

Validity and Reliability of Scales on the Implementation of Evaluation in Measuring its Antecedents and Consequences in the Malaysian Public Sector

Rafidah Mohamed Hashim, Jasmine Ahmad

To Link this Article: <http://dx.doi.org/10.6007/IJARBSS/v10-i1/6850>

DOI:10.6007/IJARBSS/v10-i1/6850

Received: 23 December 2019, **Revised:** 10 January 2020, **Accepted:** 16 January 2020

Published Online: 29 January 2020

In-Text Citation: (Hashim & Ahmad, 2020)

To Cite this Article: Hashim, R. M., & Ahmad, J. (2020). Validity and Reliability of Scales on the Implementation of Evaluation in Measuring Its Antecedents and Consequences in the Malaysian Public Sector. *International Journal of Academic Research in Business and Social Sciences*, 10(1), 264–292.

Copyright: © 2020 The Author(s)

Published by Human Resource Management Academic Research Society (www.hrmars.com)

This article is published under the Creative Commons Attribution (CC BY 4.0) license. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this license may be seen

at: <http://creativecommons.org/licenses/by/4.0/legalcode>

Vol. 10, No. 1, 2020, Pg. 264 - 292

<http://hrmars.com/index.php/pages/detail/IJARBSS>

JOURNAL HOMEPAGE

Full Terms & Conditions of access and use can be found at
<http://hrmars.com/index.php/pages/detail/publication-ethics>

Validity and Reliability of Scales on the Implementation of Evaluation in Measuring Its Antecedents and Consequences in the Malaysian Public Sector

Rafidah Mohamed Hashim, Jasmine Ahmad

Faculty of Administrative Science and Policy Studies, Universiti Teknologi MARA (UiTM), Shah Alam,
Malaysia

Abstract: This paper reports series of steps applied in the instrument building process to ensure the validity and reliability of evaluation scales developed in the study. The scales are later used in the main study to measure the implementation of evaluation on policies and programs and simultaneously measure its antecedents (evaluation capacity building (ECB) factors) and consequences (evaluation use) in the Malaysian public sector. There are eight constructs used to measure the proposed framework. Five constructs are used to measure the ECB factors, which are evaluation office (EO), internal evaluators (IE), evaluation information system (EIS), financial resources (FR), and evaluation regulatory framework (ERF). One construct is used to measure the implementation of evaluation and two constructs namely accountability and organisational learning are used to measure evaluation use. In efforts to ensure the content validity of the scales, a pre-test session with six practitioners and a content review session with three experts from the industry and academics were done prior to the pilot study commences. The pre-test session with practitioners helped to validate important constructs of the study. While the content validity session with the experts confirmed the aspects of relevance, clarity, and technical of the instrument are met. Later, a total of 50 respondents who directly involve in evaluation-related activities at selected divisions in various ministries or had previous posting experience in related divisions were chosen as pilot study samples. The reliability tests using the Statistical Package for Social Sciences Program (SPSS) version 22, revealed that the Cronbach's Alpha scores between 0.732 to 0.923 are well above the minimum set value of 0.70. Therefore, based on the feedbacks from the pre-test respondents and the content review by experts, coupled with the reliability results of the pilot test, the scales can be accepted to be valid and reliable.

Keywords: Validity, Reliability, Implementation of Evaluation, Evaluation Capacity Building (ECB), and Evaluation Use.

Introduction

Evaluating the performance of public policies and programs is considered fundamental in any policy cycle. The need for evaluation is even critical in today's era where all social arrangements, especially the ones arranged by the public organizations, are often questioned by the public. As the main study intends to measure the implementation of evaluation on policies and programs and simultaneously looking at its antecedents and consequences in the Malaysian public sector, the instruments must be first developed and tested. The development and testing of the instruments are crucial for the main study as most of the existing instruments are highly diversified and contextual based resulting into the need to adapt several existing items and combine them with new items suitable to the main research. "A sound research plan calls for a thorough discussion about the instrument or instruments – their development, their items, their scales, and reports of reliability and validity of scores on past uses" (Creswell, 2014).

The field of evaluation is increasingly identifying the importance of evaluation capacity for the promotion, conduct, and utilization of effective evaluation (Trevisan, 2002). However, measuring evaluation capacity has been a continuous challenge for evaluation scholars and practitioners. More research is, therefore needed to guide the conceptualisation and measurement of factors that are related to evaluation capacity (Preskill & Boyle, 2008). The challenge in getting ready instruments in measuring the proposed framework has forced the researcher to go through various processes in coming out with the right measurement scale since validity and reliability become the central concern of every measurement.

Hence, this paper highlights the rigor processes which involved pre-test session with practitioners, content review session with experts, and pilot test in efforts to ensure the developed instruments are valid and reliable. Generally, this study contributes towards the existing literature on evaluation by providing new and adapted scales tested in a new context of the public sector. This research also fills another knowledge gap by conducting the data collection in eight (8) different ministries in the Malaysian public sector, thus providing variability to the existing studies. Studies on organisational evaluation capacity tend to focus on a particular type of organisation, and very little work has focused on measuring evaluation capacity in different organisations thus far (Bourgeois, Whynot, & Theriault, 2015).

Literature Review

Construct Development

According to Scriven (1991:p.139), evaluation refers to "*the process of determining the merit, worth, or value of something or the product of that process*". He further clarified that evaluation could be applied to programs, policies, performance, products, personnel, and proposals. In this study context, the implementation of evaluation refers to evaluation activities carried out by ministries to determine the merit, worth, and value of government policies and programs. On the other hand, ECB is defined as the capacity of putting in place structures that support evaluation efforts within an organisation. In this study, the structural aspect of ECB becomes the main focus where the capital aspects of institutional (evaluation office), human (internal evaluators), technical (evaluation information

system), financial (financial resources), and legal (evaluation regulatory framework) are examined as to what extent they influence the implementation of evaluation activities at the ministry level. While evaluation use or sometimes also referred to as evaluation utilisation is one of the most researched themes in the literature on evaluation. In this study, evaluation use is referred as the aspect of symbolic use (accountability) and contextual use (organisational learning).

In efforts to avoid any possible confusion in the interpretation of the constructs used in the main research, the operational definition of terms that provides meanings of each construct is provided. The following sub headings listed the definition of terms used in the research:

a) *Implementation of Evaluation*

Conceptually, evaluation is defined as “a process of systematic inquiry to provide information for decision-making about some object – a program, project, process, organisation, system, or product” (Preskill & Torres, 1999). In this research, the implementation aspect is emphasised and therefore termed as the implementation of evaluation. Implementation of evaluation is referred to evaluation activities carried out to determine the merit, worth, and value of government policies and programs.

b) *Accountability*

Accountability is the process through which an organisation makes a commitment to respond and balance the needs of its diverse stakeholders in decision-making processes and activities, and deliver against this commitment (Global Accountability Report, 2008). In this research, accountability refers to the commitment of being accountable in carrying out public office responsibilities, which covers elements such as transparency, participation, evaluation, and complaints and response.

c) *Organisational Learning*

According to Hallie & Torres (1999), organisational learning is about creating continuous processes and mechanisms for learning about how to do things better. In this research, organisational learning refers to a continuous process of organisational growth and improvement through lessons drawing in the policy and program cycle. It triggers learning the culture, participatory decision-making, risk-taking, problem-solving, and change process in an organisation.

d) *Evaluation Office*

Evaluation office is a term inspired from the study by Naidoo (2011) and Mucciarone and Neilson (2011), which highlighted the importance of coercive isomorphism element proxied by the existence of oversight infrastructure or bodies. In this research, evaluation office refers to the internal entity or unit established to lead and undertake evaluation activities within organisations.

e) *Internal Evaluators*

According to Baron (2011), internal evaluators are the employees of the organisation who perform evaluation function to any degree, whether alone or in conjunction with other duties and

responsibilities. In this research, internal evaluators are defined as 'the employees of an organisation who are responsible for organisations' self-evaluation work'.

f) *Evaluation Information System*

In this research, evaluation information system is defined as a system established and used to manage on-going evaluation data and ready for public disclosure. It is inspired by the term evaluation technology, which refers to the methodological rigour or the same range of computing hardware used in program evaluation (Bamberger, 1991).

g) *Financial Resources*

According to Bourgeois and Cousins (2013), budget or financial resources refer to the stability of the evaluation budget and whether it provides sufficient funding to complete the activities outlined in the evaluation plan. In this study, financial resources refer to any kind of financial assistance, aid, budget or funds allocated for organisational evaluation activities.

h) *Evaluation Regulatory Framework*

Evaluation regulatory framework is a term inspired by the concept of coercive isomorphism in the institutional theory that focuses on the need for the pressures to be coercive, taking the form of laws, mandates, and rules. In this study, it refers to the framework that controls evaluation practices and activities in the organisation through administrative regulations, circulars, or guide. It is a mandate for evaluation office to carry out functions without fear and favour.

Development and Selection of Research Instrument

Despite growing literature on building evaluation capacity, a dynamic and complex organizational process, the field lacks empirically validated models and corresponding assessment instruments that integrate and synthesize currently agreed upon components of evaluation capacity and allow for its measurement (Labin, Duffy, Meyers, Wandersman, & Lesesne, 2012; Taylor-Ritzler, T., Suarez-Balcazar, Y., & Garcia-Iriarte, E., 2009). The majority of current instruments were developed from case studies and systematic analyses of the literature (e.g., Danseco, E., Halsall, T., & Kasprzak, 2009; Preskill & Torres, 2000; Volkov & King, 2007); none were designed to validate empirically a conceptual model of evaluation capacity, and only a few provide psychometric data (Taylor-Ritzler, Suarez-Balcazar, Edurne Garcia-Iriarte, Henry, & Balcazar, 2013).

According to Saunders et al. (2009), questionnaires are commonly used for descriptive or explanatory research that is undertaken using attitude and opinion questionnaires or questionnaires of organizational practices, in order to identify and explain the variability in different phenomena. There are three types of data variable that can be collected through questionnaires that include opinion, behavior, and attribute (Dillman, 2007). The distinctions among these three types of data variable are important as they influence the way questions are worded. Since the questionnaire's response will be generalized to the whole population, it is important for the questionnaires to suit not only the

research but also the respondents especially adapting to the language and terms comprehensible by the respondents.

In this study, there were eight variables used; evaluation office, internal evaluators, evaluation information system, financial resources, evaluation regulatory framework, implementation of evaluation, accountability, and organizational learning that are related to the opinion, behavior, and attribute of the respondents. This self-administered questionnaire was used for data collection with a total of 58 final questions, with another nine questions on demographic information. The distinctions among these three types of data variable are important as they influence the way questions are worded. The questions or measurement scales, which were developed systematically, helped to ensure that the research findings are subject to generalization.

Validity and Reliability of Instrument

In any research, validity and reliability become the central concern of any measurement where all researchers would try to ensure these requirements are met. Validity and reliability of scores on instruments lead to meaningful interpretations of data (Creswell, 2014). Validity simply means truthfulness. In simple terms, validity addresses the question of how well we measure social reality using our constructs about it (Neuman, 2014). It refers to the capability of a measurement or a research instrument to measure the true value of a concept in a hypothesis (Piaw, 2016). Validity is high when the instrument is able to measure the concepts mentioned in the operational definitions and hypothesis. In this study, validity of the measurement was ensured through the pre-test session with practitioners and the content review session with experts.

On the other hand, reliability means dependability or consistency where it suggests that the same thing is repeated or recurs under the identical or very similar conditions (Neuman, 2014). Reliable instruments can be used many times in different settings and timelines and produce explicit and consistent results. In this study, the reliability statistics using Cronbach's Alpha scores is used during the pilot test.

Methodology

This study involved thorough item development processes which are divided into three stages. The first stage started with the item generation process, which involved comprehensive literature search on the existing instrument, followed by adaptation and developing new items. This has resulted into pool of items that were verified during interview sessions with six practitioners. The first draft of questionnaire was later pre-tested by six selected practitioners from various backgrounds. Next, a content validity session was held with three experts, where two experts from the Malaysian public sector and an academician from a university were involved. The experts gave opinion on the validity of items based on relevance, clarity and technical aspects of the items.

The following stage involves a pilot survey on 50 practitioners who are currently involved in the evaluation activities or had any work experience on evaluation-related activities at the ministry level

in the past postings. Reliability analysis was done at this stage to enable the necessary correction carried out before the main survey commences. This involves the process of items deletion based on the reliability test results. The final stage was the main survey which involved a total of 372 practitioners in all line ministries in the Malaysian public sector. At this stage, another round of reliability analysis was done to ensure rigor in the measurement process.

The questionnaires consisted of multi-items and multi-scales, which previous researchers had tested in the past studies. Since there are limited empirical studies conducted on ECB and evaluation as compared to case studies and analysis, there are limited ready scales available. In the case where there is no existing scale to measure the construct, several steps were undertaken. This includes generating new items and combining items from different dimensions and constructs. Views from practitioners during the interviews, pretest and expert review sessions were also considered in the item generation processes. The details on every measurement development process were scrutinised in order to ensure the generation of items meets the minimum requirement set for each process.

In this study, the focus was on the implementation of evaluation on policies and programs, and how ECB structural factors or antecedents influenced evaluation activities and finally lead towards the use of evaluation in accountability and organizational learning. The ECB structural factors involved five variables namely evaluation office, internal evaluators, evaluation information system, financial resources, and evaluation regulatory framework. This questionnaire applies the five point Likert scale with 1 = Strongly Disagree, 2 = Disagree, 3 = Neither Agree nor Disagree, 4 = Agree, and 5 = Strongly Agree.

At the pilot stage, the Statistical Package for the Social Sciences (SPSS) version 22 was used to analyse the questionnaire's reliability and internal consistency. Out of 50 respondents, only 35 respondents responded to the survey and gave general comments on the length, volume and understanding of the items. Necessary refinement procedure was taken care based on the reliability analysis and the general comments by the respondents. The assessment of convergent validity and discriminant validity will later be applied in the main study to ensure validity of the measurement model. The overall steps undertaken in the overall instrument building process are visualized in the following Figure 1.

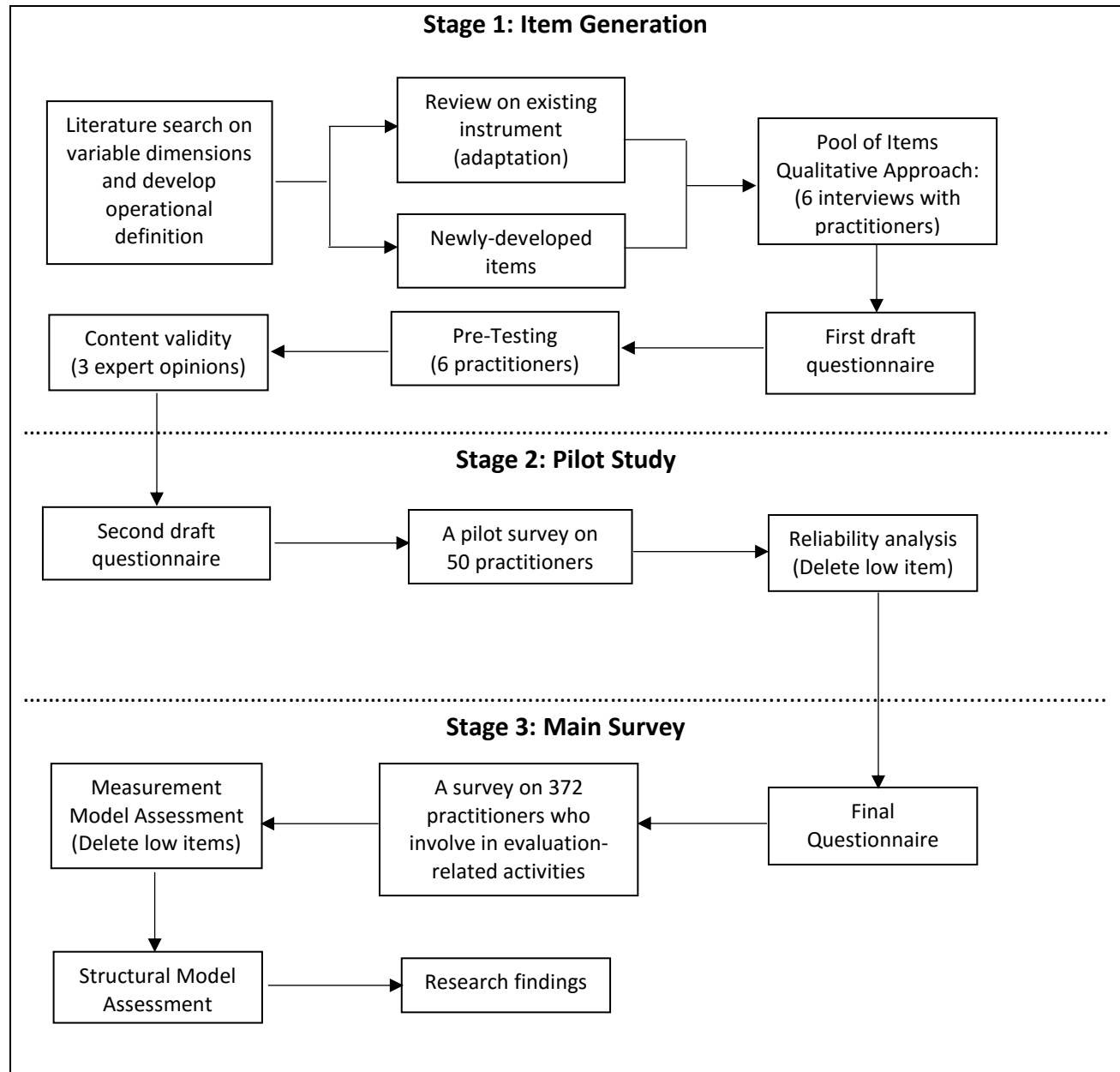


Figure 1:
 Measurement Development Processes Applied in the Study

Identifying the dimensions of measurement scales and generating items is crucial in any research process. Thus, an extensive literature review was done to investigate suitable dimensions that become a basis for constructing an operational definition of research variables and later for developing measurement scales. Identifying and defining variables are regarded as critical steps in a research study that will impact the validity and reliability of the measurement. In addition to that, interview sessions with practitioners also contribute to several newly identified dimensions of

particular variables to ensure comprehensive measurement of tested variables. There were several feedbacks and suggestions highlighted during the interview and expert review sessions that really help in the measurement development process. The process helped to establish necessary aspects to be measured in the research.

In the event where there is no existing instrument, new items were generated based on several reliable literature sources such as the 'Checklist for Building Organizational Evaluation Capacity' by Volkov and King (2007). The checklist is developed based on case study data and extensive literature review that becomes a resource for stakeholders in organizations to increase long-term capacity to conduct and use program evaluation. In measuring accountability as a variable, for example, it presents few existing quantitative studies, making it difficult to find existing instrument because of highly fragmented and non-cumulative nature of accountability definitions and concepts based on different study contexts. However, there are several established indicators for accountability developed through landmark studies and suitable to be used. The relevant indicators were identified and suited to the study context together. While for the new variables, evaluation office and evaluation regulatory framework, new items were developed based on relevant literature that suits the study context.

In developing an instrument that suits the current study context, the existing instruments were adapted accordingly with necessary modifications depending on the suitability of the measured dimensions in each variable. This had resulted in the creation of several hybrid scales, where several existing instruments were used to inspire and develop new scales to measure the tested variables. At the first stage, this study employed a self-administered questionnaire for data collection with a total of 106 items, including demographic questions as described in Table 1 below. The following table lists the constructs, the number of items used, and the adapted sources. There is quite large number of items listed during the initial stage of measurement development. This is due to various number of items in the adapted scales that represent various dimensions.

Table 1:
Measurement Items

Constructs	Number of Items	Sources
Evaluation Office	10	• Newly developed based on Naidoo (2011) and Volkov & King (2007)
Internal Evaluators	10	• Adapted from Fleischer et al. (2008) and Taylor-Ritzler et al. (2013)
Evaluation Information System	10	• Adapted from Taylor-Ritzler et al. (2013) and Preskill & Torres (2000)
Financial Resources	10	• Adapted from Walker-Egea (2014), Bourgeois et al. (2013) and newly developed based on Volkov & King (2007)

Constructs	Number of Items	Sources
Evaluation Regulatory Framework	10	• Newly developed based on Khan (1998) and Kudo (2003)
Implementation of Evaluation	12	• Adapted from Preskill & Torres (2000) and Walker-Egea (2014)
Evaluation Use (Accountability)	18	• Adapted from GAR - One World Trust (Lloyd, Warren, & Hammer, 2008).
Evaluation Use (Organisational Learning)	18	• Adapted from Botcheva, White, & Huffman (2002), Preskill & Torres (1999), and Volkov (2008)
Demographic	8	

In this study, the evaluation office is a new variable where its scales are newly developed and adapted based on Naidoo (2011) and Volkov and King (2007) past works. Initially, there were 10 items developed and adapted to measure the construct. The scales for internal evaluators were adapted based on Fleischer et al. (2008), and Taylor-Ritzler, Suarez-Balcazar, Garcia-Iriarte, Henry and Balcazar (2013) with two adapted measured dimensions. The scale for an evaluation information system, on the other hand, was a combination of measured dimensions from Taylor-Ritzler et al. (2013) and Preskill and Torres (2000). Meanwhile, the scale for financial resources was developed based on Walker-Egea (2014), Bourgeois et al. (2013), and Volkov and King (2003). The evaluation regulatory framework as a new variable used new scales developed and inspired based on Khan (1998), and Kudo (2003) works. Next, the variable implementation of evaluation used the scale developed by Preskill and Torres (2000) and Walker-Egea (2014), while the accountability variable used the established scale developed by the Global Accountability Report - One World Trust (Lloyd et al., 2008). The last variable, organizational learning, used the scale developed by Botcheva, White, and Huffman (2002), Preskill and Torres, (1999), and Volkov, (2008).

Content Validity: Pre-Test

Content validity refers to the extent to which a specific set of items for measuring variables reflects its content domain (DeVellis, 2003; Dremina et al., 2016; Ogbiji, 2018; Peprah, 2018). There are basically two goals of this process; i) to assess the content validity of various scales being developed, and ii) to identify any items which remain unclear. Upon completion of the questionnaire draft, other processes were done to ensure its validity further. This included conducting a few series of semi-structured interview sessions with relevant practitioners to have further exploration on ECB factors, implementation of evaluation and evaluation use issues in the Malaysian public sector. Short semi-structured interviews were conducted with six public sector officers at the grade 48 and above who involve in monitoring and evaluation works at the ministry level or officers with the same work experience in the previous postings.

This is crucial to assure and confirm issues in the implementation of evaluation on policies and programs, and its antecedents and consequences did exist in the targeted population. These

respondents pre-tested the questionnaires and indirectly confirmed the issues and problems that exist in evaluation-related works at the ministry level. Overall, the process facilitated great improvement on items used to measure certain variables. However, almost all pre-test respondents gave feedback that there were too many items in the questionnaire set, leaving the targeted respondents with high tendency to lose focus along the process. Therefore, it was highly suggested that the researcher to relook into the questionnaires and reduce the number of items.

In addition to that, there were few new dimensions suggested from this process, such as taking into account the element of systems integration as well as data accuracy for the evaluation information system variable. As for the variable of evaluation regulatory framework, it was recommended to have items on administrative authority through evaluation regulatory framework. It is important to identify the right dimension to be measured as this is a new variable derived in this study. On the other hand, there were few new items suggested for researcher's consideration and some items were suggested for wording improvement for better understanding among respondents. Overall, the process had helped to validate important aspects to be measured in the constructs, especially the newly developed items.

It was also suggested that some brief information that explains on the definition of every measured variable should be included at the front part of the questionnaire to give some ideas to the respondents about the measured variables before proceeding to answer the questions. This information is crucial as the research involved respondents from various background who might be senior or new officers in the divisions with different work experience. Therefore, such information is valuable, especially to the new officers in the divisions. The information is also useful, especially when several terms used in the academic reference and practical settings are sometimes different and understood differently, and therefore require some information enlightenment.

Content Validity: Expert Opinions

This process involved incorporating the opinions of experts who deal with policy evaluation in the public sector with the opinion of academics. In this aspect, expert opinion was incorporated to ensure items being developed are relevant and representing the measured variables. There were two experts selected from the public sector; a Senior Deputy Director and Principle Assistant Secretary who have vast work experience on evaluation-related activities in several ministries and departments in their past and current postings. An academician was selected to help giving views on the questionnaire building process.

The choice of experts was consistent with the ideal number of experts, as suggested by scholars, which is between a range of two to 20 (Rubio, Berg-Weger, Lee, & Rauch, 2003). Throughout the process, experts were asked to state whether they agree or not with the lists of questionnaires items by providing specific reasons, for any opinions on the instrument improvement. Details of the experts who involved in the content validity process are outlined in Table 2 below:

Table 2:
Profile of Experts (Practitioners and Academician)

Expert	Position	Institution	Interview Duration	Interview Date
1	Senior Deputy Director Research, Planning, and Policy Division	Public Service Department, Malaysia	1 hour 15 minutes	3rd March 2017
2	Principal Assistant Secretary Policy Division (Labor)	Ministry of Human Resources	45 minutes	27th February 2017
3	Research Fellow Accounting Research Institute (ARI), UiTM (Speaker for Questionnaire Building for Quantitative Studies Workshop, UiTM)	Universiti Teknologi MARA (UiTM)	1 hour 10 minutes	6th March 2017

Pilot Study

Pilot testing a survey instrument is a procedure in which a researcher makes changes (if necessary) in the instrument based on the feedback from a small number of individuals who complete and evaluate the instrument (Creswell, 2014). This is the final stage of questionnaire development, where the feasibility and suitability of actual research are roughly estimated. A rule of thumb 0.60 is used as the lower level of acceptability, as suggested by Nunnally (1978). In this research, a pilot test was conducted on a number of 50 respondents from several ministries and departments. According to Cooper and Schindler (2003), the size of the pilot group can range from 25 – 100, but it does not have to be statistically selected. In contrast, Rossi, James, and Anderson (1983) find a pilot test of 20 – 50 cases is usually sufficient to discover major flaws in the questionnaire. The pilot survey started on 20th March 2017 and carried out for a period of one month.

The questionnaire set was manually distributed to the key person of each ministry to ensure there is a personal touch in handling the research. The questionnaire was not distributed online or via email because the tendency for officers to ignore the email is especially high due to overwhelm workloads that officers need to cater through emails. The questionnaire was prepared in the English language. All of these pilot study respondents are practitioners who directly involve as well as those who had work experience in the policy planning and evaluation at the ministry level in the current or previous postings. Within a period of one month, only 37 out of 50 respondents had replied to the survey, and only 35 questionnaires are fit and usable for the final pilot data analysis.

Findings

The data from 35 respondents were then used to refine the measures by analysing their reliability and validity. The data were subjected to a further purification process. It begins with the examination of the demographic profiles of the respondents. Since the researcher personally went to the premise and met the respondents to distribute the questionnaires, majority of respondents had given their general comments on the questionnaire survey either through direct meeting or notes attached on the questionnaire survey. Therefore, item refinement procedure was later carried out mainly based on the results from the reliability analysis as well as general feedback by the respondents.

Demographic Profiles

Generally, the demographic profile of pilot respondents reflects the variability needed for each tested group, including gender, age, education level, service scheme, ministry, and service years. In terms of gender, 45.7% of the respondents were male and, 54.3% were female. In terms of age, 40% of the respondents were the officers between the age of 26 to 39, while the majority of 60% of respondents were at the age of 40 and above. The respondents also came from different academic backgrounds, where 42.9% of respondents have a bachelor's degree, 40% of the participants have a master's degree, and 17.1% of them have a PhD degree. In terms of service grade, 28.6% of the respondents came from officers at the service grade of 41 – 44, 48.6% officers at the grade of 48 – 52, and 17.1% officers from the service grade of 54. Majority of the respondents were the diplomatic and administrative officers at 54.3%, while 14.3% were the education officers, and 31.4% were the combination of various service schemes such as social, information technology, labor, higher education, and others.

On the other hand, majority of the respondents (31.4%) have served in the public service for at least 11 to 15 years, while 22.9% have served between 6 to 10 years. Another 20% of respondents served between 16 to 20 years of service, 11.4% have served more than 20 years, while balance 2.9% had served in the public service for less than 5 years. Notably, out of the total respondents, 31.4% have involved in the monitoring and evaluation works for 6 to 10 years of service while 20% involved in evaluation-related works for 11 to 15 and below 2 years respectively. Another 17.1% have involved the evaluation-related works between 3 to 5 years and a balance of 11% have involved more than 15 years. Notably, the respondents came from eight different ministries background either in the current or past postings where they involve in evaluation works.

Reliability Analysis

Basically, the reliability of research refers to the research's capability to get the same value using the same measurement. In quantitative research, the reliability concept refers to the consistency of items in any research instrument to measure the same concept (Piaw, 2016). Generally, a Cronbach's Coefficient Alpha is applied to determine the reliability of scales in the pilot study where an alpha \geq 0.7 is accepted as demonstrating a high level of homogeneity within the scale, and thereby determine whether or not the item reflects a single dimension. The details of reliability results are illustrated in the following Table 3:

Table 3:
 Reliability Test Results

Constructs	Items Code	Cronbach's Alpha	Cronbach's Alpha if the Items Deleted	Corrected Item-Total Correlation
Evaluation Entity	EE1	0.732	0.717	0.357
	EE2		0.687	0.534
	EE3		0.726	0.334
	EE4		0.699	0.460
	EE5		0.685	0.549
	EE6		0.714	0.374
	EE7		0.703	0.447
	EE8		0.774	0.028
	EE9		0.691	0.525
	EE10		0.697	0.493
Evaluators	EVA1	0.923	0.916	0.690
	EVA2		0.919	0.633
	EVA3		0.908	0.830
	EVA4		0.914	0.733
	EVA5		0.917	0.675
	EVA6		0.913	0.751
	EVA7		0.913	0.753
	EVA8		0.907	0.861
	EVA9		0.912	0.764
	EVA10		0.930	0.407
Evaluation Information System	EIS1	0.911	0.907	0.593
	EIS2		0.903	0.658
	EIS3		0.898	0.728
	EIS4		0.902	0.668
	EIS5		0.902	0.664
	EIS6		0.904	0.639
	EIS7		0.900	0.705
	EIS08		0.909	0.562
	EIS09		0.893	0.825
	EIS10		0.896	0.770
Evaluation Budget	EB1	0.923	0.910	0.802
	EB2		0.914	0.733
	EB3		0.911	0.784
	EB4		0.917	0.691
	EB5		0.916	0.699
	EB6		0.921	0.622

Constructs	Items Code	Cronbach's Alpha	Cronbach's Alpha if the Items Deleted	Corrected Item-Total Correlation
	EB7		0.913	0.755
	EB8		0.919	0.655
	EB9		0.915	0.710
	EB10		0.919	0.638
Regulatory Framework	RF01	0.892	0.878	0.676
	RF02		0.867	0.847
	RF03		0.881	0.640
	RF04		0.870	0.792
	RF05		0.875	0.720
	RF06		0.871	0.792
	RF07		0.878	0.674
	RF08		0.891	0.477
	RF09		0.905	0.317
	RF10		0.891	0.474
Implementation of Evaluation	IMPL1	0.862	0.847	0.624
	IMPL2		0.842	0.670
	IMPL3		0.866	0.271
	IMPL4		0.850	0.564
	IMPL5		0.864	0.351
	IMPL6		0.858	0.425
	IMPL7		0.862	0.386
	IMPL8		0.852	0.534
	IMPL9		0.841	0.698
	IMPL10		0.841	0.703
	IMPL11		0.846	0.628
	IMPL12		0.841	0.673
• Accountability (Evaluation Use)	ACC1	0.863	0.859	0.406
	ACC2		0.867	0.201
	ACC3		0.867	0.145
	ACC4		0.860	0.367
	ACC5		0.852	0.571
	ACC6		0.859	0.399
	ACC7		0.856	0.496
	ACC8		0.859	0.391
	ACC9		0.851	0.576
	ACC10		0.850	0.621
	ACC11		0.852	0.572
	ACC12		0.863	0.330
	ACC13		0.850	0.605

Constructs	Items Code	Cronbach's Alpha	Cronbach's Alpha if the Items Deleted	Corrected Item-Total Correlation
	ACC14		0.845	0.697
	ACC15		0.851	0.646
	ACC16		0.857	0.450
	ACC17		0.849	0.634
	ACC18		0.859	0.399
	OL1	0.861	0.852	0.539
• Organizational Learning (Evaluation Use)	OL2		0.857	0.436
	OL3		0.850	0.557
	OL4		0.856	0.440
	OL5		0.855	0.449
	OL6		0.852	0.542
	OL7		0.851	0.555
	OL8		0.852	0.522
	OL9		0.853	0.500
	OL10		0.860	0.328
	OL11		0.848	0.609
	OL12		0.856	0.436
	OL13		0.854	0.485
	OL14		0.853	0.503
	OL15		0.853	0.508
	OL16		0.847	0.643
	OL17		0.853	0.503
	OL18		0.875	-0.064

Based on this reliability test results, there were 24 items dropped from the questionnaire item list based on the scores of item-total correlations and Cronbach's alpha scores of items deleted. The item deletion process involved careful examination on the scores of these aspects in order to maintain high reliability of items of all constructs.

Pilot Respondents Comments on the Survey Questionnaire

During the pilot survey, respondents provided a few suggestions in attempts to ensure the final questionnaire for the main survey is clearer. The data gathered from these respondents were then used to further refine the measures by analyzing their reliability and dimensionality. The feedback and suggestions for improvement from the pilot survey respondents are listed in the following Table 4:

Table 4:
Respondents' Comments on the Pilot Questionnaire Survey

Bil	Questionnaire	Comments
1.	Items	<ul style="list-style-type: none"> • Too many items • Most of the items have long and complex sentences (respondents take time to read, understand and answer). There is a tendency to lose focus and attention, leaving the questionnaire incomplete. • Completion time is longer due to long and complex items (more than 15 minutes) • There are several items where respondents do not have the information to answer but just answered anyway.
2.	Questionnaire Layout	<ul style="list-style-type: none"> • The font is too small
3.	Language	<ul style="list-style-type: none"> • Should include the Malay language translation to avoid deviation from the real meaning meant by the researcher.
4.	Appropriateness of the terms used:	
	<ul style="list-style-type: none"> • Evaluation Entity 	<ul style="list-style-type: none"> • May mislead some understanding of evaluation agencies, which involve in the overall evaluation works. This involves some ministries, especially the ones that often deals with international organizations that re related to evaluation matters — suggested being replaced with 'Evaluation Office' to reflect an evaluation-related entity at the ministry level.
	<ul style="list-style-type: none"> • Evaluators 	<ul style="list-style-type: none"> • Not clear whether referred to internal or external evaluators. Respondents may not have information on external evaluators. It is suggested to use the term 'Internal Evaluators' for a more proper and specific understanding. • Many respondents seem not to agree with the term because it is not clearly laid out in the current budget system. Some other appropriate terms were suggested, such as 'Financial Resources' or 'Evaluation Aid' as these terms were more familiar in the practical context.
	<ul style="list-style-type: none"> • Evaluation Budget 	

Feedback from Content Validity Session with Experts

All of the experts were expected to give feedback in terms of relevance, clarity, and technical aspects of the questionnaire items. The relevance aspect measures whether the items appear to a good measure of every variable dimension in the study. The clarity aspect, on the other hand, measures whether the items are clearly worded and the existence of any double-barrel item that requires changes. Finally, the technical aspect measures whether the response scale is appropriate for the study. Experts could also highlight any concerns on any other relevant aspects in the other sections in the expert review form provided. Generally, the review by all experts can be summarized in Table 5 below:

Table 5:

Review Summary by Experts

Measured Aspect	Review by Expert 1	Review by Expert 2	Review by Expert 3
Relevance	<ul style="list-style-type: none"> Section 4: Evaluation Office <ul style="list-style-type: none"> -Definition of evaluation and evaluation office need to be in the questionnaire. Section 7: Financial Resources <ul style="list-style-type: none"> -The item structure should ask 'opinion level' rather than 'behaviour level'. E.g., An evaluation budget is important for any organisation to commence with any evaluation activity. Section 8: Evaluation Regulatory Framework <ul style="list-style-type: none"> -The item structure should ask 'opinion level' rather than 'behaviour level'. E.g., There is a need for the presence of a legal element in any evaluation framework to ensure the success of evaluation activities. 	<p>Good at the moment. Need to go through EFA process.</p>	<ul style="list-style-type: none"> Use only a single dependent variable (Evaluation Use) that can be measured by looking at accountability and organisational learning aspects.
Clarity	<ul style="list-style-type: none"> There is a need for a clearer statement explaining the measured variables considering various work background of officers in the policy and planning divisions. The fact that there is a combination of highly experienced and newly 	<p>Items wordings should be simple, easy to understand, and positively constructed.</p>	<p>Some of the questions need to be reworded to reflect the intended meaning.</p>

Measured Aspect	Review by Expert 1	Review by Expert 2	Review by Expert 3
	appointed officers who may require additional information on the study area cannot be ignored.		
Technical	<ul style="list-style-type: none"> • 5 point Likert scale is appropriate. • There are several items requiring changes in terms of flow due to the addition of new items and omission of existing items. • Redundant items can be identified after EFA. 	<ul style="list-style-type: none"> • Recommend 10-point Likert scale. Minimum 10 items for each construct should be sufficient. • Redundant items can be identified after EFA. 	<ul style="list-style-type: none"> • Recommend 7-point Likert scale to ensure accuracy of analysis and to avoid some issues such as normality of data. • Items for the dependent variable should be placed at the end of the questionnaire.
Others	<ul style="list-style-type: none"> • Demography Section – capture service grade category rather than merely the service category for richer information. • General information explaining the overall study framework is needed. • Clearer, better definition of every construct are needed. • There are several items that require improvement (addition and omission) to take into account the ‘opinion’ and ‘behaviour’ levels of questions. 	Overall, items constructed measure all variables under study.	<ul style="list-style-type: none"> • The quality of the questionnaire is moderate. It could be improved if the candidate makes corrections, as suggested during the discussion. Some minor changes required.

Based on an analysis of reviews and comments by the experts, necessary changes had been made to each item in the questionnaires booklet to make it clearer and understandable to the targeted respondents. The recommendation for changes and improvement was carefully studied and discussed before any amendments done to ensure it suits the overall study context.

Items Purification Procedure

The items for deletion were determined based on the correlations of items within each scale, the corrected item-to-total correlations, the items standard deviation scores, and the effects on Cronbach' alpha scores if the items were deleted. The Cronbach' alpha scores would increase if items with low item and item-scale were deleted. Based on this process, there were 24 deleted items in total. This deletion process is necessary as the reliability tests results showed that the reliability scores would increase if those items were deleted.

In addition to that, the deletion of items was also done in order to respond to the majority of pilot respondents' comments that there were too many items for each construct in the questionnaire. The total 106 items (including the demographic items) are considered too many for this type of instrument as pointed out by the experts during the expert review session. The views from the experts were further confirmed with the feedbacks from the majority of the respondents during this pilot study. Based on this information, the researcher had decided to review the overall questionnaire items and eliminate some unimportant items while retaining the dimension coverage and the desired reliability levels. Some changes on the constructs' name were performed in order to avoid misunderstanding of information among respondents. Several names of constructs might be usable and were commonly referred to as in the academic's context but might be differently understood in the practitioner's context. By taking into account this information and feedbacks gained during the pilot study, some constructs were renamed accordingly by making sure the meaning is not deviated from what the researcher wishes to measure.

In this aspect, the earlier construct's names that were modified are 'evaluation entity' to 'evaluation office', 'evaluators' to 'internal evaluators', 'evaluation budget' to 'financial resources', and 'regulatory framework' to 'evaluation regulatory framework'. The changes were done based on the respondents' feedbacks by ensuring the constructs' meanings do not deviate from the study context. After dropping down several items, the reliability test was run again to make sure that the Cronbach's alpha scores for all constructs are still within the acceptable scores range. It was found that the reliability scores for all constructs were all above 0.70 and therefore met the minimum requirement of the reliable instrument. The following Table 6 provides details on the results obtained from the pilot test.

Table 6:
 Results from the Pilot Test

Final Constructs' Names	Pilot Test				
	Number of Items	Number of Items Dropped	Cronbach' Alpha	Final Number of Items for Main Survey	Cronbach' Alpha (New)
Evaluation Office	10	4	0.732	6	0.873

Final Constructs' Names	Pilot Test				
	Number of Items	Number of Items Dropped	Cronbach' Alpha	Final Number of Items for Main Survey	Cronbach' Alpha (New)
Internal Evaluators	10	4	0.923	6	0.898
Evaluation Information System	10	4	0.911	6	0.883
Financial Resources	10	4	0.923	6	0.908
Evaluation Regulatory Framework	10	4	0.892	6	0.883
Implementation of Evaluation	12	1	0.862	11	0.932
Use of Evaluation					
• Accountability	18	8	0.863	10	0.910
• Organisational Learning	18	11	0.861	7	0.889
Demographic	8	0		9	
Total	106	40		67	

Final Instrumentation

Systematic literature search, new item generation based on the established checklist and case studies, pre-test and interview session with practitioners, and finally content validity sessions by the experts had resulted into the final instrumentation to be used in the final survey. With a total of 58 items and nine demography items, making a total of 67 items, the final instrument was then ready for the final data collection. The following Table 7 lists the final items for each construct used in the study with the final Cronbach's Alpha score. The details highlight the items which are adapted and newly developed to suit the current study context.

Table 7:
 Final Items for Each Construct

Constructs	Codes	Final Items	Item Status	Cronbach's Alpha (New)
Evaluation Office	EO1	<ul style="list-style-type: none"> There is a clear and capable evaluation office responsible for monitoring and evaluation activities in my organization. 	Newly developed based on Naidoo (2011) and Volkov & King (2007)	0.873
	EO2	<ul style="list-style-type: none"> There is a detailed written evaluation plan to assess ministries' policies. 		
	EO3	<ul style="list-style-type: none"> The evaluation function in my organization is well-developed. 		
	EO4	<ul style="list-style-type: none"> The evaluation reports to central agencies or centralized evaluation office help to improve organizational evaluation capacity. 		
	EO5	<ul style="list-style-type: none"> The evaluation reports required by central agencies or centralized evaluation office add value to my organizational works. 		
	EO6	<ul style="list-style-type: none"> The work of the evaluation office is important as it promotes accountability in the organization. 		
Internal Evaluators	IE1	<ul style="list-style-type: none"> In my organization, internal evaluators facilitate organizational learning in conducting the evaluation. 	Adapted from Fleischer et al. (2008) and Taylor-Ritzler et al. (2013)	0.898
	IE2	<ul style="list-style-type: none"> In my organization, internal evaluators are accountable to intended users and use of evaluation. 		
	IE3	<ul style="list-style-type: none"> In my organization, internal evaluators involve in the evaluation follow-up activities (e.g., dissemination of evaluation information, action planning, etc.). 		
	IE4	<ul style="list-style-type: none"> It is the responsibility of internal evaluators to build evaluation capacity in my organization. 		
	IE5	<ul style="list-style-type: none"> Internal evaluators in my organization are able to define outcome indicators of the measured policies. 		

Constructs	Codes	Final Items	Item Status	Cronbach's Alpha (New)
	IE6	<ul style="list-style-type: none"> Internal evaluators in my organization are able to develop recommendations based on evaluation results. 		
Evaluation Information System	EIS1	<ul style="list-style-type: none"> My organization provides basic technological resources for evaluation activities (e.g., computers and software). 	Adapted from Taylor-Ritzler et al. (2013) and Preskill & Torres (2000)	0.883
	EIS2	<ul style="list-style-type: none"> My organization uses specific systems (software) to manage evaluation data. 		
	EIS3	<ul style="list-style-type: none"> Employees in my organization have access to adequate technology to produce evaluation information. 		
	EIS4	<ul style="list-style-type: none"> Technical assistance is available to employees in the organization to address questions related to evaluation. 		
	EIS5	<ul style="list-style-type: none"> Evaluation information tells details about the effectiveness of policies and programs. 		
	EIS6	<ul style="list-style-type: none"> There is a system in the organization to disseminate information on evaluation. 		
Financial Resources	FR1	<ul style="list-style-type: none"> There is a plan for securing the necessary resources to develop evaluation capacity. 	Adapted from Walker-Egea (2014), Bourgeois et al. (2013) and newly developed based on Volkov & King (2007)	0.908
	FR2	<ul style="list-style-type: none"> Financial resource for evaluation activities is a priority for my organization. 		
	FR3	<ul style="list-style-type: none"> Budget costs for a specific evaluation are clearly outlined based on the proposed methodology. 		
	FR4	<ul style="list-style-type: none"> Adequate resources are available to complete evaluation projects in the organization. 		
	FR5	<ul style="list-style-type: none"> There is other available financial fund or aid for evaluation purposes such as 		

Constructs	Codes	Final Items	Item Status	Cronbach's Alpha (New)
		from international organizations, civil society organizations, etc.		
	FR6	<ul style="list-style-type: none"> Despite the budget cut, financial resources for evaluation activities remain a priority in the organization. 		
Evaluation Regulatory Framework	ERF1	<ul style="list-style-type: none"> There is a clear regulatory framework that guides the current evaluation practice. 	Newly developed based on khan (1998) and kudo (2003)	0.883
	ERF2	<ul style="list-style-type: none"> There is a clear administrative authority by my organization in exercising evaluation activity. 		
	ERF3	<ul style="list-style-type: none"> There is a clear legal ruling on the need to conduct evaluation activity in the organization (e.g., Act, Enactment or anything that derived within legal basis). 		
	ERF4	<ul style="list-style-type: none"> Central agencies release a specific ruling or directive for conducting an evaluation on any policy. 		
	ERF5	<ul style="list-style-type: none"> Following the release of any ruling on evaluation, there is a proper and standard guideline on evaluation issued. 		
	ERF6	<ul style="list-style-type: none"> The release of evaluation guidelines by the authorized agency helps in succeeding evaluation activities in the organization 		
	IMPL2	<ul style="list-style-type: none"> We integrate evaluation activities into our work in order to enhance the quality of decision making. 	Adapted from Preskill & Torres (2000) and Walker-Egea (2014)	0.932
	IMPL3	<ul style="list-style-type: none"> Management encourages us to evaluate our efforts at work. 		
	IMPL4	<ul style="list-style-type: none"> When dealing with evaluation work, there is always support from other employees. 		

Constructs	Codes	Final Items	Item Status	Cronbach's Alpha (New)
Implementation of Evaluation	IMPL5	<ul style="list-style-type: none"> Evaluation, research, and policy units coordinate efforts to minimize work duplication and leverage research. 		
	IMPL6	<ul style="list-style-type: none"> Evaluation steering committees are regularly used to guide the direction and key issues for evaluation. 		
	IMPL7	<ul style="list-style-type: none"> The organizational leadership supports the development of internal evaluation capacity. 		
	IMPL8	<ul style="list-style-type: none"> Organizational leadership is committed to supporting practices that integrate evaluation into the on-going work of the organization. 		
	IMPL9	<ul style="list-style-type: none"> Organizational leadership is committed to the development of an evaluative learning culture. 		
	IMPL10	<ul style="list-style-type: none"> Organizational leadership is committed to investing in training or professional development to increase employees' evaluation skills. 		
	IMPL11	<ul style="list-style-type: none"> The organizational leadership use evaluation data to inform their decision making. 		
Use of Evaluation: Accountability	ACC1	<ul style="list-style-type: none"> My organization is committed to being transparent. 	Adapted from GAR - One World Trust (Lloyd et al., 2008).	0.910
	ACC2	<ul style="list-style-type: none"> My organization has a policy or written guidelines on disclosure of information. 		
	ACC3	<ul style="list-style-type: none"> My organization responds to all requests for information. 		
	ACC4	<ul style="list-style-type: none"> My organization engages stakeholders in decision-making processes on matters that affect them. 		
	ACC5	<ul style="list-style-type: none"> My organization changes policy and practice as a result of engagement. 		
	ACC6	<ul style="list-style-type: none"> My organization is constantly committed to evaluating its work. 		

Constructs	Codes	Final Items	Item Status	Cronbach's Alpha (New)
	ACC7	<ul style="list-style-type: none"> My organization integrates learning from evaluation to future planning. 		
	ACC8	<ul style="list-style-type: none"> My organization is open about the results of evaluations. 		
	ACC9	<ul style="list-style-type: none"> My organization responds to all valid complaints. 		
	ACC10	<ul style="list-style-type: none"> All complaints are investigated fairly. 		
Use of Evaluation: Organizational Learning	OL1	<ul style="list-style-type: none"> The management is supportive of evaluation work. 	Adapted from	0.889
	OL2	<ul style="list-style-type: none"> Evaluation is not seen as costing too much in terms of money, time, and personal resources. 	Botcheva, White, & Huffman	
	OL3	<ul style="list-style-type: none"> Employees are confident to share their opinions even if those opinions are different from the majority. 	(2002), Preskill & Torres	
	OL4	<ul style="list-style-type: none"> Employees are provided with opportunities to reflect on their works. 	(1999), and Volkov	
	OL5	<ul style="list-style-type: none"> The degree of bureaucracy is low and makes it easy to bring about organizational change. 	(2008)	
	OL6	<ul style="list-style-type: none"> There is always support when someone tries to do something different. 		
	OL7	<ul style="list-style-type: none"> There is support from management for any needed organizational change. 		

Conclusion and Recommendations

This research has established the validity and reliability of the implementation of evaluation scale that measures ECB antecedents and consequences in the Malaysian public sector. The questionnaire set is verified valid and reliable to be employed in future researches after going through exhaustive processes that include pre-test, content review session with experts, and pilot test. The valuable feedbacks gained during the pre-test session with practitioners helped to validate the important constructs of the study. Later, the content validity session with the experts had confirmed the aspects of relevance, clarity, and technical of the instrument are met. And finally the results of reliability analysis in the pilot test had confirmed the reliability of the constructs with Cronbach' Alpha scores

above 0.70 for the all eight constructs. This questionnaire can be used in future evaluation studies to help new researchers to understand the status of evaluation implementation, ECB factors and evaluation use of policies and programs in the public sector setting. More importantly, the new constructs with the new set of items were introduced in the study and recorded acceptable Cronbach's Alpha scores of 0.862 for 'evaluation office' (5 items) and 0.846 for 'evaluation regulatory framework' (6 items). These new constructs were proven as significant ECB factors in evaluation activities in the developing country in contrast to the use of stricter laws and acts in evaluation works in the developed countries. It became a meaningful contribution to the existing knowledge in evaluation field. Future researches should attempt to test this new set of questionnaires into different public sector study settings in understanding evaluation implementation, ECB factors and evaluation use.

Acknowledgement

This publication uses data collected within the framework of the PhD thesis titled *Implementation of Evaluation on Policies and Programs in the Malaysian Public Sector: An Empirical Examination of Antecedents and Consequences* of Rafidah Mohamed Hashim submitted in November 2019 to the Faculty of Administrative Science & Policy Studies, Universiti Teknologi MARA (UiTM). This PhD scholarship is sponsored by the Public Service Department, Malaysia.

References

- Bamberger, M. (1991). The politics of evaluation in developing countries. *Evaluation and Program Planning, 14*(4), 325–339. [https://doi.org/10.1016/0149-7189\(91\)90015-9](https://doi.org/10.1016/0149-7189(91)90015-9)
- Baron, M. E. (2011). Designing internal evaluation for a small organization with limited resources. *New Directions for Evaluation, (132)*, 87–99.
- Botcheva, L., White, C. R., & Huffman, L. C. (2002). Learning Culture and Outcomes Measurement Practices in Community. *American Journal of Evaluation, 23*(4), 421–434. <https://doi.org/10.1177/109821400202300404>
- Bourgeois, I., & Cousins, J. B. (2013). Understanding Dimensions of Organizational Evaluation Capacity. *American Journal of Evaluation, 34*, 299–319. <https://doi.org/10.1177/1098214013477235>
- Bourgeois, Isabelle, Toews, E., Whynot, J., & Lamarche, M. K. (2013). Measuring organizational evaluation capacity in the Canadian federal government. *Canadian Journal of Program Evaluation, 28*(2), 1–19.
- Bourgeois, Isabelle, Whynot, J., & Theriault, E. (2015). Application of an organizational evaluation capacity self-assessment instrument to different organizations: Similarities and lessons learned. *Evaluation and Program Planning, 50*, 47–55. <https://doi.org/10.1016/j.evalprogplan.2015.01.004>
- Connie F. Walker-Egea. (2014). *Design and Validation of an Evaluation Checklist for Organizational Readiness for Evaluation Capacity Development*. University of South Florida.
- Cooper, D. R., & Schindler, P. S. (2003). *Business Research Methods* (8th Intern). New York: McGraw-Hill,.

- Creswell, J. W. (2014). *Research design: Qualitative, quantitative and mixed methods approaches 4th edition*.
- Dansecu, E., Halsall, T., & Kasprzak, S. (2009). *Readiness assessment tool for evaluation capacity building*. Ottawa, Canada.
- DeVellis, R. F. (2003). *Scale development: theory and implications*. Thousand Oaks, CA: Sage Publications.
- Dillman, D. (2007). *Mail and internet surveys: The tailored design method*. New York: Wiley.
- Fleischer, D., Christie, C., LaVelle, K., & Acres, F. (2008). Perceptions of Evaluation Capacity Building in the United States: A Descriptive Study of American Evaluation Association Members. *The Canadian Journal of Program Evaluation*, 23(3), 37–60. Retrieved from <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:PERCEPTIONS+OF+EVALUATION+CAPACITY+BUILDING+IN+THE+UNITED+STATES:+A+DESCRIPTIVE+STUDY+OF+AMERICAN+EVALUATION+ASSOCIATION+MEMBERS#4>
- Hallie, P., & Rosalie, T. (1999). Assessing an Organization's Readiness for Learning from Evaluative Inquiry. In *American Evaluation Association Annual Conference* (pp. 1–15). Orlando, Florida.
- Khan, M. A. (1998). Evaluation capacity building: An overview of current status, issues and options. *Evaluation*, 4(3), 310–328. <https://doi.org/10.1177/13563899822208626>
- Kudo, H. (2003). Between the 'Governance' Model and the Policy Evaluation Act: New Public Management in Japan. *International Review of Administrative Sciences*, 69(4), 483–504. <https://doi.org/10.1177/0020852303694005>
- Labin, S. N., Duffy, J. L., Meyers, D. C., Wandersman, A., & Lesesne, C. A. (2012). A Research Synthesis of the Evaluation Capacity Building Literature. *American Journal of Evaluation*, 33(3), 307–338. <https://doi.org/10.1177/1098214011434608>
- Lloyd, R., Warren, S., & Hammer, M. (2008). *2008 Global Accountability Report*. London.
- Trevisan, M. S. (2002). Evaluation capacity in K-12 school counseling programs. *American Journal of Evaluation*, 23(3), 291–305.
- Michael Schriener. (1991). *Evaluation Thesaurus*. Newbury Park, California: Sage Publications.
- Mucciarone, M. A., & Neilson, J. (2011). Performance Reporting in the Malaysian Government. *Asian Academy of Management Journal of Accounting and Finance*, 7(2), 35–77.
- Naidoo, I. A. (2011a). *The role of monitoring and evaluation in promoting good governance in South Africa : A case study of the Department of Social Development*. University of Witwatersrand, Johannesburg.
- Naidoo, I. A. (2011b). The Use of Question - Examples and Lessons from the Public Service Commission. In *Proceedings from the Second International Conference on National Evaluation Capacities 12 -14 September 2011* (pp. 32–38). Johannesburg, South Africa.
- Neuman, W. L. (2014). *Social Research Methods: Qualitative and Quantitative Approaches* (7th editio). United States of America: Pearson Education Limited. <https://doi.org/10.2307/3211488>
- Piaw, C. Y. (2016). *Mastering Research Method*. Shah Alam, Malaysia: McGraw-Hill Education Malaysia.
- Preskill, H., & Boyle, S. (2008). A Multidisciplinary Model of Evaluation Capacity Building. *American Journal of Evaluation*, 29(4), 443–459. <https://doi.org/10.1177/1098214008324182>

- Preskill, Hallie, & Torres, R. T. (2000). Readiness for Organizational Learning and Evaluation Survey Instrument-1.
- Rossi, P. H., James, D. W., & Anderson, A. B. (1983). *Handbook of Survey Research*. Orlando and London: Academic Press.
- Rubio, D. M., Berg-Weger, M., M., S. T., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work. *Social Work Research, 27*(2), 94–104.
- Saunders, M., Lewis, P., & Thornhill, A. (2009). *Research Methods for Business Students. Research methods for business students* (5th editio). Italy: Pearson Education Limited.
- Taylor-Ritzler, Suarez-Balcazar, Garcia-Iriarte, Henry, D. B, & Balcazar, F. E. (2013). Evaluation Capacity Assessment Instrument (ECAI).
- Taylor-Ritzler, T., Suarez-Balcazar, Y., & Garcia-Iriarte, E. (2009). Evaluation capacity: Model validation using quantitative results. In *In (Y., Suarez-Balcazar, Chair), Evaluation capacity building: Taking stock ofand advancing frameworks, strategies, outcomes and measurement. Presented at theannual meeting of the American Evaluation Association Conference*. Orlando, Florida.
- Taylor-Ritzler, T., Suarez-Balcazar, Y., Garcia-Iriarte, E., Henry, D. B., & Balcazar, F. E. (2013). Understanding and Measuring Evaluation Capacity: A Model and Instrument Validation Study. *American Journal of Evaluation, 34*(2), 190–206.
<https://doi.org/10.1177/1098214012471421>
- Volkov, B. B. (2008). *Toward Continuous Improvement in Organizations: A Case Study of Evaluation Capacity Building in the Northwest Area Foundation*. University of Minnesota. Retrieved from <http://internal-pdf//Volkov King-0968442881/Volkov King.pdf%5Cnhttp://search.proquest.com.ezp.lib.unimelb.edu.au/docview/760107380?accountid=12372> http://sfx.unimelb.hosted.exlibrisgroup.com/sfxlcl41?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:di
- Volkov, B. B., & King, J. A. (2007). A Checklist for building organizational evaluation capacity. *The Evaluation Center Western Michigan University, 1–5*. Retrieved from http://www.wmich.edu/evalctr/archive_checklists/ecb.pdf