ZULFAQAR Journal of Defence Science, Engineering & Technology
e-ISSN: 2773-5281
Vol. 7, Issue 1 (2024)
DOI: https://doi.org/10.58247/jdset-2024-0701-04
Journal homepage: https://zulfaqarjdset.upnm.edu.my

DEVELOPMENT AND VALIDATION OF RESEARCH INSTRUMENTS FOR BIG DATA ANALYTICS APPLICATION MODEL

Zam Zarina Abdul Jabar[a], Muslihah Wook[a*], Omar Zakaria[a], Suzaimah Ramli[a], Nor Afiza Mat Razali[a]

[a] Department of Computer Science, Faculty of Defence Science & Technology, National Defence University of Malaysia, Sg. Besi Camp, 57000 Kuala Lumpur, Malaysia

ARTICLE INFO

ARTICLE HISTORY
Received: 15-11-2023
Revised: 20-01-2024
Accepted: 14-02-2024
Published: 30-06-2024

KEYWORDS
Big data analytics
Big data
Big data traits
Data quality dimension
Content validation

ABSTRACT

As human beings delve deeper into the information age, the rapid growth of big data analytics (BDA) can be seen among various private and public organisations. Most organisations have implemented BDA because it provides deep data-driven insights into competitive advantages that can be used for planning their future course of action. However, when these organisations try to use and manage big data (BD), they would find that obtaining quality and actual data from the massive, diverse, and sophisticated sets of data could become a big challenge. These sets of data have numerous traits (characteristics) that require efficient mechanisms for evaluating the quality of the big data involved. This study, therefore, aimed to examine the relationships between big data traits (BDTs) and data quality dimensions (DQDs) for the implementation of BDA, specifically in the Malaysian public sector. In order to carry out this study, a research instrument will need to be developed and validated. As the validity of a research instrument is established, the data collected throughout the data collection process is strengthened, allowing for increased confidence in the survey findings. Hence, this article outlines the development and validation of the research instrument. The developed instrument was validated using the Content Validation Ratio (CVR) and Content Validation Index (CVI) methods. This research found that 54 indicators were accepted and included in the final questionnaire.

1.0 INTRODUCTION

Data has become a vital asset for most organisations since the dawn of the internet era. Currently, the amount of data obtained from various sources is increasing rapidly. Due to the large and diverse amount of data, it is impossible to analyse using traditional methods. This type of diverse data is referred to as Big Data (BD) [1], and the approach for analysing it is referred to as Big Data Analytics (BDA) [2]. At present, with the rapid advancement of big data technology and analytics solutions, BDA is becoming a hot method for analysing data in most fields, such as social media, machine learning, and artificial intelligence [3]. BDA is also used by various organisations, such as commercial businesses, banking, health care, education, government organisations [4], insurance [5], and cyber security [6]. Top-quality BD is imperative for identifying problems, improving a process, increasing productivity, supporting efficient customisation, making decisions, and optimising solutions [7-9]. In most situations, the value of data is determined based on applicability and this requirement is the most crucial factor for organisations to consider once investing in BD [10]. However, vast quantities of data do not automatically guarantee quality [10]. Despite the vast quantity, speed of delivery, and variety of data kinds, the quality of BD is far from optimal and remains unsatisfactory [11-12]. Consequently, the low quality of data would be detrimental to the organisation, including customer dissatisfaction and increased operational costs. Hence, data quality has become a major concern in BDA [4].

*Corresponding Author | Wook, M. | muslihah@upnm.edu.my

The factors that could affect BD quality are divided into four categories, namely, data, management, service, and user [3]. The data category focuses on data quality factors, such as accuracy, currency and timeliness, correctness, consistency, cohesion, coherence, validity, precision, usability, security, completeness, accessibility, accountability, complexity, redundancy, minimality, compactness, conciseness, and scalability. The management category refers to how data are managed. Meanwhile, the service category specifies how data will be used and analysed. Finally, the user category refers to how BD will be presented and visualised to the targeted user [3]. Among all categories, this study focuses on data quality factors. Data quality typically refers to whether the data meets users' expectations or is appropriate for the intended application [13]. The Data Quality Dimension (DQD) is an attribute used to manage data requirements into groups. It gives a way to keep track of and evaluate the quality of the data [14]. Organisations must not only encounter the difficulty of producing high-quality BD, but also consider the qualities of BD that may have an impact on the accuracy of the BDA [15]. Researchers have described BD using many different words, such as characteristics, traits, features, and attributes. Most researchers discuss the nature of BD in terms of its characteristics. Kitchin and McArdle (2016), on the other hand, used traits to characterise BD since datasets labelled as BD are frequently classified by comparable traits [16]. So, this study will discuss BD traits since the main goal was to find traits that can affect data quality.

With the increase in data production, BD has been revealing new traits that make quality assessment more difficult [13]. There is an emerging need for more studies on initiatives and solutions to improve BD quality, as they are still in their infancy [13]. Emphasis should be on the relationships between BDTs and DQDs, as in the context of BD, data traits make it more likely that the data quality will degrade [13]. Discussions regarding the relationships between BDTs and DQDs are limited. Noorwali (2016), Taleb (2021), and Wahyudi (2018) are among the scholars who studied the relationships between BDTs and DQDs in private organisations (business and finance) [13][15][17]. Given the lack of literature in this field, this study has focused on these relationships in the context of public sector BDA application, with a model of BDA application as an output. Several methods were selected to design and validate the instruments developed in this study, which will be described in this article. Section 2.0 presents the six phases of the research design, Section 3.0 presents the research instrument development process, Section 4.0 instrument validation, and Section 5.0 concludes the work.

2.0 RESEARCH DESIGN

The research design consists of six phases, as illustrated in Figure 1. The SEM research design provided by [18] has been used for this study.

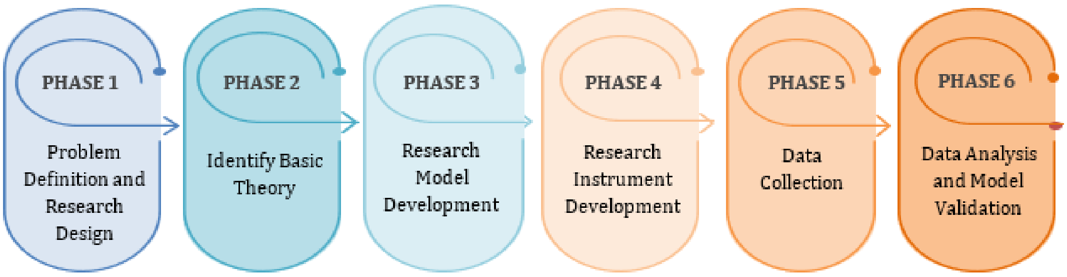
PHASE 1
Problem Definition and Research Design
PHASE 2
Identify Basic Theory
PHASE 3
Research Model Development
PHASE 4
Research Instrument Development
PHASE 5
Data Collection
PHASE 6
Data Analysis and Model Validation

Figure 1. Research design

Phase 1

As shown by the research design depicted in Figure, various issues pertinent to this study were found during the first phase. Following that, objectives and questions for the study were defined to ease the identification of variables and statistical methodologies, as well as the construction of hypotheses.

Phase 2

Subsequently, the second phase involved identifying theories, the definition of concepts, and reviewing and analysing previous research through a literature review. Prior to the development of the conceptual model, theories and concept definitions were identified through analysis and review of previous studies. To better understand the relationships between BDTs and DQDs in implementing BDA, Resource-based View (RBV), Organisational Learning Theory (OLT), Knowledge-based View (KBV), and Data Quality Framework (DQF) have been identified as important theories and frameworks for this study. Ghasemaghaei and Hassanein (2019), and Liu et al. (2017) have asserted that the DQF developed by Wang (1996) is the fundamental framework used by most researchers for identifying high-quality data across a wide range of fields and disciplines [19-21]. Meanwhile, several constructs related to BDT and DQD have been identified and selected based on their definition and field of study. The list of BDTs that have a relationship with DQD, according to Wahyudi (2018) and Merino (2016), is shown in Table 1 [17][22].

Table 1. Relationships between BDTs and DQDs

Data quality dimensions		Big data traits				
		Variety	Velocity	Veracity	Validity	Volume
Accessibility	Accessibility	[17], [22]				
	Ease of operation	[17]				
Contextual	Timeliness	[17], [22]				[22]
	Completeness	[17]	[22]		[17]	[22]
Intrinsic	Accuracy	[17]	[22]	[17]		
	Believability	[17]		[17]		
Representational	Understandability	[17]				
	Consistency	[17]				[22]

In addition to the five Vs presented in Table 1, this model includes an additional three Vs, based on the definition and their impact on the data quality. These additional Vs are value, volatility, and variability. Value is one of the most significant characteristics of BD since data without an actual value can have an adverse effect on data quality, making it useless and ineffective [23]. Volatility refers to the life span of the data; either the data remain valid and should be stored, or they become irrelevant and outdated. Furthermore, unrelated data can impair insight and decision-making [23] capabilities. Finally, variability refers to inconsistent data flow from multiple sources to the BD database and inconsistencies in the data [23]. Inconsistent data can lead to the occurrence of different data qualities.

Phase 3

The preliminary conceptual model for this study is proposed in the third phase, as depicted in Figure 2.

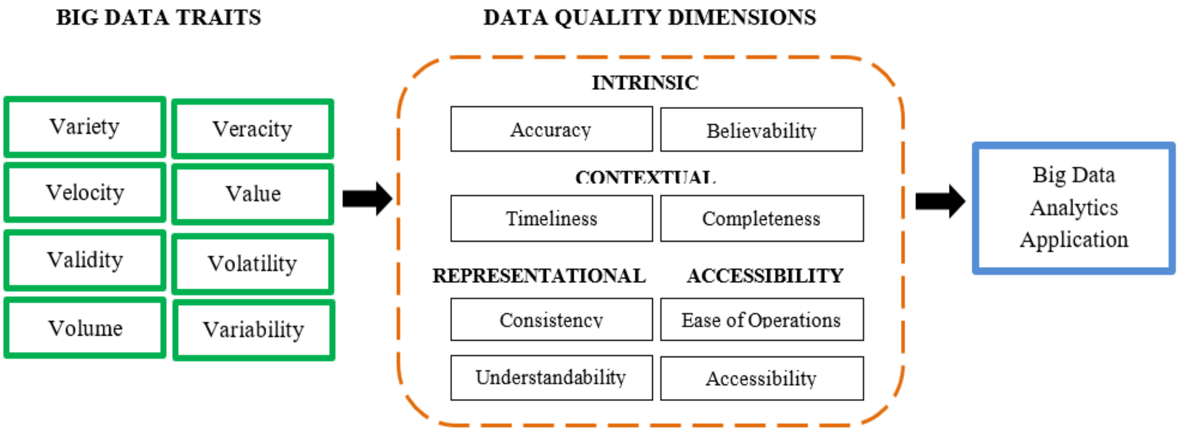

Figure 2. Preliminary Conceptual model

The conceptual model has been validated by four experts from four different government agencies in Malaysia, namely, MAMPU, the Ministry of Health (MOH), the Public Service Department (JPA), and the National Hydraulic Research Institute of Malaysia (NAHRIM). Expert opinions were evaluated using the Interquartile Range (IQR) because it is commonly used for this purpose [24]. Von der Gracht (2012) recommended keeping factors with IQRs of one or less (≤ 1) and eliminating those with IQRs of one or more (≥ 1) [24]. Table 2 summarises the findings of expert analysis using IQR calculations.

Table 2. Results of expert analysis

Factor	E1	E2	E3	E4	Median	Q1	Q3	IQR (Q3-Q1)	Result
Volume	7	7	7	7	7	7	7	0	Accepted
Velocity	7	5	6	6	5.5	5.5	6.5	1	Accepted
Variety	7	7	6	7	7	6.5	7	0.5	Accepted
Veracity	7	6	7	6	6.5	6	7	1	Accepted
Value	5	5	7	7	6	5	7	2	Eliminate
Validity	5	7	7	7	7	6	7	1	Accepted
Volatility	6	3	7	6	6	4.5	6.5	2	Eliminate
Variability	6	4	7	6	6	5	6.5	1.5	Eliminate
Intrinsic	7	7	7	7	7	7	7	0	Accepted
Contextual	7	7	7	7	7	7	7	0	Accepted
Representational	7	6	7	6	6.5	6	7	1	Accepted
Accessibility	7	6	7	6	6.5	6	7	1	Accepted

A conceptual model was developed after examining and enhancing all expert panel views. These recommendations were supported by a solid theoretical foundation. Apart from the theories used in previous studies and comments from experts, this study has also considered the literature reviews of previous researchers. Based on these reviews, several ways to improve the quality of data, one of which is data-driven. Data-driven is an approach for enhancing data quality by significantly affecting its value. This strategy needs to be implemented during the pre-processing stage [13-14]. Hence, to utilise the full capability of BD, a data-driven culture needs to be developed in the organisation. With the implementation of the data-driven culture, BD awareness should grow throughout the organisation, and BD transformation should be experienced by the entire organisation [25]. This study has accepted the data-driven culture as a moderator that could improve the relationships between DQDs and BDA applications. Hence, the conceptual model of the study is depicted in Figure 3.

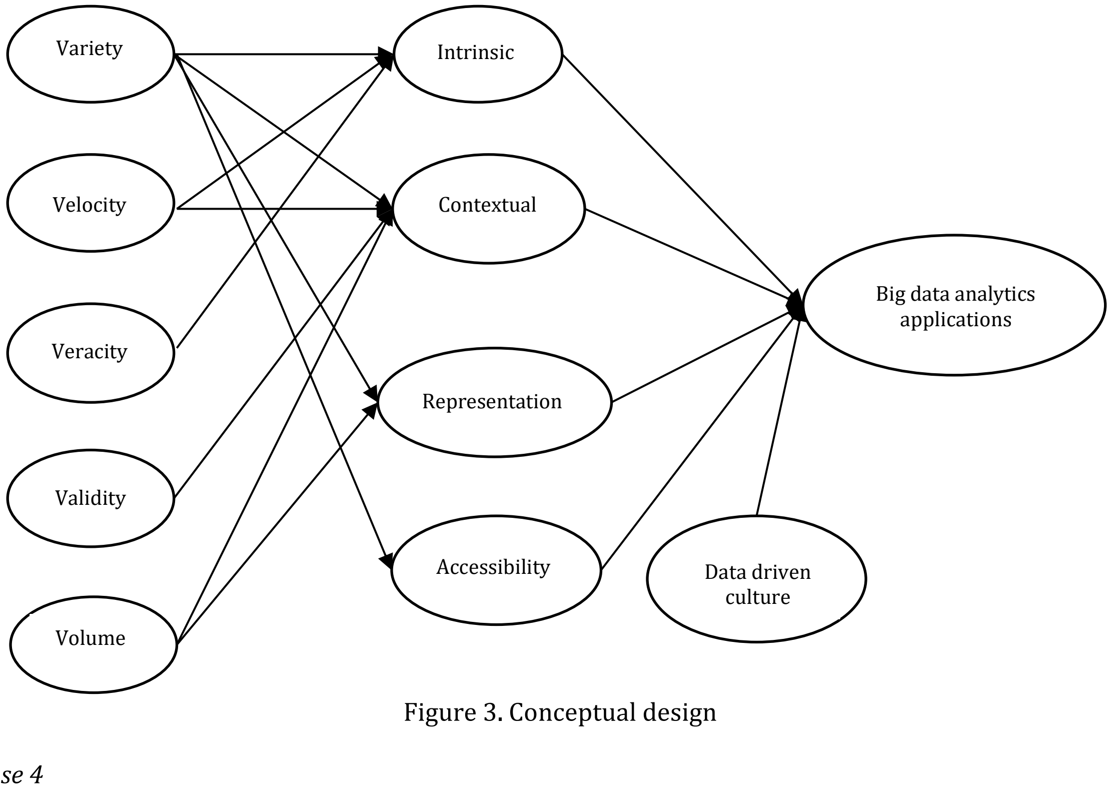

Figure 3. Conceptual design

Phase 4

Phase four began with the formulation of the study hypotheses prior to questionnaire preparation. The primary method for gathering quantitative data in this study was through a questionnaire. A questionnaire makes it possible to collect quantitative data in a consistent and logical way, making sure that the results are also consistent and logical so that they can be analysed. Questionnaires should always be tied to the research objectives, and it should be apparent from the beginning how the data will be utilised [26]. In order to develop the questionnaire, a comprehensive literature review was conducted with the intention of fully grasping the concept being studied.

Phase 5

The fifth phase of the research design was focused on data collection. The study population and sampling method were determined prior to initiate data collection activities. According to [27], a sample is a subset or a small number of elements taken from a population. The sampling procedure can be divided into three stages: i) establishing a clear target audience; ii) selecting the sampling frame; and iii) selecting a sampling method [27]. Hence, based on the objectives, the population for this study was chosen among the respondents at a designated agency in the Malaysian public sector. During the second stage of the sampling procedure, a sample was selected from the population to be included in the development of the sampling framework. A sampling framework was required to ensure a sufficient number of samples and could be re-generalised to the study population [28-29]. Hence, the sampling framework for this study was an officer in the public sector with an information background and basic knowledge of BD.

The final stage in the sampling procedure was to select a sample from the sampling framework using a well-defined sampling technique. Hence, this study employed a probability sampling strategy that incorporated stratified sampling. Probability sampling is sometimes referred to as 'random sampling'. This sampling technique ensures that all elements are equally capable of being selected [30]. Thus, the study sample was chosen by implementing stratified sampling depending on the respondent’s job grade and job scheme. The link to the questionnaire was delivered to the respondents through emails once the sampling procedure was completed.

Phase 6

In phase six, the collected data were analysed to test the established research hypotheses. Data analysis was performed to determine the relationships between the variables, as proposed in the conceptual model. Since this study has been based on the positivism philosophy, which led to the quantitative data collection, statistical techniques were used to analyse the data [31]. This study conducted a descriptive analysis using SPSS (ver. 26), with frequency and percentage values to explain the characteristics of demographical variables (gender, Ministry, job scheme, and job grade). Mean values and standard deviations were used to elucidate the 11 key variables that formed the study model. The inferential analysis was based on multivariate data analysis because of the numerous relationships between variables in the conceptual model. Structural Equation Modelling (SEM) is one of the techniques for analysing various relationships between dependent and independent variables [32]. SEM can simultaneously analyse relationships between all variables in a systematic and comprehensive manner based on a combination of factor analysis techniques and multiple regression analysis [32]. SEM allows non-measurable variables to be directly constructed and theories tested flexibly [33]. Thus, SEM was used in this study because it can empirically focus on the testability of a theory (a study based on analytical experiments and practical experiences) [18].

3.0 INSTRUMENT DEVELOPMENT

The process of developing a questionnaire begins with conducting a detailed literature review to get a better understanding of the overall concept of the research project. The construction of the questionnaire for this study is based on the method that was introduced by Sekaran and Bougie (2016), and it is applied to four primary processes, including planning and strategy, identifying the content of the questionnaire, arranging and designing the questionnaire, as well as validity and reliability, as shown in Figure 4 [29].

Choosing a suitable questionnaire was part of the planning and strategising process. The questionnaire used in this study was closed and organised to ensure that participants considered each item before responding. This kind of questionnaire was selected due to its ease of use and its relatively easy questions and answers. The questionnaire was distributed online in this study, which required respondents to fill out a web form that can be accessed through a link sent to them via email. Content validation was conducted once the content of the questionnaire had been developed to make sure the constructs were valid, clear, and reflected the content of the instrument [34-35].

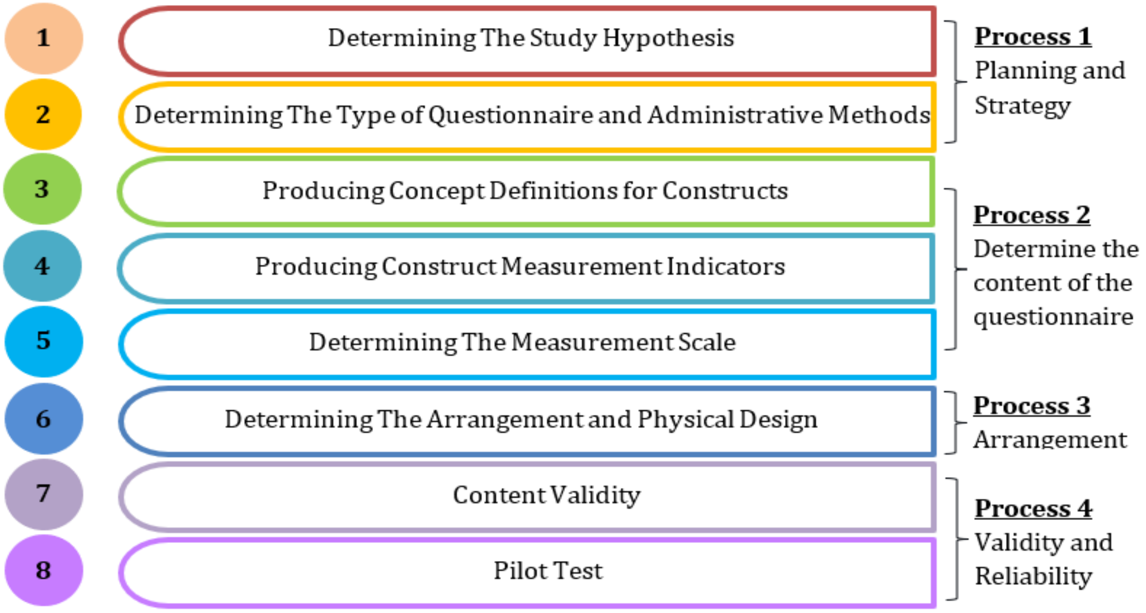
1
Determining The Study Hypothesis
2
Determining The Type of Questionnaire and Administrative Methods
3
Producing Concept Definitions for Constructs
4
Producing Construct Measurement Indicators
5
Determining The Measurement Scale
6
Determining The Arrangement and Physical Design
7
Content Validity
8
Pilot Test
Process 1
Planning and Strategy
Process 2
Determine the content of the questionnaire
Process 3
Arrangement
Process 4
Validity and Reliability

Figure 4. Questionnaire development process

Step 1: Determining the study hypothesis

Establishing a research hypothesis is essential to completing the study. It is impossible to fully understand the research process and draw comprehensive findings without first developing a research hypothesis. Research hypotheses must be expressed explicitly before the questionnaire is produced in order to ascertain the most efficient means of completing the questionnaire within the allotted time and resources. Research hypotheses can determine the connection between the researched variables, as well as the information required and the source of that knowledge [36]. Table 3 summarises the 15 research hypotheses, which are separated into 14 hypotheses for the main model and one hypothesis for the moderator.

Table 3. Research Hypotheses

	Hypothesis (H)	
H1	Variety has a significant influence on intrinsic	Var -> Int
H2	Variety has a significant influence on contextual	Var -> Con
H3	Variety has a significant influence on representational	Var -> Rep
H4	Variety has a significant influence on accessibility	Var -> Acc
H5	Velocity has a significant influence on intrinsic	Vel -> Int
H6	Velocity has a significant influence on contextual	Vel -> Con
H7	Veracity has a significant influence on intrinsic	Ver -> Int
H8	Validity has a significant influence on contextual	Val -> Con
H9	Volume has a significant influence on contextual	Vol -> Con
H10	Volume has a significant influence on Representational	Vol -> Rep
H11	Intrinsic has a significant influence on BDA application	Int -> BDAA
H12	Contextual has a significant influence on BDA application	Con -> BDAA
H13	Representational has a significant influence on BDA application	Rep -> BDAA
H14	Accessibility has a significant influence on BDA application	Acc -> BDAA
	Moderator	
H15	Data-driven culture moderates the relationship between DQD and BDAA	DDC -> BDAA

Step 2: Determining the type of questionnaire and administrative methods

In this study, participants are requested to fill out a closed-ended questionnaire after carefully reading the statements and questions. Because of its manageability and the clarity of its questions and answers, this format was selected for the study. It's important to decide in advance if the questionnaire will be sent via regular mail, sent via email, administered via phone, or conducted in a face-to-face interview. This study used emails, telephones and face-to-face interviews.

Step 3: Producing concept definitions for constructs

A concept definition for the construct ensures that it can represent the idea under consideration and is unique [37]. Questionnaire development can be done by reusing earlier constructs, but they must be modified to fit the study [29]. Existing constructs have been tested for validity and reliability and can be used to compare current and past study findings [38].

Step 4: Producing construct measurement indicators

Indicators are created based on previous research in related domains such as the big data traits, which is the study conducted by Ghasemaghaei & Calic (2019b), Arockia Panimalar et al. (2017), Côrte-Real et al. (2020) and Cai & Zhu (2015), the dimensions of data quality framework which is the study conducted by Wang and Strongs (1996), Wahyudi et al. (2018) and Ghasemaghaei & Calic (2019a), Big Data Analytics Applications which study is undertaken by Verma et al. (2018) and Akter et al. (2017) and data-driven culture which study is conducted by Gupta & George (2016), AL-Ma'aitah (2020) and Shamim et al. (2020) [7, 17, 19, 34, 39-46].

According to MacKenzie (2011), indicators with a low weighting value in a construct imply that the construct is not measured accurately (validity does not exist); therefore, identifying indicators should rely on indications with a high weighting value in the construct [37]. According to Hair et al. (2017), in order for a construct to be stable, it must have at least three indicators [32]. As a result, the minimal indicator that this study uses for each construct is equal to or greater than 4. In this study, 54 indicators were created to measure 14 variables and one moderator.

Step 5: Determining the measurement scale

In this study, a numerical rating scale with a 7-point numeric scale was utilised, with labels at the endpoints (1 indicating Strongly Disagree and finishing with 7 indicating Strongly Agree) instead of labels on each number. Using labels at the conclusion is also useful for studies that use correlation analysis, regression, or structural equation models (SEM) to analyse linear relationships between constructs [47]. Since the relationship between the constructs is the focus of the study's hypothesis and structural equation modelling is utilised to assess the relationship, a 7-point scale with a label at the end is deemed adequate for measuring the constructs in this study.

Step 6: Determining the arrangement and physical design of questions

The questions in the questionnaire have been arranged logically to facilitate a response from the respondents. The method used to organise the questions for this study is based on the recommendations made by Creswell (2018), which state that questions should progress from broad to narrow, avoid touching on controversial topics and not put an undue amount of pressure on the survey's participants [36]. Increases in response time and accuracy might result from better question layouts that make it easier for respondents to understand the concepts being assessed.

Step 7: Content validity

A content validation process needs to be carried out throughout instrument development in order to verify that the constructs developed are accurate, comprehensible, and true to their respective content [11, 35].

Step 8: Pilot test

The final phase of questionnaire preparation is the pilot test, which was conducted on a smaller scale with a subset of respondents who were representative of the target demographic [29]. The purpose of the pilot study was to assure that the measurement of indicators on constructs was accurate and valid, as well as to identify any potential issues with the questionnaire's design that may have been missed [27]. Table 4 details the findings of the test that was conducted as part of the pilot test.

Table 4. Results of pilot test analysis

Construct	No. of indicator	Reliability (Cronbach Alpha)
Variety	4	0.835
Velocity	4	0.936
Veracity	4	0.900
Validity	4	0.928
Volume	4	0.930
Intrinsic	6	0.918
Contextual	6	0.942
Representational	6	0.928
Accessibility	6	0.951
Big Data Analytics Application	4	0.940
Data-Driven Culture	6	0.938

3.0 INSTRUMENT VALIDATION

Validity refers to the degree to which a measurement (indicator) is used to accurately measure or reflect concepts and determine whether the selected indicator corresponds to a construct [48]. There are three types of validity, as described in Table 5.

Table 5. Types of validity

Types of Validity	Descriptions
Content validity	The accuracy of a research instrument gathers all aspects of a construct.
Construct validity	The intended construct is measured by a research instrument (or tool).
Criterion validity	A research instrument is connected to others that measure the same variables.

The quantitative nature of this study necessitated the use of content validity to determine whether the developed instrument could adequately cover the variables. This study adapts two methods for content validity, as described in Table 6.

Table 6. Content Validity Method

Methods	Description
Content Validation Ratio (CVR)	In this method, each construct is evaluated by a group of experts using a scale of three or five. Subject matter experts can provide additional viewpoints. In many cases, the quantity of experts is determined by the practicality of the study rather than by the number of experts. The CVR calculation criteria were developed by Lawshe (1975) [49].
Content Validation Index (CVI)	This method involves the assessment of constructs by a group of experts using a scale of four - “1 = irrelevant”, “2 = somewhat relevant”, “3 = relevant”, and “4 = highly relevant”. Expert panels are set between three to ten [50].

The content validation process of this study involved four main steps as shown in Figure 5.

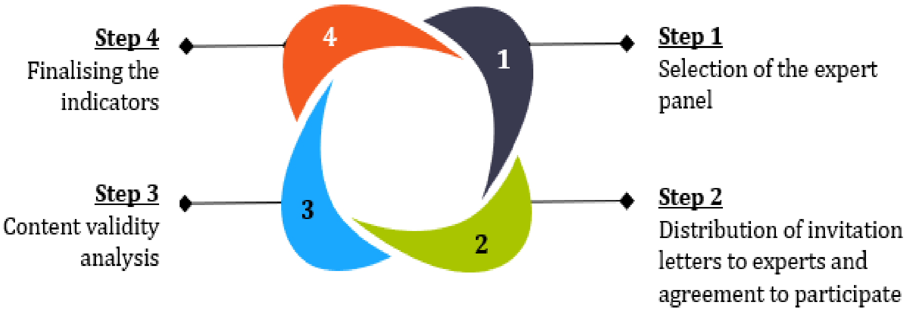
Step 4
Finalising the indicators
4
1
Step 1
Selection of the expert panel
Step 3
Content validity analysis
3
2
Step 2
Distribution of invitation letters to experts and agreement to participate

Figure 5. Content validation process

Step 1: Selection of the expert panel

The selected experts consist of professionals experienced in the implementation of BDA. The inclusion criteria are based on the following criteria: (i) Experience and involvement in BDA implementation in the public sector for at least five years; (ii) Knowledgeable in BDA; and (iii) Knowledgeable in theory, statistics or construct measurement. Table 7 details the backgrounds of the identified experts. All of the expert panel's suggestions were considered, and an enhanced conceptual model was obtained as a result.

Table 7. Expert panel background

Expert ID	Role in organisation	Work experience (years)	Expertise	Agency
E1	Senior IT Officer	25	Practitioner	National Institute of Public Administration (INTAN)
E2	Senior IT Officer	17	Practitioner	Malaysian Administrative Modernisation and Management Planning Unit (MAMPU)
E3	Senior Lecturer	20	Academic	National Defence University of Malaysia
E4	IT Officer	15	Practitioner	Ministry of Health (MOH)

Step 2: Distribution of invitation letters to experts and agreement to participate

Individual invitations to the panels were delivered through email and made over the phone in order to secure consent for participation in content validation sessions. Each expert was assured that their information would remain private and be utilised exclusively for research. The questionnaire was delivered to the experts through email after they gave their permission.

Step 3: Content validity analysis

Content Validity Ratio or Content Validation Ratio (CVR) was implemented to measure validity at the item level. For this purpose, panel feedback was statistically analysed using Microsoft Excel software. The consensus among the expert panel was measured using the CVR calculation introduced by Lawshe (1975)[49]. Answers “3” and “4” were considered relevant, while answers “1” and “2” were considered irrelevant. The following equation was used in this study:

$$CVR\ value = (2Ne/N) - 1 \tag{1}$$

* Ne = Number of experts who gave relevant answers, "3 = Agree" and "4 = Strongly Agree"
* N = Total number of experts

This equation is further explained in Table 8.

Table 8. Explanation for the equation

Equation	Explanation
If all experts give answers of "3" and "4".	CVR value is 1.00 (all agree).
If more than half of the panel (> 50%), but less than all (< 100%) answered "3" or "4".	CVR values are positive, ranging from 0.00 to 0.99.
If less than half (< 50%) of the expert panel give answers of "3" or "4".	Negative CVR values are shown.

The acceptance criteria for each indicator (minimum CVR value) depended on the panel's total number of experts. The minimum CVR value was set at a probability of 5% (p = 0.05) divided by the number of experts who participated in the study (as shown in Table 5) [49]. Given that four experts were involved, the minimum value of CVR received was 1.00. For the final questionnaire, each indicator with a value of 1.00 or higher was accepted, whereas a value of 1.00 or less was dismissed and excluded. Following the implementation of the CVR, and once the indicators have either been accepted or rejected, the content validation index (CVI) was calculated. CVI is a test to validate a questionnaire, which involves the calculation of values at the construct level [49]. CVI was calculated by averaging the CVR values of each indicator (obtained during the CVR calculation) for each construct. The following equation was used to calculate the CVI values for the Variety (VAR) construct [49]:

$$CVI_i = \frac{\sum_{j=1}^{n} CVR_i}{n} \quad (2)$$

$$i = 1, 2, \ldots.k, \qquad k\ is\ total\ number\ of\ construct$$

and

$$j = 1, 2, \ldots.n, \qquad n =\ total\ number\ of\ indicatiors\ received$$

$$CVI_{VAR} = \frac{\sum_{j=1}^{4} CVR_j}{4} = \frac{1+1+1+1}{4} = \frac{4}{4} = 1$$

Next, the following equation was used to assess the overall validity of the questionnaire:

$$Overall\ questionnaire\ validy = \frac{Total\ CVI\ value}{No.of\ constructs} \quad (3)$$

Step 4: Finalising the indicators

This study found 54 indicators, all accepted into the final questionnaire. Table 9 lists the final CVR calculation results.

Table 9. Analysis results of CVR

Construct	ID	Indicator	Ne	CVR	Results
Variety (VAR)	VAR1	My organisation analyses many types of data.	4	1	Accepted
	VAR2	My organisation uses several different sources of data to gain insights (e.g., email, web data, user-generated content, scientific data, and transactional data).	4	1	Accepted
	VAR3	My organisation examines data from a multitude of sources.	4	1	Accepted
	VAR4	My organisation processes a variety of data formats (e.g., text, audio, video, and images).	4	1	Accepted
Velocity (VEL)	VEL1	In my organisation, we get new data in the fastest time.	4	1	Accepted
	VEL2	In my organisation, we analyse data as soon as we receive them.	4	1	Accepted

Construct	ID	Indicator	Ne	CVR	Results
	VEL3	In my organisation, we analyse data speedily.	4	1	Accepted
	VEL4	In my organisation, we are fast in exploring our data.	4	1	Accepted
Veracity (VER)	VER1	My organisation analyses high actual data.	4	1	Accepted
	VER2	My organisation processes data that are believable.	4	1	Accepted
	VER3	My organisation deals with truth data.	4	1	Accepted
	VER4	My organisation processes data that are reliable.	4	1	Accepted
Validity (VAL)	VAL1	The data analysed by my organisation must be correct.	4	1	Accepted
	VAL2	The data analysed by my organisation must be valid.	4	1	Accepted
	VAL3	The data analysed by my organisation must be certain.	4	1	Accepted
	VAL4	The data analysed by my organisation must be precise.	4	1	Accepted
Volume (VOL)	VOL1	My organisation explores a substantial quantity of data.	4	1	Accepted
	VOL2	My organisation analyses a large amount of data.	4	1	Accepted
	VOL3	My organisation scrutinises abundant volumes of data.	4	1	Accepted
	VOL4	My organisation uses a great deal of data.	4	1	Accepted
Intrinsic (INT)	INT1	In my organisation, the BDA used produce accurate information.	4	1	Accepted
	INT2	In my organisation, there are a few errors in the data obtained from the BDA.	4	1	Accepted
	INT3	In my organisation, the accuracy of the data is important before they can be analysed.	4	1	Accepted
	INT4	In my organisation, the believability of the data is important before they can be analysed.	4	1	Accepted
	INT5	In my organisation, the data used in BDA are trustworthy.	4	1	Accepted
	INT6	In my organisation, the data used in BDA are unbiased.	4	1	Accepted
Contextual (CON)	CON1	In my organisation, the BDA used provides a complete set of data.	4	1	Accepted
	CON2	In my organisation, the BDA used produces comprehensive data.	4	1	Accepted
	CON3	In my organisation, the BDA used provides all the data needed	4	1	Accepted
	CON4	In my organisation, the BDA used provides sufficiently timely information.	4	1	Accepted
	CON5	In my organisation, the BDA used updates data regularly.	4	1	Accepted
	CON6	In my organisation, the BDA used provides current information for our work.	4	1	Accepted
Representational (REP)	REP1	In my organisation, the data used in BDA application contain adequate details.	4	1	Accepted
	REP2	In my organisation, the data used in BDA application are compatible with previous data.	4	1	Accepted

Construct	ID	Indicator	Ne	CVR	Results
	REP3	In my organisation, the data used in BDA applications are well-formatted.	4	1	Accepted
	REP4	In my organisation, the definitions, value domains, format, and data remain the same after processing.	4	1	Accepted
	REP5	In my organisation, during a certain time, data that have been processed remain consistent.	4	1	Accepted
	REP6	In my organisation, the data used in BDA applications are realistic.	4	1	Accepted
Accessibility (ACC)	ACC1	In my organisation, the data used in BDA can be easily obtained.	4	1	Accepted
	ACC2	In my organisation, the data used in BDA can be easily found.	4	1	Accepted
	ACC3	In my organisation, the data used in BDA can be easily downloaded.	4	1	Accepted
	ACC4	In my organisation, the data used in BDA can be easily available.	4	1	Accepted
	ACC5	In my organisation, the data used in BDA can be easily interpreted.	4	1	Accepted
	ACC6	In my organisation, the data used in BDA can be easily combined with other information.	4	1	Accepted
Big Data Analytics Application (BDAA)	BDAA1	I am excited about using the BDA in my organisation.	4	1	Accepted
	BDAA2	It is my wish to see the full utilisation and deployment of BDA in my organisation.	4	1	Accepted
	BDAA3	I am satisfied with the information generated from BDA in my organisation.	4	1	Accepted
	BDAA4	I am contented with the use of BDA in my organisation.	4	1	Accepted
Data-Driven Culture (DDC)	DDC1	In my organisation, we consider data in BDA applications as a significant asset.	4	1	Accepted
	DDC2	In my organisation, we make decisions based on facts rather than perception.	4	1	Accepted
	DDC3	In my organisation, we often provide useful data across departments.	4	1	Accepted
	DDC4	In my organisation, we have a culture of data-driven work.	4	1	Accepted
	DDC5	In my organisation, we depend on data rather than instinct when making decisions.	4	1	Accepted
	DDC6	In my organisation, we constantly train our employees to make data-driven decisions.	4	1	Accepted

The CVI values for the other constructs have been calculated, as shown in Table 10, using the same calculation method.

Table 10: Analysis results of CVI

ID	No. of construct	Total of significant construct (CVR > 0.99)	CVI
B1	4	4	1
B2	4	4	1
B3	4	4	1
B4	4	4	1
B5	4	4	1
C1	6	6	1
C2	6	6	1
C3	6	6	1
C4	6	6	1
D	4	4	1
E	6	6	1

Validity should be at a minimum of 0.99 to ensure that the questionnaire is accurate and trustworthy [49]. According to the CVI calculation results, the questionnaire in this study was legitimate and credible, with a CVI value of 1.000.

4.0 CONCLUSIONS

This paper has presented an overview of the research design and instrument validation process conducted in this study. To ascertain the quality and effectiveness of the instrument, content validity became a critical part of the development process. The CVR and CVI methods used in this study have clear procedures and accurate equations for calculation, which were simple to put into practice and understand. Once the content validation process was performed, the developed instrument became a reliable tool for measuring the relationships between BDTs and DQDs in BDA applications. Additionally, the research findings have expanded the options for utilising the measurement instrument to assess BDTs and DQDs in BDA application.

5.0 CONFLICT OF INTEREST

The authors declare no conflicts of interest.

6.0 AUTHORS CONTRIBUTION

Abdul Jabar, Z. Z. (Conceptualisation; Methodology; Validation; Formal analysis; Data curation; Resources; Visualisation; Writing - original draft; Writing - review & editing;)
Wook, M. (Conceptualisation; Writing - review & editing; Project administration; Supervision)
Zakaria, O. (Conceptualisation; Writing - review & editing; Project administration; Supervision)
Ramli, S. (Conceptualisation; Writing - review & editing; Project administration; Supervision)
Mat Razali, N. A. Conceptualisation; Writing - review & editing; Project administration; Supervision)

7.0 ACKNOWLEDGEMENTS

The authors fully acknowledge Ministry of Higher Education Malaysia (MOHE) and National Defence University of Malaysia (NDUM) which makes this important research viable and effective.

REFERENCES

[1] Acharjya, D. P., & Ahmed, K. (2016). A survey on big data analytics: challenges, open research issues and tools. International Journal of Advanced Computer Science and Applications, 7(2), 511-518.

[2] Alswedani, S., & Alsherbeeni, M. (2020). Big data analytics: importance, challenges, categories, techniques, and tools. International Journal of Advanced Trends in Computer Science and Engineering, 9(4), 5384-5392.

[3] Abdallah, M. (2019, February). Big data quality challenges. In 2019 International Conference on Big

Data and Computational Intelligence (ICBDCI) (pp. 1-3). IEEE.

[4] Gao, J., Xie, C., & Tao, C. (2016, March). Big data validation and quality assurance--issuses, challenges, and needs. In 2016 IEEE symposium on service-oriented system engineering (SOSE) (pp. 433-441). IEEE.

[5] Arumugam, S., & Bhargavi, R. (2019). A survey on driving behavior analysis in usage-based insurance using big data. Journal of Big Data, 6, 1-21.

[6] Obitade, P. O. (2019). Big data analytics: a link between knowledge management capabilities and superior cyber protection. Journal of Big Data, 6(1), 71.

[7] Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. Data science journal, 14, 2-2.

[8] Surbakti, F. P. S., Wang, W., Indulska, M., & Sadiq, S. (2020). Factors influencing effective use of big data: A research framework. Information & Management, 57(1), 103146.

[9] Kim, G. S. (2020). The effect of quality management and Big Data management on customer satisfaction in Korea's public sector. Sustainability, 12(13), 5474.

[10] Ramasamy, A., & Chowdhury, S. (2020). Big data quality dimensions: a systematic literature review. JISTEM-Journal of Information Systems and Technology Management, 17, e202017003.

[11] Akter, S., & Wamba, S. F. (2016). Big data analytics in E-commerce: a systematic review and agenda for future research. Electronic Markets, 26, 173-194.

[12] Fredriksson, C., Mubarak, F., Tuohimaa, M., & Zhan, M. (2017). Big data in the public sector: A systematic literature review. Scandinavian Journal of Public Administration, 21(3), 39-61.

[13] Taleb, I., Serhani, M. A., Bouhaddioui, C., & Dssouli, R. (2021). Big data quality framework: a holistic approach to continuous quality management. Journal of Big Data, 8(1), 76.

[14] Sidi, F., Panahy, P. H. S., Affendey, L. S., Jabar, M. A., Ibrahim, H., & Mustapha, A. (2012, March). Data quality: A survey of data quality dimensions. In 2012 International Conference on Information Retrieval & Knowledge Management (pp. 300-304). IEEE.

[15] Noorwali, I., Arruda, D., & Madhavji, N. H. (2016, May). Understanding quality requirements in the context of big data systems. In Proceedings of the 2nd International Workshop on BIG Data Software Engineering (pp. 76-79).

[16] Kitchin, R., & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. Big Data & Society, 3(1), 2053951716631130.

[17] Wahyudi, A., Farhani, A., & Janssen, M. (2018). Relating big data and data quality in financial service organizations. In Challenges and Opportunities in the Digital Era: 17th IFIP WG 6.11 Conference on e-Business, e-Services, and e-Society, I3E 2018, Kuwait City, Kuwait, October 30–November 1, 2018, Proceedings 17 (pp. 504-519). Springer International Publishing.

[18] Haenlein, M., & Kaplan, A. M. (2004). A beginner's guide to partial least squares analysis. Understanding statistics, 3(4), 283-297.

[19] Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. Journal of management information systems, 12(4), 5-33.

[20] Ghasemaghaei, M., & Hassanein, K. (2019). Dynamic model of online information quality perceptions and impacts: a literature review. Behaviour & Information Technology, 38(3), 302-317.

[21] Liu, Y., Li, Y., Zhang, H., & Huang, W. W. (2017). Gender differences in information quality of virtual communities: A study from an expectation-perception perspective. Personality and individual differences, 104, 224-229.

[22] Merino, J., Caballero, I., Rivas, B., Serrano, M., & Piattini, M. (2016). A data quality in use model for big data. Future Generation Computer Systems, 63, 123-130.

[23] Khan, N., Naim, A., Hussain, M. R., Naveed, Q. N., Ahmad, N., & Qamar, S. (2019, May). The 51 v's of big data: survey, technologies, characteristics, opportunities, issues and challenges. In Proceedings of the international conference on omni-layer intelligent systems (pp. 19-24).

[24] Von Der Gracht, H. A. (2012). Consensus measurement in Delphi studies: review and implications for future quality assurance. Technological forecasting and social change, 79(8), 1525-1536.

[25] Karaboğa, T., Zehir, C., & Karaboğa, H. (2019). Big Data analytics and firm innovativeness: the moderating effect of data-driven culture. The European Proceedings of Social & Behavioural Sciences.

[26] Roopa, S., & Rani, M. S. (2012). Questionnaire designing for a survey. Journal of Indian Orthodontic Society, 46(4_suppl1), 273-277.

[27] Bhattacherjee, A. (2012). Social science research: Principles, methods, and practices. University of South Florida.

[28] Saunders, M., Lewis, P., & Thornhill, A. (2009). Research methods for business students. Pearson education.

[29] Sekaran, U., & Bougie, R. (2016). Research methods for business: A skill building approach. john wiley & sons.

[30] T Turner, D. P. (2020). Sampling Methods in Research Design. Headache: The Journal of Head & Face Pain, 60(1).

[31] Collis, J., & Hussey, R. (2021). Business research: A practical guide for students. Bloomsbury Publishing.

[32] Leguina, A. (2015). A primer on partial least squares structural equation modeling (PLS-SEM).

[33] Haenlein, M., & Kaplan, A. M. (2004). A beginner's guide to partial least squares analysis. Understanding statistics, 3(4), 283-297.

[34] Akter, S., D'Ambra, J., & Ray, P. (2013). Development and validation of an instrument to measure user perceived service quality of mHealth. Information & management, 50(4), 181-195.

[35] Allahyari, T., Hassanzadeh, R. N., Khosravi, Y., & Zayeri, F. (2011). Development and evaluation of a new questionnaire for rating of cognitive failures at work.

[36] Creswell, J. W., & Creswell, J. D. (2017). Research design: Qualitative, quantitative, and mixed methods approach. Sage publications.

[37] MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. MIS quarterly, 293-334.

[38] Kitchenham, B. A., Pfleeger, S. L., Pickard, L. M., Jones, P. W., Hoaglin, D. C., El Emam, K., & Rosenberg, J. (2002). Preliminary guidelines for empirical research in software engineering. IEEE Transactions on software engineering, 28(8), 721-734.

[39] Ghasemaghaei, M., & Calic, G. (2019). Does big data enhance firm innovation competency? The mediating role of data-driven insights. Journal of Business Research, 104, 69-84.

[40] Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In 2013 international conference on collaboration technologies and systems (CTS) (pp. 42-47). IEEE.

[41] Côrte-Real, N., Ruivo, P., & Oliveira, T. (2020). Leveraging internet of things and big data analytics initiatives in European and American firms: Is data quality a way to extract business value?. Information & Management, 57(1), 103141.

[42] Ghasemaghaei, M., & Calic, G. (2019). Can big data improve firm decision quality? The role of data quality and data diagnosticity. Decision Support Systems, 120, 38-49.

[43] Gupta, M., & George, J. F. (2016). Toward the development of a big data analytics capability. Information & Management, 53(8), 1049-1064.

[44] AL-Ma'aitah, M. A. (2020). Utilizing of big data and predictive analytics capability in crisis management. J. Comput. Sci, 16(3), 295-304.

[45] Verma, S., Bhattacharyya, S. S., & Kumar, S. (2018). An extension of the technology acceptance model in the big data analytics system implementation environment. Information Processing & Management, 54(5), 791-806.

[46] Shamim, S., Zeng, J., Khan, Z., & Zia, N. U. (2020). Big data analytics capability and decision-making performance in emerging market firms: The role of contractual and relational governance mechanisms. Technological Forecasting and Social Change, 161, 120315.

[47] Weijters, B., Cabooter, E., & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. International Journal of Research in Marketing, 27(3), 236-247.

[48] Heale, R., & Twycross, A. (2015). Validity and reliability in quantitative studies. Evidence-based nursing, 18(3), 66-67.

[49] Lawshe, C. H. (1975). A quantitative approach to content validity. Personnel psychology, 28(4).

[50] Lynn, M. R. (1986). Determination and quantification of content validity. Nursing research, 35(6), 382-386.