number of items does the count of missing value start to increase? In Table 1, the sequencing of the tests is presented (Column 1), the number of items in the specific test (Column 2), the cumulative number of items in the questionnaire (Column 3), the number of missing items in the specific test (Column 4), and, most importantly, the percentage of missing items per test (Column 5).

Table 1

Missing values over time

Sequence of tests	Items per subtest	Cumulative number of items in battery	Missing values	Missing values per items in subtest
1	20	20	113	5.65
2	9	29	31	3.44
3	10	39	20	2.00
4	21	60	219	10.42
5	16	76	72	4.50
6	8	84	72	9.00
7	18	102	136	7.55
8	21	123	46	2.19
9	14	137	57	4.07
10	17	154	407	23.94

From Table 1 it can be observed that the number of missing values, as per the percentage missing values per number of items in the test, increased quite dramatically for the last test, after 137 questions had been answered.

Figure 1 presents the same information graphically. On the y-axis, the number of items completed (# items) is presented along with the percentage of missing cases (# missing). The x-axis presents the test number.

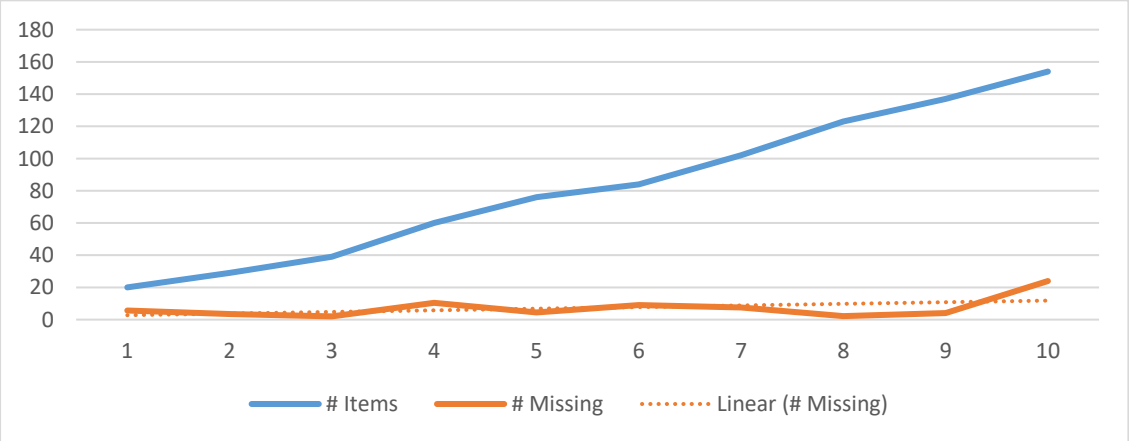

Figure 1: Cumulative number of items completed and number of missing values

From Figure 1 it can be observed that the number of missing values fluctuates, and that the strongest increase in missing values occurs following test 9, after 137 items. The trend line (dotted) shows a slight positive trajectory.

Results pertaining to the second hypothesis

The second hypothesis could read as follows: After what number of items does the count of zero and low variance responses start to increase?

Table 2 is identical to Table 1 in respect of the first three columns. Column 1 of Table 2 presents the sequencing of the tests, Column 2 reports the number of items in the specific test, and Column 3 details the cumulative number of items in the questionnaire. The number of respondents who reported with zero variance is reported in Column 4 and the number of respondents with low variance is presented in Column 5.

Table 2
Zero variance and low variance responses over time

Sequence of tests	Items per subtest	Cumulative number of items in battery	Individuals with 0 variance responses (a)	Individuals with low variance (b)	a + b
1	20	20	0.18	1.06	1.25
2	9	29	7.04	10.66	17.70
3	10	39	3.08	6.98	10.06
4	21	60	2.32	5.97	8.30
5	16	76	0.22	3.74	3.96
6	8	84	1.41	10.34	11.76
7	18	102	0.97	5.09	6.06
8	21	123	2.45	3.14	5.59
9	14	137	5.56	3.42	8.99
10	17	154	0.34	2.32	2.67

Figure 2 presents the same information as presented in Table 2 graphically. The y-axis shows the percentage of respondents presenting either zero variance or low variance, as well as the total – that is the sum of zero variance and low variance. The x-axis presents the test number.

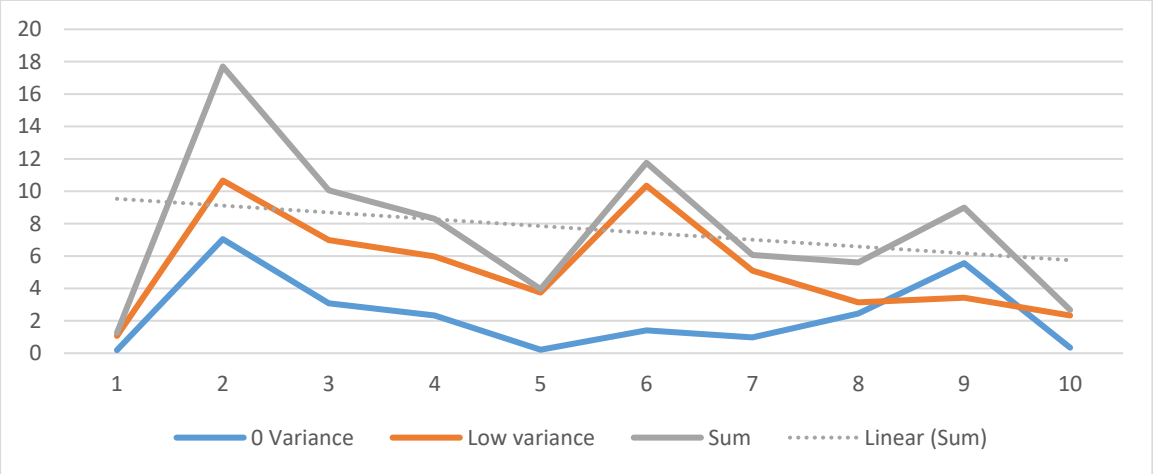

Figure 2: Percentage of 0 variance responses and low variance responses

The trend line (dotted), for the sum of the zero variance and low variance responses, shows a negative trajectory. It is evident from Figure 2 that zero variance response was the highest at Test 2 and at Test 9, whilst low variance response was the highest at Test 2 and Test 6.

In order to gain more insight into the nature of zero variance response, the zero-variance response for Test 2 was analysed in greater detail. The findings are presented in Table 3, below.

Table 3

Analysis of responses with zero variance

Response style	Frequency	Percentage of zero variance responses	Percentage in total group
0, 0, 0, 0, 0, 0, 0, 0, 0	5	2.26	0.15
1, 1, 1, 1, 1, 1, 1, 1, 1	2	0.90	0.06
2, 2, 2, 2, 2, 2, 2, 2, 2	2	0.90	0.06
3, 3, 3, 3, 3, 3, 3, 3, 3	49	22.17	1.54
4, 4, 4, 4, 4, 4, 4, 4, 4	19	8.59	0.59
5, 5, 5, 5, 5, 5, 5, 5, 5	52	23.52	1.63
6, 6, 6, 6, 6, 6, 6, 6, 6	92	41.62	2.89
Total	221	100.00	6.94

The information from Table 3 is presented below in chart form.

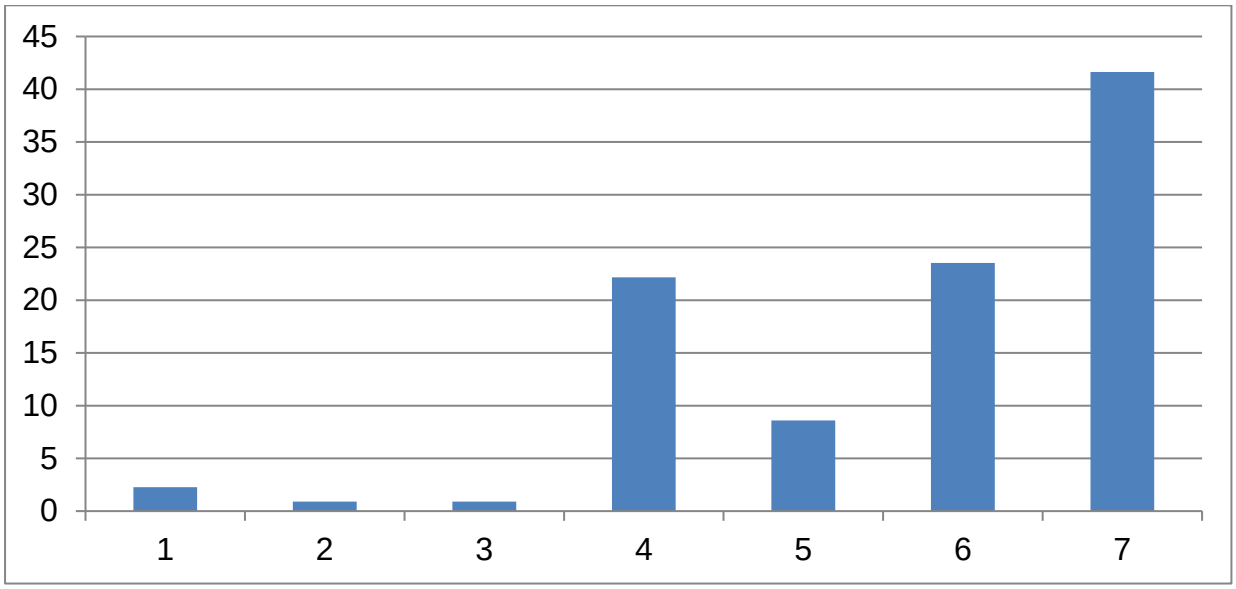

Figure 3: Consistent response pattern for zero variance cases

Note: In the chart above, 1 represents Option 0, 2 refers to Option 1 and so forth. This is based on the software in which the chart was created.

From Figure 3 it can be observed that the middle value (4 or Option 3) is frequently selected, but not as often as either Option 5 or Option 6.

Discussion

The responses which were analysed in this research were those of individuals who had little interest in the outcomes of the survey. The quality of their responses could thus be deemed similar to the quality of responses submitted by individuals who complete questionnaires

following a holiday experience. The assumption is thus that the results obtained with this sample will be similar to the results garnered from a sample of tourists.

The first hypothesis stated that missing values would increase as the number of items increased. Despite the small increase in missing values, particularly at Test 10, the trend line for missing values was relatively flat. These results complement the findings of Ackerman and Kanfer (2009), as well as those of Tulsky and Zhu (2000), who found that performance did not decline as fatigue set in.

The second hypothesis was analogous to the first hypothesis, suggesting that zero and low variance responses would increase over time. The trend line showed a negative trajectory, contrary to what was expected. The peaks at Test 2, early in the questionnaire, and at Test 6, in the middle of the questionnaire, suggested that test-specific characteristics, rather than when the test appeared on the timeline, may have influenced the response pattern.

When considering the response pattern in Test 2, as depicted in both Table 3 and Figure 3, it is interesting to note that the middle option (as mentioned in the method section of this paper) was not the option most often endorsed. The high score (Option 6) was selected most often. This also reflects that the type of test may play an important role in the response style of those who complete tests.

What conclusion might be reached about the ideal length of questionnaires and the design of surveys for the tourism industry? Two aspects might be considered here: Firstly, the length of the questionnaire and the fatigue associated with completing the questionnaire may not play a significant role in the respondent's response style. Secondly, the response style may be a function of the type of items used, more so than a function of the number of items included in the questionnaire.

Recommendations

Previous research, as well as the empirical part of this research, has indicated that the length of questionnaires does not influence the performance of the respondents and, as such, designers of questionnaires should not limit themselves unnecessarily in order to manage this perceived restraint. The empirical research conducted for this article additionally suggests that designers should be aware that the type of items they use may influence the response style of those individuals who complete the questionnaires.

From the literature gathered for this article, several recommendations can be made to those interested in designing questionnaires:

- Include enough items to make precise-enough inferences from the results. Also include enough items to make sure the central question is not misinterpreted. Posing the same question in different ways may lead to additional items in the questionnaire, but it could also improve the quality of the response. Furthermore, be very clear on what is asked when designing items, so as to ensure that the interpretation of questions is uniform. It may also be necessary to conduct a pilot study before introducing a questionnaire.
- To improve the accuracy of the assessment, and to improve the calibration of the response, a scale with more points adds precision and reduces the number of items needed.
- Ensure that the content is assessed comprehensively, addressing all aspects of the phenomenon. Satisfaction with a hotel booking may be more than a unidimensional satisfaction with the hotel; it could also include aspects such as the quality of the room, the speed of service, and the check-in process. Such a questionnaire will require more items than a less comprehensive one.
- Linked to the aforementioned is the construct validity of the questionnaire. Include only items which relate to the behaviour you want to influence. For example, if returning customers constitute your target, items on the location of the hotel may be irrelevant

as this is not something you can control. Considering this aspect may shorten the length of the questionnaire.

- Criterion variables should also be included so as to statistically test whether the items of the questionnaire relate to the behaviour it intends to predict. This will result in additional items being included in the questionnaire.

It is further suggested, that future research should focus on online questionnaires, programmed in such a manner as to accurately detect when those who abandon questionnaires do so. This will provide first hand data on when respondents start to disengage completely.

Limitation of the research

The central limitation in this study was the number of assumptions made in order to proceed with the research. The assumptions were grounded in intuition which was necessary due to incomplete knowledge or information. While this may be problematic, it was done in order to allow the investigation to proceed. Perhaps the most important of these assumptions was the supposition that data collected within the context of general psychometric enquiry may be transferred to the tourism setting. Arguing that both groups have limited interest in the outcomes of the assessment and that this makes them comparable may be an oversimplification of the matter. Future researchers should avoid such generalisations.

Reference List

Ackerman, P. L. & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied,* 15(2), 163–181. http://dx.doi.org/10.1037/a0015719

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anastasi, A. & Urbina, S. (1997). *Psychological testing* (7th edn.). Upper Saddle River, NJ: Prentice-Hall.

Cohen, R. J., Swerdlik, M. E. & Sturman, E. D. (2013). *Psychological testing and assessment: An introduction to test and measurement* (8th edn.). New York, NY: McGraw-Hill.

Cronbach, L. J. & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedure. *Educational and Psychological* Measurement, 64(3), 391–418.

DeVellis, R. F. (2012). *Scale development: Theory and applications* (3rd edn.). Thousand Oaks, CA: Sage.

Fitzpatrick, A. R. & Yen, W. M. (2010). The effects of test length and sample size on the reliability and equating of tests composed of constructed-response items. *Applied Measurement in Education*, 14(1), 31–57. http://doi/abs/10.1207/S15324818AME1401_04

Gregory, R. J. (2011). *Psychological testing: History, principles, and applications* (6th edn.). Boston, MA: Pearson.

Kaiser, H. F. & Michael, W. B. (1975). Domain validity and generalizability. *Educational and Psychological Measurement*, 35, 31–35.

Kanfer, R. (2011). Determinants and consequences of subjective cognitive fatigue. In P. L. Ackerman, *Cognitive fatigue: Multidisciplinary perspectives on current research and future applications* (pp. 189–207). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/12343-009

Lord, F. & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lunz, M. E. (2009). *Test length and test reliability for multiple choice examinations. measurement research associates test insights.* https://www.rasch.org/mra/mra-02-09.htm

Moerdyk, A. (2015). *The principles and practice of psychological assessment* (2nd edn.). Pretoria: Van Schaik.

Novick, M. R. & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, 32(1), 1–13.

Nunnally, J. C. (1978). *Psychometric theory* (2nd edn.). New York, NY: McGraw-Hill.

Nunnally, J. C. & Bernstein, I. H. (1994). Psychometric theory (3rd edn.). New York, NY: McGraw-Hill.

Shaughnessy, J. J., Zechmeister, E. B. & Zechmeister, J. S. (2009). *Research methods in psychology* (8th edn.). New York, NY: McGraw-Hill.

Stanley, J. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (2nd edn.), Washington, DC: American Council on Education, 356-442.

Tredoux, C. & Durrheim, K. (2013). *Numbers, hypotheses & conclusions: A course in statistics for the social sciences* (2nd edn.). Cape Town: Juta.

Tulsky, D. S. & Zhu, J. (2000). Could test length or order affect scores on letter number sequencing of the WAIS-III and WMS-III? Ruling out effects of fatigue. *The Clinical Neuropsychologist*, 14(4), 474–478.

Wagner-Menghin, M. M. & Masters, G. N. (2013). Adaptive testing for psychological assessment: How many items are enough to run an adaptive testing algorithm? *Journal of Applied Measurement*, 14(2),106-117.

Wells, C. S. & Wollack, J. A. (2003). *An instructor's guide to understanding test reliability.* Madison, WI: University of Wisconsin's Testing & Evaluation Services. https://testing.wisc.edu/Reliability.pdf

Wright, B. D. (1992). What is the "right" test length? *Rasch Measurement Transactions,* 6(1), 205. https://www.rasch.org/rmt/rmt61g.htm [Accessed 11 April 2017]