Development of Beyond-Compliance Behaviour Questionnaire: How to Address Unexpected Results

Beta Romadiyanti
Doctoral Leadership and Policy Innovation Program, Graduate School, Universitas Gadjah Mada, Indonesia
Wahyudi Kumorotomo (PhD)
Department of Social and Political Sciences, Universitas Gadjah Mada, Indonesia
Sumaryono (PhD)
Department of Psychology, Universitas Gadjah Mada, Indonesia
Wakhid Slamet Ciptono (PhD)
Department of Economics and Business, Universitas Gadjah Mada, Indonesia

Beyond compliance behaviour is individuals behavior which compliant in implementing policies and do extra effort to ensure policies can be appropriately implemented. This research was conducted to develop the Beyond-Compliance Behaviour Questionnaire because measurement tools are needed to complement previous studies. This study consists of several stages; starting with developing questions from the literature review and operationalisation results then second study was a piloting of the questionnaire that took a minimum of 40 respondents in each trial stage using convenience sampling and voluntary filling. The piloting stage was conducted at procurement policy institutions in Indonesia, and after the validity and reliability analysis, the second study was carried out. The second study was distributed through the social media of the public procurement community. The respondents obtained in this study were 58 at the piloting stage and 75 at the second stage. Construct validity and reliability analyses were carried out using SEM PLS *(Structural Equation Modelling Partial Least Square*) and item reliability and scale validity using the Rasch Model. The results showed an improved questionnaire related to construct validity value of all dimensions. Average variance extracted (AVE) value of more than 0.5 shows convergent validity and heterotrait-monotrait (HTMT) below 0.90 show divergent validity. Composite reliability (CR) value of a new instrument is 0.95 which shows good internal consistency. The Cronbach Alpha of the improvement

questionnaire's item is 0.96 with separation at about 4.93 show that the instrument could be applied to other samples with a high level of steadiness. The evaluation of 6-point likert scale indicate observed average and andrich thresholds increase monotonically from category 1 to 6, show the scale validity.

Keywords: Beyond-Compliance Behaviour, Questionnaire Development, Validity, Reliability, Scale.

*Correspondence concerning this article should be addressed to: Beta Romadiyanti, MSc, University of Gadjah Mada, Indonesia, Email: betaroma70@gmail.com.

Introduction

There have been many studies focusing on behaviour related to policy implementation. Much of the previous research focused on the compliant behaviour of individuals towards regulations or policies. (Nobbie & Brudney, 2003; Kim & Oh, 2015; Flynn, 2018; Anthony et al., 2019; Jensen, 2020). However, the obstacles and challenges in policy implementation require beyond-compliance behaviour (Wang et al., 2021; Yoong et al., 2021; Nizigiyimana et al., 2022; Romadiyanti et al., 2024;). Beyond compliance behaviour is individuals behavior which compliant in implementing policies and do extra effort to ensure policies can be appropriately implemented (Romadiyanti et al., 2024). Using homegrown goods/services in public procurement is one of the policies that has challenges in its implementation (Wells & Hawkins, 2010; Esteves & Barclay, 2011; Kazzazi & Nouri, 2012; Ovadia, 2012; Collins, 2018; Hansen, 2020; Kalyuzhnova et al., 2022).

Beyond-compliance behaviour exceeds the role requested in implementing the policy (Romadiyanti et al., 2024). Based on the literature, there are two dimensions of beyond compliance behaviour: compliance and extra-role behaviour (Romadiyanti et al., 2024). Reliable measurement tools are needed to measure these dimensions (Yang et al., 2004) and provide evidence for two distinct dimensions.

One way to measure behaviour is through self-reporting using a valid and reliable questionnaire (George et al., 2006). Self-reporting is favoured by researchers (George et al., 2006; McDonald, 2008). Nevertheless, one must ascertain accurate measurement and extent of validity of the construct (Lajunen & Summala, 2003; McDonald, 2008).

Operationalization of behavioural dimensions of beyond-compliance into measuring tools are needed (B. Yang et al., 2004). Since the behaviour of a person differs from one place to another, it is not right to pin any single particular behaviour on someone as their own. Human nature may be distinguished by differing behaviour across contexts such as a unique or complex system (Colquhoun et al., 2017; R. Davis et al., 2015). Based on a study conducted by Romadiyanti et al. (2024), beyond-compliance behaviour is necessary to face policy challenges. A measurement tool need to be developed related to beyond-compliance behaviour. This study seeks to develop a questionnaire to measure beyond-compliance behaviour as a continuation of the concept that has been developed by Romadiyanti et al. (2024). This research was conducted in the context of Indonesia's homegrown or domestic goods/services utilization policy for public procurement.

Measurement is very important in a study. Phenomena that occur need to be quantified, especially variables related to behaviour (Brown & Room, 2021). Romadiyanti et al.'s research (2024) introduced a new concept of behaviour in policy implementation that exceeds its proper role, namely beyond compliance behaviour. Objectivity and standardisation of measurement are needed so that research related to these behaviours can be carried out quantitatively (Creswell & Creswell, 2018). This research is a follow-up research to compile and develop a beyond compliance behaviour questionnaire from a previously developed concept (Romadiyanti et al., 2024).

Method

This study was designed to development of a questionnaire for Beyond-Compliance Behaviour to address unexpected results as well as determine their psychometric properties, through the process of validity and reliability analysis

The first part of this research, which consisted in the literature review and dimension operationalisation was undertaken to develop indicators for beyond-compliance behaviour as well separate the dimensions which were amalgamated into a preliminary questionnaire. A piloting study has been prepared to test the initial questionnaire. Improvements were made based on the piloting, and a second data

collection was conducted for further questionnaire testing. Details of each stage are reported in the research results. This research started with the preparation of the operationalised questionnaire. Previous studies have carried out Literature review and operationalisation (Romadiyanti et al., 2024).

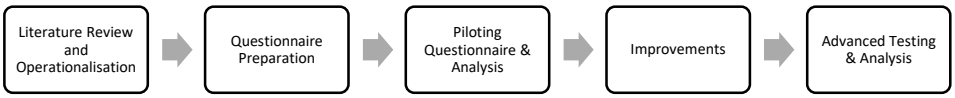

Figure 1. Flow chart depicting the stages followed for the questionnaire development.

Initial Questionnaire

The initial questionnaire in this study was developed based on the literature review and operationalisation presented in previous research by Romadiyanti et al. (2024). The compliance dimension is a behaviour that reflects approval or rejection of policy implementation (Christine & Nielsen, 2017; Meyer, 2021; Romadiyanti et al., 2024). Meanwhile, the extra-role behaviour dimension is creative behaviour through plans, strategies, and voluntary input in implementing policies (Hsu et al., 2015; Srivastava & Dhar, 2019; S. S. Kim, 2020; Romadiyanti et al., 2024). These definitions are then translated as indicators of each dimension and operationalised through interviews (Romadiyanti et al., 2024). The operationalisation results are then presented as questions in the initial questionnaire presented in Table 1.

Table 1

Initial Questionnaire with 14 items

Dimensions	Operationalisation Results	Question	Code
Compliance	Use the homegrown goods/services, following regulations/policies.	I always use homegrown goods/services following applicable regulations.	Q1
	Fulfilment organisation needs with the use of homegrown goods/services.	I always fulfil the agency needs of goods/services with domestic products.	Q2
	Ensure fulfilment of domestic product requirements.	I always ensure the fulfilment of domestic product requirements in public procurement.	Q3
	Considering homegrown products/services in public procurement.	I always consider the use of homegrown goods/services in public procurement.	Q4

	Prioritise for using homegrown goods/services in public procurement.	I always prioritise for using homegrown goods/services in public procurement without being influenced by other aspects.	Q5
Extra Role Behaviour	Interaction and cooperation with relevant parties.	I always communicate and coordinate with relevant parties to ensure the suitability of the use of homegrown goods/services in public procurement.	Q6
	Quality consideration when selecting homegrown goods/services in public procurement.	I always assess the quality of goods/services in selecting homegrown goods/services in public procurement.	Q7
		I always assess the sales service of goods/services in selecting homegrown goods/services in public procurement.	Q8
	Encourage users to use homegrown goods/services.	I always encourage users to use domestic products in public procurement.	Q9
	Proactive towards business actors.	I have always been proactive in encouraging domestic business actors to participate in public procurement.	Q10
	Efforts to update information and/or regulations.	I always update information and/or regulations related to policy implementation.	Q11
	Provide input/suggestions to the organisation, superiors or *users.*	I always provide input/suggestions to the organisation to improve the implementation of using homegrown goods/services in public procurement.	Q12
		I always provide input/suggestions to superiors to improve the policy implementation on using homegrown goods/services in public procurement.	Q13
		I always provide input/suggestions to Users to improve the policy implementation on using homegrown goods/services in public procurement.	Q14

Participants and Sampling

This study captured respondent data at the piloting and follow-up stages, where the two groups differed. Each trial stage took a minimum of 40 respondents (Hertzog, 2008). The respondents must be involve in procurement minimum two years to ensure sufficient experience and behaviours. Criteria limitations were placed on the research questionnaire so that respondents who did not meet the criteria could not fill out the

questionnaire. The piloting stage was conducted at procurement policy institutions in Indonesia, considering that in addition to filling out the questionnaire, they could also provide input on the suitability of the questionnaire items tested. Both piloting and second stages use convenience sampling or voluntary filling. The second data collection stage was blasting questionnaires on procurement actors' social media groups. The respondents obtained in this study were 58 at the piloting stage and 75 at the second stage.

Procedure

This research is piloting, analysing, and improving the questionnaire according to Figure 1. After the questionnaire had been developed, it was tested on the first research sample. Furthermore, a series of analyses was conducted to determine construct validity and reliability, item reliability, and scale validity. If the results of the analysis were not as expected, a series of analyses was carried out to improve the questionnaire. The improved questionnaire was then tested on the second sample, and the same analyses were conducted.

Measure and Data Analysis

The study used a seven-point semantic differential evaluation scale on the initial questionnaire ("Never"-"Always"). Data analysis in this study proved the validity and estimated the reliability of the question items. This study analyses construct validity and reliability, item reliability, and scale validity.

The construct validity investigated in this study was convergent and discriminant validity. Convergent validity measures how well two tests measure the same thing or how sound indicators converge with their respective constructs. (Carlson & Herdman, 2016; Cheung et al., 2023). Discriminant validity indicates the empirical difference between a construct and another (Shaffer et al., 2016). Construct Reliability shows the internal consistency of indicators in measurement and the impact of scale error on the construct (Hair et al., 2019; Raykov & Grayson, 2003). Construct validity and reliability in this study were investigated by Confirmatory Factor Analysis (CFA) with SEM PLS (Structural Equation Modelling Partial Least Square). The application used in this analysis is Smart PLS 3.0. Convergent and discriminant validity were investigated using the average variance extracted (AVE) method (Garson, 2013).

Item reliability and scale validity in this study were analysed using the Rasch Model. Item reliability is an internal consistency test commonly used in item response theory for binary items (Garson, 2013). Meanwhile, The validity tested is the rating scale validity. Rating scale validity is an analysis to verify whether the scale choices confuse respondents (Sumintono & Widhiarso, 2014). The Rasch model analysis application used in this study is Winstep Rasch 5.4.1.

Results

This study presents the testing process sequentially, from piloting data analysis, questionnaire refinement, and further testing results. Measurement standards are explained in the piloting research explanation. Follow-up questionnaire improvements are expected to provide insight for other researchers who face similar problems in their research.

Construct Validity and Reliability

Construct validity indicates that a construct's assessment score reflects the construct's level, so construct validity determines whether or not a study is valid. (Flake et al., 2022). Construct validity consists of two types, namely convergent and discriminant validity. (Hair et al., 2019). Convergent validity explains the unity of the construct with its items (Carlson & Herdman, 2016; Cheung et al., 2023; Hair et al., 2019). In addition, discriminant validity is an empirical demonstration of how the actual difference between a construct and other constructs can explain (Hair et al., 2019; Shaffer et al., 2016). On the other hand, construct reliability refers to the internal consistency between indicators in a construct (Hair et al., 2019).

Convergent Validity

It can be seen from the AVE (average variance extracted) value of each construct whether we have reached convergent validity. (Hair et al., 2019). Hair et al.(2019) and Zhang & Li (2020) suggested that the AVE >= 0.5 is considered having construct validity pledged to it, the value describes that the construct can account for at least 50 percent of item variance. The AVE value of beyond-compliance behaviour 0.66, compliance dimension 0.78, and extra-role behaviour 0.69. The AVE value is higher than 0.5 for all constructs in the original scale, so sall constructs in this study are valid.

Discriminant Validity

Heterotrait-monotrait (HTMT) can be used for discriminant validity testing (Friman et al., 2019; Hair et al., 2019; Henseler et al., 2015). It can accommodate constructs that are conceptually similar (Hair et al., 2019; Henseler et al., 2015). The recommended HTMT value for conceptually similar constructs is below 0.90, while constructs conceptually completely different are below 0.85. (Henseler et al., 2015). Henseler and colleagues (2015) have proposed the most lenient criteria for assessing discriminant validity, even when the HTMT value is almost perfect (close to 1.0). However, a high HTMT value only sometimes implies low discriminant validity. Homogeneous loadings or large sample sizes can be a consideration in the case of HTMT values close to 1.0 (Henseler et al., 2015). Compliance and Extra Role Behaviour are similar constructs, so the Threshold value used is 0.90.

The results of the HTMT analysis on the initial questionnaire are in Table 2. In Table 2, the HTMT value between the Compliance and Extrarole Behaviour Dimensions is a bolded number, showing a value of 0.85. Both values are below the conceptually similar construct threshold of 0.90 (Henseler et al., 2015). Interestingly, the assessment results with HTMT on beyond compliance with compliance and extra-role behaviour have HTMT values greater than 0.90. This result is related to the repeated indicator approach used to test the construct validity and reliability in SEM-PLS analyses (Sarstedt et al., 2019). The indicators used in Beyond-Compliance Behaviour are repetitions of compliance and extra-role behaviour indicators, so high values on HTMT sub-indicators and indicators are predicted. (Sarstedt et al., 2019).

Table 2

Discriminant Validity Assessment using HTMT Criterion

	Beyond - Compliance Behaviour	Compliance	Extra-role Behaviour
Beyond Compliance Behaviour			
Compliance	0.97		
Extra-role Behaviour	1.02	**0.85**	

Construct Reliability

Construct reliability can be determined based on the composite reliability (CR) value, which shows internal consistency. (Hair et al., 2019;

Sujati et al., 2020). Cronbach alpha is another indicator of internal consistency but it's value will always be smaller than CR value (Hair et al., 2019). Reliability must have CR greater than 0.70 (Sujati et al., 2020). On the other hand, CR values beyond 0.95 can also be problematic, indicating that associations are too high and reducing construct validity to a certain extent as well (Hair et al., 2019). For exploratory research CR value from 0.6 to 0.7 is still acceptable (Hair et al., 2019).

In the initial questionnaire, CR value of beyond-compliance behaviour was 0.96. The constructs of compliance and extra-role behaviour exhibited a CR value of 0.95 in the initial questionnaire set. This indicates that the general construct satisfies at least the required minimum CR. The CR value of the beyond-compliance construct is too high at above 0.95, and this might be a threat to its validity.

Item Reliability and Scale Validity

The Rasch model is a popular method to measure psychometric reliability and validity (Aryadoust et al., 2021). According to the Rasch model, people and items predict an answer on a measuring instrument (Quintão et al., 2013). Rasch's model is not fixated on items alone, but the calculations and analyses also consider human factors (Aryadoust et al., 2021; Quintão et al., 2013). This characteristic is the advantage of the Rasch model over other models.

In Rasch measurement, the validity of a construct can be indicated by unidimensionality, which can also be performed on multi-dimensional variables (Aryadoust et al., 2021; Briggs & Wilson, 2003; Quintão et al., 2013). However, the Rasch model research in this study focused on investigating item reliability and rating scale validity.

Item Reliability

Estimates of model stability in new samples in Rasch modelling are indicated by item separation and reliability estimates. (Bond et al., 2021; Van Zile-Tamsen, 2017). Reliability values below 0.6 are not acceptable, 0.6-0.8 are less acceptable, and more than 0.8 have high reliability (Rahayah Ariffin et al., 2010; Sumintono & Widhiarso, 2014). Item separation index is an estimate of the separation of items in the variable being measured (Bond et al., 2021). The higher the item reliability and separation index value, the more convincing the items will be replicated (Bond et al., 2021). An item separation index of more than 3.0

and reliability of more than 0.90 is estimated to have good stability when measured on other samples. (Van Zile-Tamsen, 2017).

The initial questionnaire's item reliability index was 0.16, with a separation of 0.44. The initial questionnaire has unacceptable item reliability; the value is minimal compared to the separation index. This result indicates that there is a problem with the questionnaire items. Improvements to the questionnaire must be made to make the item reliability acceptable.

Rating Scale Validity

The validity of the rating scale in this study is an assessment of the relevance of the categories used in the rating scale (Andrich, 2011). The Rasch Model's category relevance assessment uses the Andrich Rating Scale Model (Andrich, 2011; Van Zile-Tamsen, 2017). Andrich (2011) underlines that, in general, the thresholds of the categories should be correct, ordered and show a consistent structure. Rating scale categories are appropriate if the Andrich threshold increases as the category level increases. In addition, the assessment of the rating scale can also be seen from the observed average in the analysis of the rating scale in the Rasch model (Linacre, 2002; Van Der Wal et al., 2012). The observed average is an empirical indicator of the category and is expected to increase monotonically as the category increases (Linacre, 2002).

The initial questionnaire used a seven-point semantic differential evaluation scale ("Never" - "Always"). The results of the scale evaluation in the initial questionnaire are presented in Table 3. The Observed Average shows that the value of Category 2 is greater than Category 3. At the same time, the Andrich Threshold value shows that the value of Category 4 is higher than Category 5. This result shows that the scale used in the initial questionnaire still needs to be clarified for respondents.

Table 3

Evaluation of Initial Questionnaire Rating Scale

	1	2	3	4	5	6	7
Observed Average	-0.92	1.87	0.02	0.86	1.21	2.07	3.03
Andrich Threshold	NONE	-1.86	-0.33	0.1	0.04	0.93	2.35

Improvement Questionnaire: What Improved?

Improvements that can be made in the piloting questionnaire include evaluating the scale's ability and improving the wording to delete items (Kishore et al., 2021). The modifications of the questionnaire were based on the validity and reliability analyses.

The construct validity and reliability of the first questionnaire were good and the visible problems were item reliability and scale validity. Results of the Rasch model analysis underpinning improvement of existing items. The Rasch model can perform item analyses such as item fit/misfit and polarity (Bond et al., 2021; Rahayah Ariffin et al., 2010; Van Zile-Tamsen, 2017).

Item polarity is used to analyse the relationship between items and constructs (Bond et al., 2021; Othman et al., 2014; Rahayah Ariffin et al., 2010). The point-measure correlation coefficient (PTMEA Corr) is the value used as a reference of item polarity (Bond et al., 2021; Rahayah Ariffin et al., 2010). This study shows a relatively strong correlation with PTMEA Corr values of 0.71 to 0.82 (Othman et al., 2014). These results indicate that the item polarity of the initial questionnaire is as expected; no items are opposite to the construct (Bond et al., 2021; Rahayah Ariffin et al., 2010).

Item fit is evaluating how suitable an item for measuring variable (Bond et al., 2021; Rahayah Ariffin et al., 2010). The results of this study indicate that the items in Table 1 that have Infit and Outfit values outside the range of 0.6 -1.4 are Q8, Q10, Q11, and Q13; the item can be eliminated (Bond et al., 2021; Rahayah Ariffin et al., 2010).

Furthermore, the research questionnaire was improved; no elimination of the questionnaire was carried out. The improvements made were rewording and clarifying the measurement scale. The sentence structure is as straightforward as possible to avoid confusing the respondent. Item wording has a significant role in the reliability of items and the effectiveness of questionnaire completion (Eys et al., 2007; van Sonderen et al., 2013). The first change made is to clarify the period observed; in this study, the observed time is two (2) years; this considers the latest policy issued in Indonesia regarding the use of domestic products. Furthermore, the second change is the removal of the word "always", which is considered to lead respondents’s answer.

In addition to rewording, scale improvements were also made in the revised questionnaire. The previous measurement scale consisted of a seven-point semantic evaluation scale from "never" to "always". Measurement of behaviour by self-report requires frequency-specific scale measures (Fishbein & Ajzen, 2011). The measurement scale was then changed to be more specific, to a Likert 6-point scale with answer options: Never (0%), Rarely (20%-39%), Sometimes (40%-59%), Often (60%-79%), Almost Always (80%-99%) and Always (100%).

Second Study (Improvement Questionnaire)

After improvement, validity and reliability analyses were conducted again. As in the initial questionnaire, convergent validity, discriminant validity, construct reliability, item reliability and scale validity analyses were conducted. The average variant extracted (AVE) value on the beyond-compliance construct is 0.60, compliance is 0.74 and extra-role behaviour is 0.62; all constructs have the minimum AVE value to show convergent construct validity (Hair et al., 2019; Zhang & Li, 2020). Furthermore, the heterotrait-monotrait (HTMT) value between the compliance and extra-role behaviour variables is 0.87, which meets the discriminant validity requirement of below 0.90 (Hair et al., 2019; Henseler et al., 2015). The CR value of the improvement questionnaire is 0.95. The results of the overall test of validity and reliability showed that, despite changing items, the constructs were conceptually valid and reliable (Hair et al., 2019).

The reliability index for the questionnaire items was retested after improvements, and it became a Cronbach alpha of 0.96 with separation at about 4.93. The reliability index and item separation of the revised questionnaire surpassed those of original one, suggesting that this instrument could also be applied to other samples with a high level of steadiness (Van Zile-Tamsen, 2017). Furthermore, infit value of the items in improvement questionnaire were between 0.78 and 1.40 whereas outfit ranged from between 0.67 and 1.27, meaning that basically each item of the improvement questionnaire suitable to measure the variable. The robust point-measure correlation coefficient (PTMEA Corr) that ranged from 0.69 - 0.72 and have positive values show that the items fulfil the polarity requirements (Othman et al., 2014).

The results of the evaluation using a 6-point likert scale were also very good. The observed average increase monotonically from category 1

to 6. Monotonically increasing Andrich thresholds were observed, varying from − 2.54 to 2.22. All of these results show that the scale used is clear to respondents.

Discussion

A methodological and systematic procedure must be followed to ensure a good quality questionnaire (Kishore et al., 2021). It is crucial to conduct a pilot study in order to evaluate the feasibility and adequacy of an instrument that has been developed (Hertzog, 2008). Pilot studies also provide an opportunity to improve items in the questionnaire (Kishore et al., 2021). The proposed constructs and items in the questionnaire used should be tested. This study focuses on investigating construct validity and reliability, item reliability and scale validity by conducting a pilot study and further testing after refinement.

The pilot questionnaires exhibited good results for construct validity and reliability. Despite that internal consistency in the first questionnaire was high and above 0.95, construct validity was met as well. This study did not see the potential reduction of construct validity due to excessive internal consistency (Hair et al., 2019). Discriminant validity between the compliance and extra-role behaviour constructs also met the HTMT requirement of below 0.90 (Hair et al., 2019; Henseler et al., 2015). Although there is a violation in the discriminant validity of compliance and extra-role behaviour against beyond-compliance behaviour, this is predictable due to the repeated indicator approach in the relationship of constructs and sub-constructs (Sarstedt et al., 2019).

The initial questionnaire was improved because shallow item reliability values were discovered. The meagre reliability value prompted further analysis of item fit and polarity. Many studies have looked at the quality of measurement items using the Rasch model, but most of them tested ready-made questionnaires with good reliability (Briggs & Wilson, 2003; Ip et al., 2018; Quaigrain & Arhin, 2017; Rahayah Ariffin et al., 2010; Van Zile-Tamsen, 2017; W. C. Wang et al., 2006). This research presents findings that are different from what is expected and makes improvements once the expected item analysis results.

This research also indirectly shows how the characteristics of behaviour are particular. Behaviour itself is multi-dimensional and flexible so that the analysis can be adjusted to the characteristics of the behaviour under study (Meneses & Palacio, 2016; Flynn, 2018; Gkargkavouzi et al., 2019; S. Yang et al., 2019; Lambe & Craig, 2020). This research shows how the concept of beyond-compliance behaviour needs to be

operationalised first to fit the context under study. Operationalisation also facilitates the preparation of questions in the questionnaire. Operationalisation is very important, this is also reinforced by the fact that the challenges and issues that occur in the implementation of the use of local products in public procurement are very varied (Wells & Hawkins, 2010; Esteves & Barclay, 2011; Ovadia, 2012; Kazzazi & Nouri, 2012; Collins, 2018; Hansen, 2020; Kalyuzhnova et al., 2022; Romadiyanti et al., 2024).

The development of the measurement scale in this study also reinforces the theory of behavioural scales that have been developed (Fishbein & Ajzen, 2011). Many previous studies have used 4, 5, 6 and 7-point Likert scales (De Jong & Den Hartog, 2010; Lambriex-Schmitz et al., 2020; Messmann & Mulder, 2012). However, specific frequency is emphasised for measuring behaviour (Fishbein & Ajzen, 2011). This study added a percentage frequency to clarify the frequency boundaries of the behaviour further.

The samples in the piloting and follow-up studies in this study are different. However, both samples fulfil the minimum requirements for piloting questionnaires and testing through SEM-PLS (Hertzog, 2008; Kishore et al., 2021; Kock, 2018; Sujati et al., 2020). Although testing using SEM-PLS generally requires a large sample size, a pilot study does not require a large sample size (Kock, 2018) and samples as low as 30 are enough for piloting (Hertzog, 2008; Kishore et al., 2021; Kock, 2018; Sujati et al., 2020).

This research has limitations due to the small number of samples. In addition, this study did not discuss and analyse all criteria in the SEM-PLS or Rasch Model. In SEM-PLS, for example, structural models, factor loading, and collinearity are not discussed. (Hair et al., 2019). Meanwhile, in the Rasch model, this study did not discuss person measures and dimensionality (Bond et al., 2021). The SEM-PLS test represents the relationship between constructs and the internal consistency of constructs in this study. At the same time, the Rasch Model is used to investigate the quality of items and scales used. The SEM-PLS and Rasch Model can be combined and complement each other (Bond et al., 2021).

Conclusion

This study describes the questionnaire development process, from developing operationalised interview questions to testing the initial questionnaire, refining the questionnaire, and conducting further testing. This research shows how steps should be taken if the piloting results find results that differ from what is expected. The paper outlines the

questionnaire phrasing and measurement modifications, achieving the expected item reliability.

The research also illustrates the importance of operationalising the proposed behavioural constructs in the context of policies on using domestic products in public procurement. It is crucial to operationalise behaviour in a particular context. This research also illustrates how to develop a questionnaire from the results of operationalising variables.

This study also illustrates the importance of scale clarity in behavioural measurement. Its results are expected to provide a complete picture for other researchers who will develop behavioural questionnaires in various contexts. The improved questionnaire in this study can also be replicated in other relevant studies. The improved questionnaire's test results have shown construct validity and reliability, item reliability, and scale fit.

This study used a very small sample size. Research with a larger sample size and involving several sample groups may be needed in the future. Research with other policy contexts can also be developed to enrich and strengthen the concept of beyond compliance behaviour further. The results of this study can be used as a reference for developing beyond compliance behaviour questionnaires in different policy contexts and as an illustration for developing other behaviour questionnaires.

References

Andrich, D. (2011). Rating scales and Rasch measurement. *Expert Review of Pharmacoeconomics and Outcomes Research*, *11*(5), 571–585. https://doi.org/10.1586/erp.11.59

Anthony, J., Goldman, R., Rees, V. W., Frounfelker, R. L., Davine, J., Keske, R. R., Brooks, D. R., & Geller, A. C. (2019). Qualitative Assessment of Smoke-Free Policy Implementation in Low-Income Housing: Enhancing Resident Compliance. *American Journal of Health Promotion*, *33*(1), 107–117. https://doi.org/10.1177/0890117118776090

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, *38*(1), 6–40. https://doi.org/10.1177/0265532220927487

Bond, T. G., Yan, Z., & Heene, M. (2021). *Applying the Rasch Model: Fundamental Measurement in The Human Sciences* (4th ed.). Routledge. https://doi.org/10.4324/9781410614575

Brown, F. D., & Room, H. (2021). Scale development: Theory and applications.

Briggs, D. C., & Wilson, M. (2003). An Introduction to Multidimensional Measurement using Rasch Models. *Journal of Applied Measurement*, *4*(1), 87–100.

Carlson, K. D., & Herdman, A. O. (2016). Understanding the Impact of Convergent Validity on Research Results. *Organizational Research Methods*, *000*(00), 1–16. https://doi.org/10.1177/1094428110392383

Cheung, G. W., Cooper-Thomas, H. D., Lau, R. S., & Wang, L. C. (2023). Reporting reliability, convergent and discriminant validity with structural equation modeling: A review and best-practice recommendations. In *Asia Pacific Journal of Management* (Issue 0123456789). Springer US. https://doi.org/10.1007/s10490-023-09871-y

Christine, P., & Nielsen, V. L. (2017). Compliance: 14 questions. *Regulatory Theory: Foundations and Applications*, 217–232. https://library.oapen.org/bitstream/handle/20.500.12657/31596/626829.pdf?sequen#page=253

Collins, D. (2018). Government procurement with strings attached: The uneven control of offsets by the world trade organization and regional trade agreements. In *Asian Journal of International Law* (Vol. 8, Issue 2, pp. 301–321). Cambridge University Press. https://doi.org/10.1017/S2044251316000278

Colquhoun, H. L., Squires, J. E., Kolehmainen, N., Fraser, C., & Grimshaw, J. M. (2017). Methods for designing interventions to change healthcare professionals' behaviour: A systematic review. *Implementation Science*, *12*(1), 1–12. https://doi.org/10.1186/s13012-017-0560-5

Creswell, J. W., & Creswell, J. D. (2018). Research Design: Qualitative, Quantitative, and Mixed Methods Approaches (5th ed.). SAGE Publications, Inc.

Davis, K. (2016). A method to measure success dimensions relating to individual stakeholder groups. *International Journal of Project Management*, *34*(3), 480–493. https://doi.org/10.1016/j.ijproman.2015.12.009

Davis, R., Campbell, R., Hildon, Z., Hobbs, L., & Michie, S. (2015). Theories of behaviour and behaviour change across the social and behavioural sciences: a scoping review. *Health Psychology Review*, *9*(3), 323–344. https://doi.org/10.1080/17437199.2014.941722

De Jong, J., & Den Hartog, D. (2010). Measuring innovative work behaviour. *Creativity and Innovation Management*, *19*(1), 23–36. https://doi.org/10.1111/j.1467-8691.2010.00547.x

Esteves, A. M., & Barclay, M. A. (2011). Enhancing the benefits of local content: Integrating social and economic impact assessment into procurement strategies. *Impact Assessment and Project Appraisal*, *29*(3), 205–215. https://doi.org/10.3152/146155111X12959673796128

Eys, M. A., Carron, A. V., Bray, S. R., & Brawley, L. R. (2007). Item wording and internal consistency of a measure of cohesion: The group environment questionnaire. *Journal of Sport and Exercise Psychology*, *29*(3), 395–402. https://doi.org/10.1123/jsep.29.3.395

Fishbein, M., & Ajzen, I. (2011). Predicting and changing behavior: The reasoned action approach. In *Predicting and Changing Behavior: The Reasoned*

Action Approach. Psychology Press, Taylor & Francis Group. https://doi.org/10.4324/9780203838020/PREDICTING-CHANGING-BEHAVIOR-MARTIN-FISHBEIN-ICEK-AJZEN

Flake, J. K., Davidson, I. J., Wong, O., & Pek, J. (2022). Construct Validity and the Validity of Replication Studies: A Systematic Review. *American Psychologist*, *77*(4), 576–588. https://doi.org/10.1037/amp0001006

Flynn, A. (2018). Investigating the implementation of SME-friendly policy in public procurement. *Policy Studies*, *39*(4), 422–443. https://doi.org/10.1080/01442872.2018.1478406

Friman, M., Westman, J., & Olsson, L. E. (2019). Children's Life Satisfaction and Satisfaction with School Travel. *Child Indicators Research*, *12*(4), 1319–1332. https://doi.org/10.1007/s12187-018-9584-x

García-Fernández, M. (2015). How to measure knowledge management: Dimensions and model. *Vine*, *45*(1), 107–125. https://doi.org/10.1108/VINE-10-2013-0063

Garson, G. D. (2013). *Validity & Reliability*. Statistical Publishing Associates.

George, J., Mackinnon, A., Kong, D. C. M., & Stewart, K. (2006). Development and validation of the Beliefs and Behaviour Questionnaire (BBQ). *Patient Education and Counseling*, *64*(1–3), 50–60. https://doi.org/10.1016/j.pec.2005.11.010

Gkargkavouzi, A., Halkos, G., & Matsiori, S. (2019). A Multi-dimensional Measure of Environmental Behavior: Exploring the Predictive Power of Connectedness to Nature, Ecological Worldview and Environmental Concern. *Social Indicators Research*, *143*(2), 859–879. https://doi.org/10.1007/S11205-018-1999-8/FIGURES/1

Hair, J. F., Risher, J. J., Sarstedt, M., & Ringle, C. M. (2019). When to use and how to report the results of PLS-SEM. *European Business Review*, *31*(1), 2–24. https://doi.org/10.1108/EBR-11-2018-0203

Hansen, M. W. (2020). Toward a strategic management perspective on local content in African extractives: MNC procurement strategies between local responsiveness and global integration. *Africa Journal of Management*, *6*(1), 24–42. https://doi.org/10.1080/23322373.2020.1717283

Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, *43*(1), 115–135. https://doi.org/10.1007/s11747-014-0403-8

Hertzog, M. A. (2008). Considerations in Determining Sample Size for Pilot Studies. *Research in Nursing & Health*, *31*(2), 180–191. https://doi.org/10.1002/nur

Hsu, J. S.-C., Shih, S.-P., Hung, Y. W., & Lowry, P. B. (2015). The Role of Extra-Role Behaviors and Social Controls in Information Security Policy Effectiveness. *Information Systems Research*, *26*(2), 282–300. https://doi.org/10.1287/isre.2015.0569

Ip, P., Tso, W., Rao, N., Ho, F. K. W., Chan, K. L., Fu, K. W., Li, S. L., Goh, W., Wong, W. H. sang, & Chow, C. B. (2018). Rasch validation of the Chinese parent–child interaction scale (CPCIS). *World Journal of Pediatrics*, *14*(3), 238–246. https://doi.org/10.1007/s12519-018-0132-z

Jensen, J. (2020). The Current NIMS Implementation Behavior of United States Counties. In *Journal of Homeland Security and Emergency Management* (Vol. 8, Issue 1). De Gruyter. https://doi.org/10.2202/1547-7355.1815

Kalyuzhnova, Y., Azhgaliyeva, D., & Belitski, M. (2022). Public Policy Instruments for Procurement: An Empirical Analysis. *Technological Forecasting and Social Change*, *176*, 121472. https://doi.org/10.1016/J.TECHFORE.2022.121472

Kazzazi, A., & Nouri, B. (2012). A conceptual model for local content development in petroleum industry. *Management Science Letters*, *2*, 2165–2174. https://doi.org/10.5267/j.msl.2012.05.031

Kim, J., & Oh, S. S. (2015). Confidence, knowledge, and compliance with emergency evacuation. *Journal of Risk Research*, *18*(1), 111–126. https://doi.org/10.1080/13669877.2014.880728

Kim, S. S. (2020). Exploitation of shared knowledge and creative behavior: the role of social context. *JOURNAL OF KNOWLEDGE MANAGEMENT*, *24*(2), 279–300. https://doi.org/10.1108/JKM-10-2018-0611

Kishore, K., Jaswal, V., Kulkarni, V., & De, D. (2021). Practical guidelines to develop and evaluate a questionnaire. *Indian Dermatology Online Journal*, *12*(2), 266–275. https://doi.org/10.4103/idoj.IDOJ_674_20

Kock, N. (2018). Minimum sample size estimation in PLS-SEM: An application in tourism and hospitality research. *Applying Partial Least Squares in Tourism and Hospitality Research*, 1–16. https://doi.org/10.1108/978-1-78756-699-620181001

Lajunen, T., & Summala, H. (2003). Can we trust self-reports of driving? Effects of impression management on driver behaviour questionnaire responses. *Transportation Research Part F: Traffic Psychology and Behaviour*, *6*(2), 97–107. https://doi.org/10.1016/S1369-8478(03)00008-1

Lambe, L. J., & Craig, W. M. (2020). Peer defending as a multidimensional behavior: Development and validation of the Defending Behaviors Scale. *Journal of School Psychology*, *78*, 38–53. https://doi.org/10.1016/J.JSP.2019.12.001

Lambriex-Schmitz, P., Van der Klink, M. R., Beausaert, S., Bijker, M., & Segers, M. (2020). Towards successful innovations in education: Development and validation of a multi-dimensional Innovative Work Behaviour Instrument. *Vocations and Learning*, *13*(2), 313–340. https://doi.org/10.1007/s12186-020-09242-4

Linacre, J. (2002). Understanding Rasch measurement: Optimizing Rating Scale Category Effectiveness. *Journal of Applied Measurement*, *3*, 85–106.

McDonald, J. D. (2008). Measuring Personality Constructs: The Advantages and Disadvantages of Self-Reports, Informant Reports and Behavioural

Assessments. *Enquire*, *1*(1), 1–19. https://doi.org/10.1126/science.237.4815.599

Meneses, G. D., & Palacio, A. B. (2016). Recycling Behavior: A Multidimensional Approach. *Http://Dx.Doi.Org/10.1177/0013916505276742*, *37*(6), 837–860. https://doi.org/10.1177/0013916505276742

Messmann, G., & Mulder, R. H. (2012). Development of a measurement instrument for innovative work behaviour as a dynamic and context-bound construct. *Human Resource Development International*, *15*(1), 43–59. https://doi.org/10.1080/13678868.2011.646894

Meyer, M. (2021). Putting the onus on authority: A review of obedient behavior and why we should move on. *New Ideas in Psychology*, *60*. https://doi.org/10.1016/j.newideapsych.2020.100831

Nizigiyimana, A., Acharya, D., Morillon, G. F., & Poder, T. G. (2022). *Predictors of Vaccine Acceptance, Confidence, and Hesitancy in General, and COVID-19 Vaccination Refusal in the Province of Quebec, Canada*. https://doi.org/10.2147/PPA.S376103

Nobbie, P. D., & Brudney, J. L. (2003). Testing the Implementation, Board Performance, and Organizational Effectiveness of the Policy Governance Model in Nonprofit Boards of Directors. *Nonprofit and Voluntary Sector Quarterly*, *32*(4), 571–595. https://doi.org/10.1177/0899764003257460

Othman, N., Salleh, S. M., Hussin, H., & Wahid, H. Ab. (2014). Assessing Construct Validity and Reliability of Competitiveness Scale Using Rasch Model Approach. *The 2014 WEI International Academic Conference Proceedings*, 113–120.

Ovadia, J. S. (2012). The dual nature of local content in Angola's oil and gas industry: development vs. elite accumulation. *Journal of Contemporary African Studies*, *30*(3), 395–417. https://doi.org/10.1080/02589001.2012.701846

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education*, *4*(1), 1–12. https://doi.org/10.1080/2331186X.2017.1301013

Quintão, S., Delgado, A. R., & Prieto, G. (2013). Validity study of the beck anxiety inventory (Portuguese version) by the rasch rating scale model. *Psicologia: Reflexao e Critica*, *26*(2), 305–310. https://doi.org/10.1590/S0102-79722013000200010

Rahayah Ariffin, S., Omar, B., Isa, A., & Sharif, S. (2010). Validity and reliability Multiple Intelligent item using Rasch measurement model. *Procedia - Social and Behavioral Sciences*, *9*, 729–733. https://doi.org/10.1016/j.sbspro.2010.12.225

Raykov, T., & Grayson, D. (2003). A test for change of composite reliability in scale development. *Multivariate Behavioral Research*, *38*(2), 143–159. https://doi.org/10.1207/S15327906MBR3802_1

Romadiyanti, B., Kumorotomo, W., Sumaryono, & Ciptono, W. S. (2024). Beyond-Compliance Behaviour: Concept and Operationalisation in the Context of Using Domestic Product Policy in Public Procurement. *Journal of Policy Studies*, *39*(1), 17–27.

Sarstedt, M., Hair, J. F., Cheah, J. H., Becker, J. M., & Ringle, C. M. (2019). How to specify, estimate, and validate higher-order constructs in PLS-SEM. *Australasian Marketing Journal*, *27*(3), 197–211. https://doi.org/10.1016/j.ausmj.2019.05.003

Shaffer, J. A., DeGeest, D., & Li, A. (2016). Tackling the Problem of Construct Proliferation: A Guide to Assessing the Discriminant Validity of Conceptually Related Constructs. *Organizational Research Methods*, *19*(1), 80–110. https://doi.org/10.1177/1094428115598239

Srivastava, A. P., & Dhar, R. L. (2019). Authentic Leadership and Extra Role Behavior: a School Based Integrated Model. *Current Psychology*, *38*(3), 684–697. https://doi.org/10.1007/S12144-017-9634-4/TABLES/6

Sujati, H., Sajidan, Akhyar, M., & Gunarhadi. (2020). Testing the construct validity and reliability of curiosity scale using confirmatory factor analysis. *Journal of Educational and Social Research*, *10*(4), 229–237. https://doi.org/10.36941/JESR-2020-0080

Sumintono, B., & Widhiarso, W. (2014). *Aplikasi model Rasch untuk Penelitian Ilmu-Ilmu sosial*. Trim Komunikata Publishing House. http://eprints.um.edu.my/11413/

Van Der Wal, M. B. A., Tuinebreijer, W. E., Bloemen, M. C. T., Verhaegen, P. D. H. M., Middelkoop, E., & Van Zuijlen, P. P. M. (2012). Rasch analysis of the Patient and Observer Scar Assessment Scale (POSAS) in burn scars. *Quality of Life Research*, *21*(1), 13–23. https://doi.org/10.1007/s11136-011-9924-5

van Sonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of Reverse Wording of Questionnaire Items: Let's Learn from Cows in the Rain. *PLoS ONE*, *8*(7), 1–7. https://doi.org/10.1371/journal.pone.0068967

Van Zile-Tamsen, C. (2017). Using Rasch Analysis to Inform Rating Scale Development. *Research in Higher Education*, *58*(8), 922–933. https://doi.org/10.1007/sl

Wang, H., Li, J., Mangmeechai, A., Su, J., & Linking, J. (2021). Linking Perceived Policy Effectiveness and Proenvironmental Behavior: The Influence of Attitude, Implementation Intention, and Knowledge. *International Journal of Environmental Researchand Public Health*, *18*, 1–17. https://doi.org/10.3390/ijerph18062910

Wang, W. C., Yao, G., Tsai, Y. J., Wang, J. Der, & Hsieh, C. L. (2006). Validating, improving reliability, and estimating correlation of the four subscales in the WHOQOL-BREF using multidimensional rasch analysis. *Quality of Life Research*, *15*(4), 607–620. https://doi.org/10.1007/s11136-005-4365-7

Wells, J., & Hawkins, J. (2010). Increasing 'local content' in infrastructure procurement. Part 1. *Proceedings of Institution of Civil Engineers: Management, Procurement and Law*, *163*(2), 65–70. https://doi.org/10.1680/mpal.2010.163.2.65

Yang, B., Watkins, K. E., & Marsick, V. J. (2004). The construct of the learning organization: Dimensions, measurement, and validation. *Human Resource Development Quarterly*, *15*(1), 31–55. https://doi.org/10.1002/hrdq.1086

Yang, S., Su, Y., Wang, W., & Hua, K. (2019). Research on developers' green procurement behavior based on the theory of planned behavior. *Sustainability (Switzerland)*, *11*(10). https://doi.org/10.3390/su11102949

Yoong, S. L., Hall, A., Stacey, F., Nathan, N., Reilly, K., Delaney, T., Sutherland, R., Hodder, R., Straus, S., & Wolfenden, L. (2021). An exploratory analysis to identify behavior change techniques of implementation interventions associated with the implementation of healthy canteen policies. *Translational Behavioral Medicine*, *11*(8), 1606–1616. https://doi.org/10.1093/tbm/ibab036

Zhang, Y., & Li, L. (2020). Intention of Chinese college students to use carsharing: An application of the theory of planned behavior. *Transportation Research Part F: Traffic Psychology and Behaviour*, *75*, 106–119. https://doi.org/10.1016/j.trf.2020.09.021

Received October 15th, 2024

Revisions Received December 31st, 2024