# Human and AI Trust: Trust Attitude Measurement Instrument

Retno Larasati<sup>a</sup>, Anna De Liddo<sup>a</sup>, Enrico Motta<sup>a</sup>

<sup>a</sup>Knowledge Media Institute, The Open University UK, Walton Hall, Milton Keynes, United Kingdom

# Abstract

With the current progress of Artificial Intelligence (AI) technology and its increasingly broader applications, trust is seen as a required criterion for AI usage, acceptance, and deployment. A robust measurement instrument is essential to correctly evaluate trust from a human-centered perspective. This paper describes the development and validation process of a trust measure instrument, which follows psychometric principles, and consists of a 16-items trust scale. The instrument was built explicitly for research in human-AI interaction to measure trust attitudes towards AI systems from layperson (non-expert) perspective. The use-case we used to develop the scale was in the context of AI medical support systems (specifically cancer/health prediction). The scale development (Measurement Item Development) and validation (Measurement Item Evaluation) involved six research stages: item development, item evaluation, survey administration, test of dimensionality, test of reliability, and test of validity. The results of the six-stages evaluation show that the proposed trust measurement instrument is empirically reliable and valid for systematically measuring and comparing non-experts' trust in AI Medical Support Systems.

Keywords: Human-AI Trust, Human-AI Interaction, Trust Factors, Trust Measurement

# 1. Introduction

Nowadays, the study of trust in Artificial Intelligence (AI) is of great concern in computer science and cognitive systems engineering and is also becoming a hot discussion topic in the media. The call for trustworthy AI was made by formal national institutions around the world (Europe [1], USA [2], China [3]). Trustworthiness in AI systems is increasingly becoming an ethical and societal need. Trust is a crucial factor in all kinds of relationships, such as human-human social interactions or human-AI system interactions. Trust is humans' primary reason for acceptance [4]. A 2018 survey conducted by Intel shows that 36% of patients lack trust in AI and identify trust as a key barrier to AI adoption [5]. People in general are sceptical about AI, and in those instances when AI did something wrong, such as, the Google Photos algorithm classifying black people as gorillas <sup>1</sup>, or the Microsoft chatbot that turned racist in a day <sup>2</sup>, the general public could not understand why the AI did it, and were therefore left only with a sense of distrust toward such systems. This process will only worsen if the public keeps receiving news about the harm of AI. For example, there was an uproar in 2017 over the UK's National Health Service (NHS) allegedly illegally handing 1.6 million patient records to Google's DeepMind, as part of a trial [6]. This news sparked conversations about privacy and ethics. In 2018, a government-backed AI healthcare application, Babylon, also received criticism for the inaccuracies in diagnosis, [7] which brought the medical regulator, the Medicines and Healthcare products Regulatory Agency (MHRA), into the spotlight. These controversies only add to the reported general unwillingness of people to engage with AI when it gets to their healthcare needs [8].

However, research also shows that people who are already AI users tend to easily take algorithmic outputs as accurate and valid and even prefer an algorithmic decision to human advice [9]. People that are more accustomed to AI-facilitated processes have been shown to over-trust AI recommendations, even when the AI systems have been proved to malfunction or when the use of the system caused harm. As AI is increasingly embedded in all sorts of largely adopted systems, research evidence indicates that users tend to over-trust and

 $<sup>\</sup>hline ^{1} \rm https://mashable.com/2015/07/01/google-photos-black-people-gorillas/07/01/google-photos-gorillas/07/01/google-photos-black-people-gorillas/07/01/google-photos-gorillas/07/01/google-photo$ 

 $<sup>^2</sup> https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist$ 

continue to rely on a system even when it malfunctions [10]. This phenomenon is known as automation bias, which occurs when people tend to over-trust and accept system outputs 'as a heuristic replacement of vigilant information seeking and processing' [11, 12]. People often neglect automation bias and tend to trust a system when they think the answer came from an algorithm rather than another person [9]. This misplaced trust (distrust or over-trust) has motivated research on trustworthy AI, whether via developing new forms of explainable AI or AI transparency [13, 14, 15, 16].

A huge challenge into advancing research in this crucial research field is the issue of comparability. Currently, there is no general trust measurement or evaluation method for research in AI trust. In evaluating trust, literature measures users' confidence [17, 18, 19], reliance [20, 21], and also straightforward trust rating [18, 22, 23] amongst other evaluation methods. Aside from the context specific nature of trust [24], the difference in evaluation approach and how trust is measured stemmed from variations in trust definition, which lead to different trust metrics, and therefore prevents from meaningful scientific comparisons. Another important point to note is the broad definition of Artificial Intelligence (AI). To put it simply, Artificial Intelligence is artificially constructed intelligence, it ranges from prediction to recommendation system; to physical materialization, such as, robots and automated machines. This variety results in different trust measurements which rely on questionnaires derived from research in the human-robot trust <sup>3</sup> [25], human-automation trust [26, 27, 28], and humantechnology trust [29], which have been used interchangeably between sub-fields. Combination of existing questionnaires was also implemented to achieve usable measurement tool [30, 31, 32]. However, most of the research has not conducted any validation test [33], or only conducted one test (reliability[30] and predictive validation [32]), to their adapted questionnaires. This lack of measurement

<sup>&</sup>lt;sup>3</sup>The "-" symbol is used as an indicator of the trustor and trustee in the interaction between both. Human-robot indicates interaction between human as a trustor and robot as a trustee in the interaction.

validation raises concerns and can undermine the validity of research findings achieved using said measurements. Moreover, valid measurement instruments play a significant part in the progress of trustworthy AI design, development, and research; how can we design trustworthy AI systems if we cannot measure the effect of our design choices in a reliable and comparable way?

In this study, we developed and validated a trust measure instrument following the psychometric principles, methodological concepts, and techniques in scale development and validation research [34, 35, 36, 37, 38, 39, 40, 41, 42]. The instrument is specifically built for research in human-AI interaction, to measure trust attitude towards AI systems, from a layperson (non-expert) perspective. The use-case we used was in the AI medical support system context (cancer/health prediction). The development (Measurement Item Development) and validation (Measurement Item Evaluation) involved six stages, which are: item development (Section 4), item evaluation (Section 5), survey administration (Section 6), test of dimensionality (Section 7), test of reliability (Section 8), and test of validity (Section 9).

Our work provides three main contributions to the field. Firstly, we demonstrate a methodological approach to develop and thoroughly evaluate a trust measurement for human-AI interaction. Secondly, we propose a trust measurement instrument to evaluate trust in human-AI interaction for laypeople, which future researchers can use and adapt to evaluate their AI systems. Finally, our work contributes to the ongoing conversation regarding trust and trust measurement in human-AI interaction.

# 2. Background and related work

# 2.1. Trust Concepts

Mayer et al. conceptualised trust as a willingness to be vulnerable based on the expectation that another party (the trustee) will perform certain actions that are important to the trust giver (trustor), regardless of the ability to monitor or control the trustee [43]. Although the context of Mayer et al.'s trust concept is human-human trust in organisations, this definition was widely applied and adapted in the context of human-technology trust, such as, trust in automation [44][45], trust in information systems [46][47], and trust in robots[48][49]. This definition of trust, and its adaptations, emphasise the components of competence and vulnerability within it and consider them as factors that can influence trust.

Mayer et al. noted the distinction between trust, as in the factors that influence trust (trust factors), with trust-related behaviour, irrespective of the relationship between these two. Trust as an attitude does not always translate into trust-related behaviours, such as, dependence and [50], and should be measured separately. In contrast, although the concepts of trust and trust factors are easily distinguished, the measurement aspects are quite connected. Since trust is regarded as an attitude, which is a "psychological construct, a mental and emotional entity attached to or characterising a person" [51], it is said to be externally non-observable [43, 52]. Psychological constructs are determined by psychological factors and can therefore be measured using self-reports, attitude scales, or questionnaires, for example, the Likert Scale [53]. The Likert scale is one of the scales that has been widely used and supported by the attitude measurement literature [40, 54, 55]. This study focuses on measuring trust as an attitude through the means of trust factors using Likert scales.

# 2.2. Measuring Human Trust in AI Systems

In general, there is a considerable trust measurement literature, be it behavioural trust (trust-related behaviour) or attitudinal trust (trust as an attitude) <sup>4</sup>. Several disciplines, such as psychology [56] and management [43], have been looking at human trust in technology. In particular, much work has been done investigating trust in human-automation interaction [44, 45, 57, 58, 59] and human other technologies interaction [29, 47]. However, only some of these

<sup>&</sup>lt;sup>4</sup>Since we have established that behavioural trust and attitudinal trust are different, and therefore measured differently, the discussion below covers only the attitudinal trust-related literature

studies have included measurement scales [60, 26, 61, 25]. Several trust measurement scales have become recurring trust scales used in human-AI research. One such scale, Jian et al.'s [26], is reported to be the most cited trust scale in human factors research.

A closer look at the existing measurement scales show that some of these trust measurements are very specific to certain applications. For example, the scale developed by Schaefer [25] refers specifically to the context of human dependence on robots in team settings, with one of the questions: "Does the robot act as part of the team?". Dzindolet et al. [61] developed a trust measurement scale for AI systems, with one of the questions asked being: "How many mistakes do you think you will make over 200 trials?". These questions are highly specific to the task, to their application context, and to their user's expertise. More general scales were also developed [26, 62, 46]. While the scale by Jian et al. was developed for human-automated systems trust [26], the survey questions comprising the scale are very generic, which can be one of the main reasons for its re-usability. The scale dimensions and items were developed through elicitation of trust definitions, followed by cluster and factor analyses. Jian et al.'s proposed scale contains 12 items, with seven items to measure trust and five items to measure distrust. A similarity matrix was used to prove inter-rater reliability. However, no validity for the scale was established. As mentioned previously, most of the research which utilised trust measurement instrument has not conducted any validation test [33], or only conducted one validation test [30, 32], which is not adequate. Measurement instrument or measurement scale should demonstrate internal consistency, and different types of validity: content validity, construct validity, and criterion-related validity, to be sufficient [38].

Madsen and Gregor [62] developed a more generalised measurement for human-computer trust, and tested the measurement reliability and validity. The dimensions used in this measurement were common factors that influence trust. The factors were constructed using the Nominal Group Technique, and then compared to constructs from previous trust research. Through the scale validation process, high internal consistency (Cronbach's alpha ¿ 0.94) and con-

struct validity were established, with poor criterion-related validity. The final scale proposed by Madsen and Gregor consists of five main trust factors: such as, perceived reliability, perceived technical competence, perceived understanding, confidence, and personal attachment, with five questions proposed for each dimension. McKnight developed another trust measurement instrument to capture the trust relationship between users and specific technologies [46]. This scale was developed based on an understanding of trust in the broader context of society and previous research on human-human trust. After conducting several trust-related studies with different information systems, and by covering a large literature, including trust in humans, McKnight defined trust as a construct consisting of three components: propensity to trust, institution-based trust, and trust in specific technologies. The scale proposed by McKnight has eight main dimensions: perceived reliability, perceived functionality, perceived usefulness, situational normality, structural assurance, confidence, and trusting attitude. Three to four questions are proposed to measure each dimension. These measurement instruments show good reliability (Cronbach's alpha ; 0.9), construct validity, and criterion-related validity.

To date, however, there has been no developed trust measurement instrument intended specifically for AI non-expert users/laypeople. Although the more general trust measures described above have been used for laypeople, appropriate modifications to fit the context of our study are still required. Therefore, in the next section, we describe the steps we followed to achieve a suitable yet generalisable measurement instrument for non-expert users' trust in AI medical systems.

# 3. Measurement Instrument Development Process

To develop a sound human trust measurement instrument, we followed recommendations by previous research in psychometric [34, 35]. Six key research stages (Fig 1) were carried out to develop the items, and thoroughly evaluate them (through the assessment of each of the item individually, the overall scale,

and the possible correlation between items). We finally carried out validity and reliability tests. According to the Standards for Educational and Psychological Testing, a guideline approved by American Psychological Association (APA), an appropriate operational definition of the construct a measure aims to represent should include a demonstration of content validity, criterion-related validity, and internal consistency [38]. The complete steps of the method and analysis we carried out is shown in Fig 1. The detailed description of each steps will be described in the next sections.

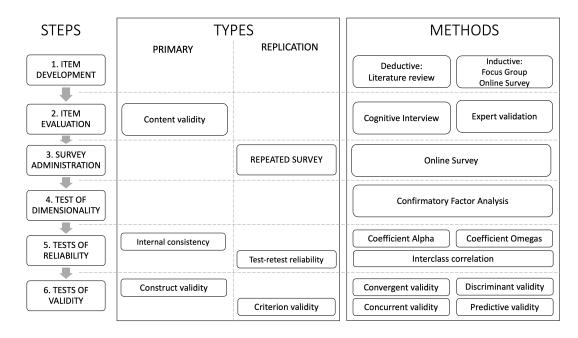


Figure 1: Six Steps Process To Develop a Sound Human Trust Measurement Instrument.

## 1. Item Development

The first step to developing a measurement instrument is to determine the measurement domain that will be used to identify measurement items. A measurement domain (or sometimes referred to as a measurement construct) is the concept or attribute which is the measurement target. In this study, the domain is a set of trust factors, as trust (the target) will be measured using various fac-

tors (attributes) that influence human trust in AI systems. The trust factors and the relevant items were generated using deductive and inductive approaches. As an example, Perceived Understandability is one of the trust factors with one of the relevant items is a statement "I know what will happen the next time I use the system because I understand how it behave". The deductive approach requires a literature review to develop a domain definition with a theoretical foundation. Trust factors proposed in the literature were assessed and selected based on context relevance. The inductive approach requires exploratory research to develop items from dimensions that may not be easily identified in the conceptual basis. A mixed methods study comprising an online survey and a focus group were conducted to help explore other dimensions that may not have been considered previously. The study also helped contextualise and revise the items in the context of a concrete healthcare scenario. It should be noted that, while the measurement instrument was developed to be as general as possible to enable future use in AI system interaction evaluation, the use-case we used was an AI medical support system (cancer/health prediction).

#### 2. Item Evaluation

Once the measurement domains had been defined and the measurement items have been developed, we carried out an evaluation process. In the first evaluation process, the initial set of items was reviewed by experts. The measurement domains (trust factors) were presented and the experts were asked to provide a review and rating for each measurement item. The expert validation resulted in a number of items being reconstructed or being removed from the measurement instrument. The revised items were then further evaluated by the target population, in this case the general public, through cognitive interviews. The cognitive interview assessed how the target population understood the measurement domain and the mental processes behind the answers given from the measurement instrument.

#### 3. Survey Administration

At this stage, the measurement items have been evaluated and revised. We then administered the survey as the main measurement instrument test. The sample size of the survey was carefully considered. The survey questions were presented to the participants after a description of random AI system. The measurement items were designed to quantitatively measure the trust factor, and we used a 7-point Likert scale for this survey. As a replication process, the survey was administered in a repeated manner. Replication was deemed necessary to increase the generalisability of the measurement instrument [35].

#### 4. Dimensionality Test

Dimensionality testing is the first part of measurement instrument evaluation. It evaluates the hypothesised factor or factor structure, and can be performed using confirmatory factor analysis, bifactor modelling, or measurement invariance. Confirmatory Factor Analysis (CFA) was conducted to quantitatively assess the measurement dimensions, thereby confirming that thorough analyses have been conducted to develop the measurement instrument.

# 5. Reliability Test

Reliability is the degree of consistency shown when a measurement is repeated under the same conditions. Different ways to assess reliability are: inter-temporal reliability (test-retest reliability) and inter-item reliability (internal consistency)[63]. Internal consistency was assessed with alpha and omega coefficients, and stability across time was assessed with test-retest reliability.

# 6. Validity Test

Validity is the extent to which "evidence and theory support the interpretation of test scores required by the proposed use of the test" [38]. In this step, we assessed the Construct Validity and Criterion Validity of the instrument. Construct Validity is "the extent to which an instrument assesses a construct of interest and is linked to evidence measuring other constructs in that domain and measures specific real-world criteria" [41] and Criterion Validity is "the extent to which there is a relationship between a given test score and performance on another measure of special relevance, usually referred to as a criterion" [41].

# 4. Item Development

#### 4.1. Methods

#### 4.1.1. Deductive Method

The literature on human-computer trust was reviewed as the first step in the item development step. It is important to identify what we wanted to measure and what type of measurement instrument we wanted to develop. Next, we should identify whether validated measurement instruments are already available before developing new ones. From a review of the literature in the relevant field, no available instruments were identified to measure trust in the context of human-AI from the perspective of non-expert user/laypeople.

We decided to develop a quantitative measurement, in the form of a survey, to measure human trust in AI systems based on factors that may influence human trust. As the construct of interest has been determined, we collected the literature related to trust from various fields, such as, Human-Computer Interaction, Information System, and Human-Factors. Indeed, much of the relevant literature cited and informed this research are not necessarily in the field of AI for the general public; therefore, an early user study to contextualise, extend and revise the theoretical framework of human-AI trust needs to be conducted. Hence, an inductive approach was also undertaken to discover factors that may have been missed from the literature review (deductive step).

#### 4.1.2. Inductive Method

A mixed method approach of online surveys and focus groups was selected to gather people's general opinions. We chose online surveys because online surveys can reach a wide population of individuals with various characteristics, which better matches our target population of non-expert users [64]. We chose to use online surveys, rather than paper surveys, and certain crowd participation

systems (such as Mechanical Turk) to provide access to a broader group, a larger number of potential participants, who could be reached in a timely and efficient manner" [65]. An initial version of the measurement instrument was developed to conduct the online survey using the initial domains and items.

At the same time, since the answers from online surveys are usually quite short, we conducted additional data collection with focus groups to improve the depth of analysis. Focus groups are an effective way to collect rich data due to the nature of the interaction between participants [66]. For example, unlike in the online survey, we explicitly asked participants what they thought the main challenges and opportunities were in using AI systems in healthcare, which helped improve our understanding of the underlined patterns identified in the online survey. To help contextualise the issues to healthcare scenarios, we provided participants in the online survey and focus groups with dramatised vignettes to read and reflect on. We aimed to evoke and elicit participant insights and chose the vignette technique as a method of enquiry. The vignette technique is a method that can elicit perceptions, opinions, beliefs, and attitudes from responses or comments to stories describing scenarios and situations that may be fictitious or adapted from real events" [67]. According to Finch, a vignette is "a short story about a hypothetical character in specific circumstances, to which situation the interviewee is invited to respond [68]. Vignettes are particularly appropriate for obtaining feedback from users regarding the implications of unrealised future possibilities. Therefore, this method is well suited for our research, where we need to elicit people's opinions on the public use of AI-assisted healthcare for early cancer diagnosis that does not yet exist.

We wrote a vignette depicting a fictitious scenario where AI-assisted health assessment is available and accessible to everyone for early cancer diagnosis. Although the vignette was fictional, the story still had to seem plausible and real to [67] participants and had to be built around actual experiences. Our vignette was based on real experiences shared by breast cancer survivors on

the breast health centre website <sup>5</sup>. The dramatisation of this vignette was designed, with the aim to stretch participants' thinking towards opposing views, extreme scenarios, and promote reflection on contentious actions and unforeseen consequences.

For the online survey, questions were presented using a Google Form formed as long text answers. The online survey was published through Amazon Mechanical Turk. Our target was 80 participants, with 40 participants from the general public and 40 participants from workers in the healthcare field. We placed "master worker" as the selection criteria for participation and also placed one check-in question within the survey to validate if participants read the vignettes carefully. The Mechanical Turk task lasted for a week, and in the end, we had 53 participants.

For the focus group, eight participants were recruited through the forum and asked to attend a face-to-face meeting on the Open University campus. Participants were Open University students and staff. The focus groups ran for one hour and were audio-recorded. The recordings of the focus group discussions were then transcribed, and any personal information was removed from the transcripts.

Both the focus group transcripts and the online survey open-ended responses were analysed using a grounded theory approach. Several studies on health-care systems and technologies have successfully applied the grounded theory method [69, 70]. Grounded theory is an appropriate approach to use in inductive methods because it allows the generation of new statements, hypotheses, or relationships.

# 4.2. Results: Trust Factors and Items Developed

# 4.2.1. Deductive Method

As mentioned previously, the aim of this study is to develop a scale to measure trust attitude towards an AI system based on trust factors. We have

 $<sup>^5 \</sup>mathrm{http://www.breastlink.com/}$ 

the reviewed literature on trust and identified factors that could affect trust.

Research on trust in human-automation systems' interaction, suggested that trust consist of human-related trust, environment-related trust, and learned trust [44][71][45]. Human-related trust is, as the name suggest, related to the human trustee, such as individual personalities, backgrounds, and capabilities. As the name suggest, environment-related trust is related to the environment or situation of the task and the system. Learned trust relates to the system itself, such as its behavior, reliability, transparency, and performance. Other research also proposed similar form of trust in different trust context, composed by similar concepts with different names. For example, in human-robot context, trust is composed of human factors, environment factors, and robot factors [48, 49]. In the context of information systems, trust consists of basic personality trust, basic institutional trust, and basic system trust. [46, 47]. To put it simply, a person trust towards an object (person, robot, AI) is built from their personality/attributes/characteristics, the environment surrounding the person and the interaction, with the object and the attributes/characteristics of the object.

A different perspective by Morrow et al theorised trust under two bases: cognitive and affective [72]. Cognitive base trust is trust resulted from a pattern of careful and rational thinking, while, affective base trust is trust that results from feelings, instincts and intuition. Moreover, this theory was included in the literature on human-computer trust [62] and in human-automation [45] under the human-related factors affecting trust. The link between human-related factors of trust and categorisation of cognitive-affective base trust, shows that trust is a multidimensional concept and can be categorised in different ways. An inter-disciplinary exploration on trust theory concluded that trust concepts are actually similar, overlapping, and sometimes only divided by different jargon [73]. Therefore, we looked at the literature on trust factors while not overly considering discipline-specific categorisations.

Understanding how the system works is part of cognitive factors in trust in automation [45, 44]. Not only for human-automation, human trust is often dependent upon such an understanding of a system/machine in general [74]. Empirical evidence has shown that understandability affect users' trust and confidence [75][76]. Cognitive compatibility, or understandability, was also found to be a trust factor in the study of trust in medical assistance devices [77].

Trust is dynamic, and can change over time as a result of experience with the system [74]. Research found that trust is typically built up in a gradual manner [78, 79], and when a system offers consistent experience the system is seen to be "reliable" or "predictable". In socio-psychology, reliability is seen as a factor of trust [56, 80]. Similarly, reliability is considered as one promoting factor of trust in automation [44, 45, 71]. At first, a person will judge the predictability of a system by assessing the consistency of its behaviours. The more consistent the system is, the more predictable it will appear to be. Trust will be directly related to the stability of the system, or system reliability.

At the end of repeated positive experiences, a person may develop "faith" towards a system. Even when people cannot know that a system will continue to be reliable in the future, they need faith to continue to depend on the system in the future. Faith means emotional security or confidence despite potential future uncertainty, and is considered as a factor of trust [56] which is driven by the affective/emotive side of a person [80]. Faith is also seen as a factor of trust in human-automation and human-computer context [45, 62].

On the one hand, a positive experience could affect cognitive base trust and, over time, forms affective base trust. On the other hand, a negative experience could also affect trust judgement. People who initially trusted a system could turn to distrust when they encounter system faults or errors[61]. Since errors and system faults indicate the lack of system competency, lack of trust is then expected because competency is considered one of the antecedents of trust [43]. The display of system technical ability/competence, in the form of confidence level, also affects the user's trust in automatic notification devices [19]. Technical ability to correctly perform its tasks is frequently regarded as one of the most important factors for human-machine trust [74, 62] and human-automation trust [71, 45, 44].

Other than competency, benevolence is also seen as a trust antecedent in human-human trust [43, 56]. Benevolence means that the trustee is believed to want to do "good" to the trustor. In human-automation and human-AI trust research, benevolence is seen as a system's purpose [44, 45]. However, since technology has no agency (want) nor intention (purpose) [46], we could argue that benevolence can also be seen as helpfulness in the context of human-technology trust. Therefore, helpfulness means that trustee is believed to be able to do good to the trustor, and is included as a the trust factor.

In human-human trust, in a situation in which trustee and trustor have similar purpose or intent could also evoke emotive/affective trust [81], which is contingent on personal preference [82]. Personal preference is shaped by culture and experiences [83] and culture is seen as a component that could affect trust in technology [45]. In human-computer trust, when a person finds a system agreeable or suits their personal taste, strong preference for the system and emotional attachment could be formed and affect the person trust [62].

The cross-discipline literature above identifies various concepts and factors that could affect trust in different contexts. However, even though multi-disciplinary source have added valuable insights to the construct, the diversity in wording and terms used should be unified. Therefore, we merged factors with similar meaning, and six trust factors were chosen to be the domains of the initial measurement instrument. The Table 1 below summarises our definition of these domains, which are based on the literature. We merged domains that overlapped in meaning and modified some of their descriptions into the final six trust metrics: perceived understandability, perceived reliability, faith, perceived technical competence, perceived helpfulness, and personal attachment.

Perceived Technical competence means that the system is perceived to perform the tasks accurately and correctly, based on the input information. Perceived Understandability means that user can form a mental model and predict future system behaviours. Perceived Reliability means that the system is perceived to be functioning consistently. Perceived helpfulness means that the system is perceived to provide adequate, effective, and responsive help. Faith

Table 1: Human-AI Trust Measurement Domains: Trust Factors and Descriptions

Trust Factors	Description		
perceived technical competence	system is perceived to perform the tasks accurately		
	and correctly based on the information that is input.		
perceived reliability	system is perceived to be, in the usual sense of re-		
	peated, consistent functioning.		
perceived understandability	user can form a mental model and predict future sys-		
	tem behaviour.		
personal attachment	user finds using the system agreeable, preferable,		
	suits their personal taste.		
faith	user has faith in the future ability of the system to		
	perform even in situations in which it is untried.		
perceived helpfulness	system is perceived to provide adequate, effective,		
	and responsive help.		

means that the user is confident in the future ability of the system to perform, even in situations in which has never used the system before. Finally, personal attachment means that users find using the system agreeable, and consistent with their personal taste.

# 4.2.2. Inductive Method

# Online Survey

As mentioned previously, an online survey was conducted to help explore other possible trust factors that were not considered previously. The online survey also aimed to test if any domain we included is not considered relevant or important by the general public in a healthcare context. We developed a dramatising vignette as a tool to help contextualise the measurement domains and items. The vignette technique is a method that can elicit perceptions, opinions, beliefs and attitudes from responses or comments to stories depicting scenarios and situations [67]. Vignette is also appropriate to elicit users' feedback on the

implications of possible futures yet to be realised. Because AI assisted healthcare for preliminary cancer diagnosis that is able to provide explanation does not yet exist, this method is suitable for our study.

Variable	Value	Frequency	%
Gender	Male	25	47.1
Gender	Female	28	52.8
A	<30	16	30.1
Age	30-40	23	43.3
(median = 35)	>40	14	26.4
Total		53	

Table 2: Initial Survey Demographic

The demographic of initial online survey participants can be seen in Table 2. We asked participants to first read a dramatising vignette and then to reflect on it and write down the potential main issues/problems (challenges) and the potential main advantages/benefits (opportunities) of AI systems in healthcare. The potential issues pointed out by participants mostly fell under Competency group, such as, misdiagnosis, false positive, false negative, or inaccuracy (34 instances). This finding can be put as a supporting argument that perceived technical competence is an important consideration for user to use an AI system. In the relation to explanation, participants mentioned their concern on the lack of information or information that is not laypeople friendly. "Yes. More information is always better for a diagnosis"- P24. "Vague information or terms an average person may not use,..."- P52. This finding can be linked to the perceived understandability and perceived helpfulness domains.

For the potential main advantages/benefits (opportunities) of AI system in healthcare, specifically AI breast cancer self-diagnosis system, participants mentioned Fast Result (18 instances) compared to the traditional diagnosis process. "Get some answers quickly, which gives you peace of mind"- P20. The AI system could also offers information which can be beneficial for users as a Second

Opinion (10 instances) provider, or provider for More Information (9 instances) in general, or both. "If used regularly, it can also be a great tool to track changes in your condition without having to see a doctor."- P50.

The topic of trust came up several times (13 instances), specially disproportionate level of trust, such as, Over-trusting (7 instances) and Distrust (6 instances) on the AI system. "People could put too much weight on the result giving it the same value as a true diagnosis."-P12. "People don't trust it enough."-P48. However, overall, we found no new dimension of trust factor from the open-ended answers. Since the nature of free-text survey response is often lack attention to context and conceptual richness [84], the answers rarely produce a rich qualitative data.

We also did measurement instrument pre-testing, as a part of domains and items selection and/or reduction, and also to test the form of measurement instrument. We created the initial measurement instrument based on the literature with six domains of trust factor and three statement items for each domain. Based on our knowledge, there is no rule of thumb for the number of items should be included in the measurement scale, as long as it's not too long that inspired participation fatigue or motivation [85]. Since this initial online survey is quite long, we decided to only use three statements for each domain, making it 18 statement in total. Participants were asked to rate, in Likert 7-point scale, the importance of 18 item statements before and after reading the dramatising vignette. For example, to evaluate perceived technical competence, participants were asked to rate the following statement: "The application would use appropriate methods to get results based on the information I input." (See Appendix A).

We inspected mode, median, and mean values for each item and internal consistency of each domain was then measured, using Cronbach's alpha (See Table 3). Based on the median rating, most of the domains are rated as very (rating = 6) or extremely (rating = 7) important by the survey respondents. The only item rated negatively (rating ; 4) was from personal attachment domain, with the statement: "you feel a sense of loss if the app is suddenly unavailable

	re	liabili	ty		technical understandabi		dability	personal attachment		helpfulness		ess	faith					
	r1	r2	r3	tc1	tc2	tc3	u1	u2	u3	p1	p2	р3	h1	h2	h3	f1	f2	f3
mode	7	7	7	7	7	7	7	7	7	2	5	5	7	7	6	5	7	6
median	7	7	7	7	7	7	6	7	6	3	5	5	7	6	6	5	6	6
mean	6.2	6.0	6.4	6.3	6.2	6.0	5.6	6.0	5.7	3.0	5.2	5.3	6.1	5.8	5.7	5.2	5.5	5.4
α	0.81			0.82			0.77			0.75			0.90			0.76		

Table 3: Importance Rating: Initial Measurement Scale with Six Domains and Three Statement Items Each

to use". Perceived reliability, perceived technical competence, and perceived helpfulness domains demonstrated excellent internal consistency, with their alpha coefficient ¿ 0.8 [86][40]. Meanwhile, perceived understandability, personal attachment, and faith domains' internal consistency can be regarded as acceptable. However, the overall initial measurement is still demonstrated excellent reliability Cronbach's alpha ¿ 0.94. From this result, we argued that the initial measurement scale is a good starting point, with some refinement need to be done in perceived understandability, personal attachment, and faith items.

# Focus Group

The focus group analysis was used to capture any missing human-AI trust factor that was not captured by the literature. During the focus group participants were asked to read the dramatising vignette and then have an open discussion on human-AI trust and the factors affecting this relationship. The codes emerged can be grouped to these core variables: User Needs, Communication, User Concern, AI usage, and Trust. In the following, we describe one of the core variable: Trust, as a part of inductive method in this item generation phase.

Based on the questions asked about trust before and after the vignette, trust is affected by communication and credibility. AI systems' credibility could be proven with license or certification. License and certification are required for all medical tools, and AI in healthcare should too." Overseeing bodies, both in the U.S. and here, and elsewhere. I think (here) it's BMC, the royal colleges.

Things that you have to be able to practice medicine in most countries. Integrating AI into that system somehow. Whether it's, through having to release your bug report and knowing what the control condition is you run the amazing test results. "-P5." What if AI goes through a residency with real physicians. The tool itself needs to be continuously improved for a period of two years by physicians, by experts, in clinical practice."-P3

Autonomy is an important principle in medical ethics as perceived by patients, and it means individuals demand to be free to choose whether and what kind of treatment to receive [87]. Autonomy is relevant for both interactions between medical professionals and patients, and between AI systems and users. In the AI healthcare system context, users should have the right to make decisions for themselves; and should be put in the right conditions that enable them to make those decisions in a well informed but autonomous way. The decisions mentioned by participants vary from decisions regarding treatment, to the decision regarding whether or not they want to use the system, or even decisions about the conditions that enable them to make decisions, in this case, is the decision about what kind of information users want to be given in the explanation. "I would like to be able to invoke it or turn it off at my choice."-P1."you could be given references if you wanted to do research. but it's also up to you"-P6.

From these results, two main additional trust factors emerged: Institutional Credibility and User Autonomy (See Table 4). For Institutional Credibility, participants meant the users' trust is placed in the institution which regulates the certification of the AI system. We previously did not include institution-based trust in the initial domain, because according to McKnight et al.'s paper, institution-based trust is outside of the overall concept of trust in specific technology [46]. Since trust towards an AI system can be considered as trust in specific technology, we did not include institutional credibility, or structural assurance supporting the technology use.

User Autonomy is also deemed as important by participants. As mentioned above, user should be able to control their decision regarding treatment or regarding whether or not to use the system. This is in line with previous research on trust in healthcare, patient's trust will improve when doctors give patient autonomy by letting them manage their disease [88] [89].

Table 4: Human-AI Trust Measurement Additional Domains: Trust Factors and Descriptions

Trust Factor	Description
institutional credibility	user beliefs that the technology has been tested or certified by
	overseeing bodies.
user autonomy	user knows and able to decide for their own decision.

#### 5. Item Evaluation

#### 5.1. Methods

After the initial development process, we further review and revise initial domains and items. Since it is recommended to create a large number of items in the early stage of item development [35], we created five item statements for each domain, including two new added domains, as our second version of measurement instrument (See Appendix). To assess the new measurement instrument, an evaluation by experts and target population was carried out.

# 5.1.1. Expert Validation

Evaluation by expert entails analysis of content validity, which is the degree to which items of a measurement instrument are relevant to and representative of the targeted construct for a particular measurement purpose [63, 42]. This is one of the most important of measurement instrument validations [34]. Content validity is also used to observe the correct grammar and appropriate wording in items and appropriate scoring [90]. In summary, experts review the clarity (the question is clear and specific. the question make sense to the reader), coherence (the question is logical, consistent, and reasonable in the context of the research

problem being addressed), and completeness (the statement presented is fully represented by the definition) of the measurement instrument items [91]. To conduct expert validation, we went through several stages: content validation form development, expert selection, validation administration, and expert rating analysis.

Similar to any measurement instrument, the content validation form should be developed with the appropriate items to allow the experts to have a clear expectation and understanding about the task. As defined above, content validity is the degree of relevance and representativeness [63] and clarity of the targeted domain item [90]. Therefore, the form should include at least two evaluations: content relevancy and content clarity. We presented the domains and its definitions, followed by five items related to them to the experts. The experts were requested to critically review the domain and its items before providing rating on each item using a 5-point Likert scale of agreement. For each item, we asked if "the statement is clear, consistent, specific, and a non-expert reader would be able to make sense of it" to provide evidence of content clarity, and if "the statement fully represents the definition of perceived understandability" to provide evidence of content representativeness and relevancy. Additionally, we asked experts if they understand the domain definition because unambiguously defined domain can also be considered as one of conditions of content validity [92]. Since the measurement instrument review requires verbal comments from experts, a mixed data collection method was administered, where qualitative data was collected via interview and quantitative data was collected via content validation form.

The next stage is experts selection. The experts involved in this stage should be highly knowledgeable about the domain of interest and measurement instrument development [42][93]. Since this measurement instrument was developed in the context of AI systems in healthcare, the domain includes; AI/computing and medicine/healthcare. Therefore, we aimed to recruited AI/computing experts and medical experts, in addition to measurement instrument development expert.

Invitations for interview were sent via email to possible experts via email from our personal research network. The number of expert recommended by the literature is two at the minimum [94] and ideally range between five to seven experts [95][96]. In the end, we selected seven experts: two scale development experts with psychology background, two scale development experts from computing field, one AI expert, and two medical experts. The interviews ran for one hour and were transcribed in the process.

The last stage of expert validation is expert rating analysis. The Content validity index (CVI) was calculated to measure proportional agreement [97][96]. Although CVI is broadly used for content validity, this index has been criticised for not considering possible inflated values caused by chance agreement [98]. Therefore, we also calculated Cohen's coefficient kappa ( $\kappa$ ) [99] which adjusts for chance agreement and is considered as the most efficient [98]. Items with low value of CVI and  $\kappa$  were considered invalid and removed from the measurement instrument. Lastly, comments from experts worked as a guide and consideration to refine the trust factors and revise the rest of the items.

#### 5.1.2. Cognitive Interview

Not only from experts, evaluation on the target population is also important. We conducted Cognitive Interview, to evaluate if the items reflect the domain of study by ensuring that the target population understand the item statements and/or questions[100] and help refine the measurement instrument items. Cognitive Interview can help improve clarity, identify confusing and problematic item, indicate problematic item order, reveal the thought process of participants, ensure the intended data are produced [101][102], and is the recommended method to evaluate the measurement instrument before Survey Administration [34].

It is recommended to run 5-15 interviews from target population sample [101][100]. The interview technique combined think-aloud approach with verbal probing. In the interview, first, we described the study and introduced participants to AI technology and AI in healthcare. Since, the target population is

non-experts/laypeople this introduction process is important. Participants were then asked to read the item statement, answer if they can understand and make sense of it, and explain what does the statement mean using their own words. Based on the answer, verbal probing might occur. Invitations for this interview were sent via email and electronic messages from our personal and professional network.

#### 5.2. Results

### 5.2.1. Expert Validation

We conducted interviews with seven experts separately. The experts include professionals and academics: two scale development experts from psychology field, two scale development experts from computing field, one AI expert, and two medical experts. In the first round, CVI was calculated in item level by dividing the number of experts giving a rating 4 or 5 to the representativeness of each item with the total number of experts. Evaluation criteria for CVI is "Excellent" for CVI > 0.79 [94, 103, 96] and "For Revision" for CVI > 0.7 [104]. After CVI for all instrument items were calculated, kappa was calculated using numerical values of probability of chance agreement (PC) and CVI of each item in following formula:

$$K = (CVI - PC)/(1 - PC).$$

Evaluation criteria for kappa value are "Excellent" for  $\kappa \geq 0.74$ , "Good" for  $0.74 > \kappa \geq 0.6$ , and "Fair" for  $0.59 > \kappa \geq 0.40$  [105]. Among the 40 instrument items, 16 items with CVI lower than 0.7 and  $\kappa$  score lower than 0.4 were interpreted as "Invalid" and removed from the measurement instrument (See Table 5 and Table 6). However, the number of items removed from each dimension were not equal and range from three items (perceived understandability and perceived helpfulness) to one item (user autonomy and faith). Thus, to have the same number of item for each domain, we selected the two best items from each domain. When domain has two or more items passed the CVI and  $\kappa$  "Excellent"

threshold, items with "For Revision" scores were removed, making it 20 items in the end.

In the second round, we looked at the Cohen's  $\kappa$  of the Clarity ratings from the rest of 20 items, to decide if the item needs revision or not. The evaluation criteria for kappa value is the same as above. Items with  $\kappa < 0.4$  were considered "Poor" and removed from the item pool, and items with "Excellent" clarity were accepted without major revision. Since only two best items were selected for each domain, one of *faith* items with "Fair" clarity and one of user autonomy items with EA < 7 were removed. In the end, we have the measurement instrument consists of 16 items from 8 domains, with comments from experts that helped refine and revise the items accordingly.

# 5.2.2. Cognitive Interview

We conducted qualitative interviews with nine participants. All participants are laypeople, with age range: six participants were below 30, three participants were in 30-45, and three participants were above 45 years old. Each cognitive interview lasted one to two hours in semi-structured format. As described above, participants' were expected to think-aloud their understanding on each item statement using their own words. Since no specific data analysis method is recommended by cognitive interview literature [101][100], general view on participants' understanding and in-depth look on participants' cognitive processing were noted.

In general, participants claimed that the items are understandable and make sense. When participants explained their interpretation on the item statements, the description of their mental process allowed participants to answer in a manner that reflects their experience, which indicates their understanding [101]. The participants were able to understand correctly the specifications of the items and, crucially, the interpretation were consistent across participants reflect the trust factor definition.

However, some inconsistencies between participants' interpretation were found on two item statements from perceived technical competence and personal at-

Dimension Items	EA	CVI	PC	$\kappa$	Interpretation (CVI)	Interpretation $(\kappa)$	Clarity
understandability					()	(**)	
u1	7	1.000	0.008	1.000	Excellent	Excellent	Excellent
u2	6	0.857	0.055	0.849	Excellent	Excellent	Excellent
u3	1	0.143	0.055	0.093	Invalid	Invalid	
u4	4	0.571	0.273	0.410	Invalid	Fair	
u5	3	0.429	0.273	0.214	Invalid	Invalid	
technical competence							
tc1	6	0.857	0.055	0.849	Excellent	Excellent	Excellent
tc2	5	0.714	0.164	0.658	For Revision	Good	Poor
tc3	3	0.429	0.273	0.214	Invalid	Invalid	
tc4	5	0.714	0.164	0.658	For Revision	Good	Excellent
tc5	4	0.571	0.273	0.410	Invalid	Fair	
reliability							
r1	6	0.857	0.055	0.849	Excellent	Excellent	Fair
r2	5	0.714	0.164	0.658	For Revision	Good	Excellent
r3	5	0.714	0.164	0.658	For Revision	Good	Poor
r4	2	0.286	0.164	0.146	Invalid	Invalid	
r5	4	0.571	0.273	0.410	Invalid	Fair	
helpfulness							
h1	7	1.000	0.008	1.000	Excellent	Excellent	Excellent
h2	6	0.857	0.055	0.849	Excellent	Excellent	Good
h3	3	0.429	0.273	0.214	Invalid	Invalid	
h4	1	0.143	0.055	0.093	Invalid	Invalid	
h5	4	0.571	0.273	0.410	Invalid	Fair	

Table 5: Expert Validation Review Ranking: Dimension Items' Experts in Agreement (EA), Content Validity Index (CVI), Probability of Chance agreement (PC), Cohen's coefficient kappa  $\kappa$ , CVI evaluation interpretation,  $\kappa$  evaluation interpretation,  $\kappa$  item Clarity (Part 1)

Dimension Items	EA	CVI	PC	$\kappa$	Interpretation	Interpretation	Clarity
Dimension Items	ĽA	CVI		_ ^	(CVI)	$(\kappa)$	Clarity
personal attachment							
pa1	7	1.000	0.008	1.000	Excellent	Excellent	Excellent
pa2	6	0.857	0.055	0.849	Excellent	Excellent	Good
pa3	5	0.714	0.164	0.658	For Revision	Good	
pa4	4	0.571	0.273	0.410	Invalid	Fair	
pa5	3	0.429	0.273	0.214	Invalid	Invalid	
user autonomy							
ua1	7	1.000	0.008	1.000	Excellent	Excellent	Excellent
ua2	6	0.857	0.055	0.849	Excellent	Excellent	Excellent
ua3	7	1.000	0.008	1.000	Excellent	Excellent	Excellent
ua4	5	0.714	0.164	0.658	For Revision	Good	
ua5	4	0.571	0.273	0.410	Invalid	Fair	
faith							
f1	6	0.857	0.055	0.849	Excellent	Excellent	Fair
f2	6	0.857	0.055	0.849	Excellent	Excellent	Excellent
f3	7	1.000	0.008	1.000	Excellent	Excellent	Excellent
f4	4	0.571	0.273	0.410	Invalid	Fair	
f5	5	0.714	0.164	0.658	For Revision	Good	
institution credibility							
ic1	6	0.857	0.055	0.849	Excellent	Excellent	Good
ic2	6	0.857	0.055	0.849	Excellent	Excellent	Good
ic3	1	0.143	0.055	0.093	Invalid	Invalid	
ic4	4	0.571	0.273	0.410	Invalid	Fair	
ic5	5	0.714	0.164	0.658	For Revision	Good	

Table 6: Expert Validation Review Ranking: Dimension Items' Experts in Agreement (EA), Content Validity Index (CVI), Probability of Chance agreement (PC), Cohen's coefficient kappa  $\kappa$ , CVI evaluation interpretation,  $\kappa$  evaluation interpretation,  $\kappa$  item Clarity (Part 2)

tachment domains. One statement from perceived technical competence: "The AI system uses appropriate methods to get results and to reach decisions based on the information I input" was understood differently. The cause of this inconsistency was the word "appropriate". Some interpretations on "appropriate method" were: ethical method, method that gets the job done, method human professionals uses, or method where the data is taken without the user's permission. One of the participants claimed that the word appropriate is "a bit too fluid".

In the statement from the personal attachment factor, "I find the AI system suitable for my style and I would feel a sense of loss if I could no longer use it.", the word "style" was interpreted differently. As described previously, personal attachment measures the degree to which the user finds using the system agreeable, preferable, and suits their personal taste. The sentence was developed based on the style's definition from Cambridge Dictionary as a way of doing something, especially one that is typical of a person. Even though some participants interpreted style correctly, most participants interpreted style as lifestyle. Thus, when they reflected on their experience, they drove it from their lifestyle. "I think style is lifestyle, the app is suitable for my lifestyle. Like I like to do yoga so the apps that suit my style are health yoga app. "- P8. One of the participants even appeared confuse when reading the statement. "Hmm. That's an interesting one. Yeah. Suitable for my style. And I would feel a sense of loss if I could no longer use it. Um, "my style" is intriguing the way you worded that."-P2.

Lastly, the word "vendor" in the institution credibility statement was found to be confusing. "I am confident in the AI system capability because it is developed by a reputable institution, and backed by valid vendors and consumer protections.", "I'm not sure what vendor means. [...] Vendors makes me think of food."-P5. Based on this result, none of the item statements were dropped and some modifications on the word choice were applied.

# 6. Survey Administration

#### 6.1. Method

After all item statements and domains were evaluated, we administered the main survey to further evaluate the measurement instrument. A survey could identify the characteristics of a broad population of individuals if a clear research question inquiring about the nature of the target population is present [64]. We chose to use an online survey, because an online survey takes advantage of the Internet to provide access to broader groups and efficient in time [65].

The online survey consists of; demographic questions (gender and age group), their general trust towards AI in healthcare, their likelihood to use AI in healthcare, and the trust measurement instrument at the end (See Table 7). Participants were asked to read the information page and filled the consent form before proceeding to the online survey. The measurement instrument was presented after videos of available AI medical support systems. The online survey contains two sets of questions with 20 items in total. The first set of questions contains two demographic items (gender and age group) and two trust propensity questions. The second set of questions contains 16 statements from the measurement instrument. Between the first and the second set of questions, participants were assigned to watch a two-minutes video of cancer detection/risk assessment applications available on the market, such as, SkinVision (skin cancer detection), Braster (breast cancer detection), and Alexa Babylon (health assessment). Participants were then asked to rate their agreement with the statements from the measurement instrument based on the application that they just saw using 7point Likert scales. The online survey was developed using the Google Forms platform and published via Amazon Mechanical Turk. To minimise submission from bots, we only accept master worker and put different codes in the survey to submit at the end.

To decide the sample size or the number of participant, we looked at the guidelines given by the literature. The overall recommendation is the larger sample size/participants the better. The number of participants can be determine

using the ratio of participant and instrument item, such as, 5:1 participant-to-item ratio [106] and 10:1 participant-to-item ratio [107]. The number of participants can also be determined without taking the number of item into account. The 200-300 range is argued to be appropriate [108], and other literature considered 300 participants as good [109][110]. Thus, we aimed to have 300 participants.

#### 6.2. Results

The Mechanical Turk tasks were up for two weeks and data from 300 participants were collected. The participants were 52.7% male, 47% female, 0.3% prefer not to say their gender. Almost half of the participants were between 30-40 years old (46%), with the rest 22% were between 40-50 years old, 18.7%between 20-30 years old, 13% above 50 years old, and 0.3% below 20 years old. Further analyses for measurement instrument evaluation, methods and results, are described in the following sections. To determines if the responses given with the sample are adequate, we performed the Kaiser-Meyer-Olkin (KMO) test [111]. KMO measures the sampling adequacy (MSA) with criterion: 0-0.49 as unacceptable, 0.50-0.59 as miserable, 0.60-0.69 as mediocre., 0.70-0.79 as middling, 0.80-0.89 as meritorious, and 0.90-1.00 as marvelous [111]. Additionally, we ran Bartlett's test of sphericity test [112] to test the null hypothesis: the correlation matrix is an identity matrix. If the significance level is less than 0.05, the null hypothesis is rejected, which means the items are suitable for structure detection. As depicted in Table 8, KMO results of high value (0.9458) implied the adequacy of the sampling data and a significant test statistic (0.00) by Bartlett's test of sphericity indicated that a factor analysis may be useful with this data.

Domain	Statement				
Understandability	I understand how the AI system works and I feel confident I will be able to use				
	it in the future.				
	I understand how the AI system behaves, how it can assist me, and what I can				
	expect from using it in the future.				
Technical Competence	The AI system uses appropriate methods to get results based on the information				
	I input.				
	The AI system correctly uses the information I input to provide accurate re-				
	sults.				
Reliability	The AI system consistently provides the results it is expected to produce.				
	The AI system responds the same way under the same conditions at different				
	times.				
Helpfulness	When I need help, the AI system responds to my needs effectively and respon-				
	sively.				
	The AI system provides me with the effective and responsive help I need.				
Personal Attachment	I find the AI system suits my preference and I would feel a sense of loss if I				
	could no longer use it.				
	I like using the AI system because it suits me, and always want to use it.				
User Autonomy	I feel in control when operating the various functions and features of the AI				
	system.				
	The AI system has functionalities and features I can control.				
Faith	When I am unsure about the AI system's result, I believe in the AI system				
	rather than myself.				
	Even if I am not sure about the result and the actual performance, I am con-				
	fident that the AI system will provide the best result.				
Institutional Credibility	I feel assured using the AI system because it is made by a reputable institution				
	and therefore already went through a credible regulation process.				
	I am confident in the AI system capability because it is developed by a reputable				
	institution, and backed by valid companies and consumer protections.				

 ${\it Table 7: 16 \ Survey \ Questions/Statements \ (Two \ Statements \ for \ Each \ Domain}$ 

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.9458
	Approx. Chi-Square	3634.78
Bartlett's Test of Sphericity	$\mathrm{d}\mathrm{f}$	120
	Sig.	0.00

Table 8: Kaiser-Meyer-Olkin (KMO) sample adequacy test.

#### **Measurement Instrument Evaluation**

# 7. Test of Dimensionality

# 7.1. Method

To test the dimension of the proposed measurement instrument, we used Factor Analysis. Two main classes of factor analysis are Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA). EFA seeks to discover the measurement model and, as the name suggest, exploratory in nature. Meanwhile, CFA starts with a hypothesis model, such as, how many factors and which items load on which factors. Our current model consists of eight (8) dimensions of trust factors: perceived reliability, perceived technical competence, perceived understandability, faith, personal attachment, perceived helpfulness, user autonomy, and institution credibility; with two items on each dimension. Given that we have hypothesised the trust model, confirmatory factor analysis (CFA) was performed in this stage. CFA investigates how well the hypothesised factor structure (model) fits with the data [113]. Some of the fit indices are: Comparative Fit Index (CFI >0.95), Tucker Lewis Index (TLI >0.95), Root Mean Square Error of Approximation (RMSEA <0.06), Standardized Root Mean Square Residual (SRMR <0.08) and low Chi-square [114] [35][34].

# 7.2. Results

Table 9 shows the fit indices for the hypothesised model. Based on the root mean square error of approximation (RMSEA), standardized root mean square residual (SRMR), comparative fit index (CFI), and Tucker–Lewis index

(TLI) reported, the hypothesised model fits well and does not need additional alteration.

Chi-square	124.209
Comparative Fit Index (CFI)	0.987
Tucker-Lewis Index (TLI)	0.979
Root Mean Square Error of Approximation (RMSEA)	0.046
Standardized Root Mean Square Residual (SRMR)	0.023

Table 9: Confirmatory Factor Analysis of Trust Model

#### 8. Test of Reliability

#### 8.1. Methods

# 8.1.1. Internal Consistency

As mentioned previously, reliability refers to the degree of consistency demonstrated when a measurement is repeated under the same conditions [115]. Multiple assessments of reliability have been developed, and among these assessment methods, the coefficient alpha, specifically Cronbach's alpha [86], has been the most widely used measure of reliability. Coefficient alpha is commonly used to estimate one type of reliability: internal consistency. Internal consistency represents the degree to which the measurement items are inter-correlated, or if they assess the same construct consistently. However, many studies have criticised coefficient alpha, pointing out that coefficient alpha can only be treated as scale reliability when specific conditions hold [116], which unlikely to hold in practice[117, 118]. These conditions are: (1) items are unidimensional; (2) the average factor loading is above .7; (3) the differences between individual factor loadings and average factor loading are less than .2 [116]. Failure to hold these three conditions could resulted in biased result.

The proposed alternative to coefficient alpha is coefficient omega, which has been argued to be a more sensible measurement of internal consistency [116, 118]. Moreover, the formulation of coefficient omega is deemed to match the definition of reliability [119]. However, the difference between coefficient omega and coefficient alpha was found to be small in applications [120, 37], and coefficient alpha can be treated as identical to coefficient omega when the conditions (1-3) above hold [116]. Thus, we used both coefficient alpha and omega to assess the reliability of our measurement instrument, with Cronbach's alpha [86], Raykov's [41], Bentler's [121], and McDonald's omega [119]. If the value of McDonald's omega is similar to the other two omegas (Bentler's and Raykov's), it indicates that the model fits the data well [121].

#### 8.1.2. Test-retest reliability

To assess the instrument temporal stability, test-retest reliability was conducted. Test-retest reliability looks at the reliability across time, in which the same participants are able to perform similarly in different times [116]. The tests usually quantified using correlation [63], such as, Intra-class Correlation Coefficient [122] and Pearson product moment correlation coefficient (Pearson r) [123]. Even though reliability was often assessed using Pearson r [124], Pearson r was not recommended for assessing test-retest reliability [125], especially for non-continuous data [126]. If the correlation value is high (close to 1), it indicates high test-retest reliability; and if the correlation value is close to zero, it indicates low reliability.

# 8.1.3. Replication: Repeated Survey

The test-reliability is a part of replication processes, where the survey was administered in two (or more) different times to the same group of people. In order to do this, we collected additional data and repeated the online survey. The repeated survey consists of demographic questions (gender and age group), their general trust towards AI in healthcare, and the trust measurement instrument (See Table 7). Participants were then asked to interact with AI healthcare prototypes and rate their agreement with the statements from the measurement instrument using 7-point Likert scale. The online survey was developed using

the Google Forms platform and published via Amazon Mechanical Turk. Similar with the main survey, to minimise data from bots, we only accept master worker and put different codes in the survey to submit at the end.

#### 8.2. Results

# 8.2.1. Internal Consistency

The coefficient omegas and coefficient alpha were calculated in R. Table 10 depicts that all alphas and omegas values are above 0.7, indicating internal consistency in all dimensions and overall measurement [40]. Additionally, the results show that all Raykov's, Bentler's, and McDonald's coefficient omega are similar, suggesting that the model fits the data well. This support the finding in previous Test of Dimensionality stage.

	Cronbach's alpha	Raykov's omega	Bentler's omega	Mcdonald's omega
reliability	0.7232	0.7244	0.7244	0.7244
technical competence	0.8371	0.8383	0.8383	0.8383
understandability	0.7860	0.7938	0.7938	0.7938
personal attachment	0.8388	0.8393	0.8393	0.8393
helpfulness	0.8683	0.8699	0.8699	0.8699
faith	0.8285	0.8306	0.8306	0.8306
user autonomy	0.8289	0.8311	0.8311	0.8311
institution credibility	0.9136	0.9136	0.9136	0.9136
overall measurement	0.9481	0.9512	0.9512	0.9504

Table 10: Reliability tests for the measurement instrument.

In relation to the different reliability assessment measures, the results shown that the difference in coefficient alphas and omegas are small (less than .1). To see if this is the case where coefficient alpha can be treated as coefficient omega, we tested the three conditions of coefficient alpha. However, we could not conclude if the conditions hold. We fit the data to unidimensional model (single-factor model) with CFA and applying the same fit indices: Comparative Fit Index (CFI >0.95), Tucker Lewis Index (TLI >0.95), Root Mean Square Error

Parameter		Difference
Average loading	0.737	
Individual loading		
under standability 1	0.567	0.17
under standability 2	0.623	0.114
technical competence1	0.753	-0.016
technical competence2	0.794	-0.057
reliability1	0.759	-0.022
reliability2	0.663	0.074
helpful1	0.762	-0.025
helpful2	0.824	-0.087
personal attachment1	0.697	0.04
personal attachment2	0.819	-0.082
user autonomy1	0.740	-0.003
user autonomy2	0.677	0.06
faith1	0.691	0.046
faith2	0.780	-0.043
$institution\ credibility 1$	0.812	-0.075
institution credibility2	0.838	-0.101

Table 11: Average loading, Individual item loadings, and Difference between the two, in Fitted Unidimensional (Single Factor) Model.

of Approximation (RMSEA <0.06), Standardized Root Mean Square Residual (SRMR <0.08) and low Chi-square [114, 35, 34]. Only SRMR (0.068) passed the threshold, (CFI = 0.836, TLI = 0.811, RMSEA = 0.138), making the condition (1) not hold. Table 11 shown that both conditions (2 and 3) hold, with average factor loading is above .7 and the differences between individual factor loadings and average factor loading are less than .2. Further test and evaluation should be done to confirm this. Nonetheless, we conclude that our measurement instrument is reliable based on the coefficient alpha and omegas.

# 8.2.2. Test-retest reliability from Repeated Survey

The Mechanical Turk tasks were up for one month and data from 304 participants were collected. The participants were 53.6% male, 45.4% female, 0.7% prefer not to say their gender. Half of the participants were between 30-40 years old (52.6%), with the rest 19.4% were between 20-30 years old, 15.8% were between 40-50 years old, and 12.2% above 50 years old. As mentioned previously, the test was quantified using Intra-class Correlation Coefficient (ICC) between the ratings given by the same participants answered at closely spaced points in time (30 minutes - one hour). The test-retest reliability was established with ICC value = 0.7377.

## 9. Test of Validity

#### 9.1. Methods

A measurement instrument should not only be reliable but also valid, because reliability is a necessary but not sufficient condition for validity. Validity refers to the degree to which an instrument accurately measures the dimension or construct for which it was designed [41]. Due to the wide range of validity assessment metrics, many terms are used, such as, concurrent validity, construct validity, content validity, convergent validity, criterion validity, discriminant validity, divergent validity, face validity, and predictive validity. However, validity assessment can be summarised in three main forms:

- Content validity (including: face validity)
- Construct validity (including: convergent validity, discriminant validity)
- Criterion validity (including: predictive validity, concurrent validity)

Up until this stage, we have evaluated the content validity and face validity in the Item Evaluation stage. Content validity is the degree to which the instrument measure what it designed to measure. Content validity is not only assessed using statistical procedures but also relies on the reasoning [40], which is why Expert Validation and Cognitive Interview were conducted.

### Construct Validity

After survey has been administered, construct validity of a measurement instrument can be examined. Construct validity refers to the degree to which an instrument assesses a construct of real-world concern and is associated with evidence that measures other constructs [41]. Convergent validity is the extent to which a construct measured in different ways produces similar results. Literature suggest that convergent validity is established when the average variance extracted (AVE) is above 0.5 [127, 128, 129]. Other literature suggest that alongside AVE value, composite reliability (CR) also need to be considered, and convergent validity is established when the CR is above 0.7 [55].

There are various definitions of discriminant validity in the existing studies, one is the extent to which a measure is novel and not just a reflection of some other measurement [39, 41, 34]. The other is the extent to which two constructs were empirically distinguishable [130][131]. Since our main objective is to evaluate the validity of our measurement instrument, we assessed discriminant validity through our trust factors (domains) and trust propensity. However, we also looked at the discriminant validity between our domains to see if our domains are empirically different one from another. To evaluate discriminant validity, Fornell-Larcker criterion and Heterotrait-monotrait (HTMT) ratio of correlation can be examined. Fornell-Lacker criterion compares the square root of the AVE with the correlation of latent constructs, and a greater value of each AVE indicates discriminant validity [132]. The Heterotrait-monotrait (HTMT) criterion calculates HTMT ratio, and value close to 1 indicates a lack of discriminant validity [133]. Literature proposed threshold values for HTMT ratio are 0.9 [134] and 0.85 [135]. We examined both methods, since it is recommended to employ more than one method [39], and literature on discriminant validation have used multiple distinct measurement methods [136, 137].

## Criterion Validity

The content and construct validity are considered as validity of measurement [36], and included in the objective tests for psychological instrument [138]. In con-

trast, criterion validity is recognised as validity for decision, instead of validity of a measurement [36]. Validity for decision means that the measurement can lead to a correct/right decision, such as, psychological test for recruitment criteria. Criterion validity refers to the extent of the relationship between a particular criteria on other related measures, whether it's with the "gold-standard" measurement (concurrent validity) or other related measurement in the future (predictive validity) [41]. Even though our trust measurement was not designed to be a decision-making aid nor designed to predict any future decision, we still assessed the criterion validity of our measurement to explore our measurement aptitude. The criteria we used was trust, and since currently there is no "gold-standard" for human-AI trust measurement, the concurrent validity was assessed with single-question trust measurement from Replication survey. Concurrent validity is established when the constructs of our measurement instrument are significantly correlated to the trust rating, which were measured at the same time. Predictive validity is established by regressing some future outcome on the established construct, which we assessed with regression analysis to the single-question trust measurement from Replication survey.

# 9.2. Results

We were looking at the construct validity, where we evaluate if an instrument measures a construct that is not directly observable [40]. Construct validity composed of convergent validity, when the domains address the same construct; and discriminant validity, when the domains address different aspects of the construct.

# Construct Validity: Convergent Validity

A measurement can established convergent validity when the measurement domain/construct correlates highly with each other. As depicts in Table 12, all composite reliability (CR) values are above 0.7 [55], all average variance extracted (AVE) are above 0.5, suggesting the convergent validity of measurement instrument. This result suggest that our measurement instrument could

	CR	AVE	r	tc	u	pa	h	f	ua	ic
reliability (r)	0.724	0.585	0.764							
technical competence (tc)	0.838	0.733	0.895*	0.855						
understandability (u)	0.794	0.665	0.772*	0.815	0.815					
personal attachment (pa)	0.839	0.729	0.795*	0.748	0.552	0.853				
helpfulness (h)	0.870	0.769	0.939*	0.870*	0.747	0.746	0.876			
faith (f)	0.831	0.708	0.803*	0.683	0.479	0.833	0.729	0.841		
user autonomy (ua)	0.831	0.713	0.758*	0.764	0.639	0.751	0.731	0.694	0.844	
institution credibility (ic)	0.914	0.846	0.779*	0.774	0.552	0.836	0.744	0.870*	0.743	0.920

Table 12: Composite reliability (CR), the average variance extracted (AVE), the square root of AVE (in bold), and correlations between constructs (off-diagonal).

examines or measures trust in different ways while still yields similar results.

# Construct Validity: Discriminant Validity

The discriminant validity of inter-construct was first assessed with Fornell and Larcker criterion [132]. As described above, this method compares the AVE with the correlation of latent constructs, and a greater value of each AVE indicates discriminant validity. However, there are different interpretation in the literature to conclude the discriminant validity. Some literature applied Henseler et al. [133] interpretation, where the comparison is made from the square root of each AVE with the correlation coefficients for each construct [139][130]. Based on this interpretation, as shown in the Table 12, the correlation coefficients between reliability - all constructs, helpfulness-technical competence, and institution credibility-faith, are greater than the AVEs (marked with \*) which means those constructs do not hold discriminant validity.

Another interpretation claimed that discriminant validity is only established when the AVEs for both constructs are bigger than the squared factor correlation/shared variance (SV) between them, and Henseler et al. interpretation is considered one of the misapplications of Fornell-Larcker criterion [140]. The misapplication rooted in the usage of only one of the two AVE values or the

AVE/SV	r	tc	u	pa	h	f	ua	ic
reliability (r)	0.585							
technical competence (tc)	0.755*	0.733						
understandability (u)	0.568	0.642	0.665					
personal attachment (pa)	0.604*	0.547	0.287	0.729				
helpfulness (h)	0.854*	0.745*	0.543	0.545	0.769			
faith (f)	0.627*	0.467	0.228	0.759*	0.523	0.708		
user autonomy (ua)	0.556	0.569	0.403	0.575	0.530	0.491	0.713	
institution credibility (ic)	0.576	0.579	0.295	0.690	0.543	0.748*	0.537	0.846
trust propensity	0.180	0.203	0.133	0.374	0.163	0.433	0.242	0.328

Table 13: Fornell-Larcker Table: The average variance extracted AVE (in bold), the square root of correlations between constructs SV (off-diagonal).

average of the two AVE values, instead of both AVE values when comparing to the SV. Based on this original interpretation, as shown in the Table 13, the SVs between reliability-technical competence, reliability-personal attachment, reliability-helpfulness, reliability-faith, helpfulness-technical competence, faith-personal attachment, and institution credibility-faith, are greater than both constructs AVEs (marked with \*) which means those constructs do not hold discriminant validity [140].

Lastly, we examined discriminant validity with the Heterotrait-monotrait (HTMT) criterion. HTMT criterion was proposed as a superior method that successfully achieved higher specificity and sensitivity rates (97-99%) compared to the older criterion, including Fornell-Lacker (20.82%). As mentioned previously, lack of discriminant validity usually indicated with HTMT value close to 1, or HTMT higher than set threshold. Table 14 shows HTMT results, and using both 0.85 [135] and 0.9 [134] threshold, reliability-technical competence, reliability-helpfulness, helpfulness-technical competence, faith-personal attachment, and institution credibility-faith passed the threshold (marked with \*). Based on this criterion, discriminant validity was not established on all items.

	r	tc	u	pa	h	f	ua	ic
reliability (r)	1.000							
technical competence (tc)	0.890*	1.000						
understandability (u)	0.778	0.823	1.000					
personal attachment (pa)	0.783	0.745	0.541	1.000				
helpfulness (h)	0.945*	0.873*	0.742	0.737	1.000			
faith (f)	0.799	0.689	0.484	0.872*	0.723	1.000		
user autonomy (ua)	0.765	0.765	0.645	0.762	0.729	0.705	1.000	
institution credibility (ic)	0.769	0.772	0.550	0.832	0.740	0.868*	0.736	1.000
trust propensity	0.425	0.455	0.368	0.612	0.404	0.658	0.492	0.573

Table 14: The Heterotrait-Monotrait (HTMT) Ratio Table.

In summary, using different methods and criterion, the discriminant validity was not established on all inter-constructs (domains) of our measurement instrument. Since the dimensionality and the internal consistency of our measurement has been established, similarity between dimensions can be examined further in future research, specifically on trust factors model.

To evaluate the discriminant validity of our trust measurement, we assessed the relation between our measurement and participants' trust propensity level. As we can see in the last row of Table 13 and Table 14, based on Fornell and Larcker criterion and HTMT ratio, the results suggested low similarity between our trust factors and trust propensity. These results supported the discriminant validity of our measurement instrument, which was different from trust propensity.

## Criterion Validity: Concurrent Validity

To established concurrent validity, our measurement should be significantly correlated to some outcome measured at the same time, meaning the trust factors should be correlated to the trust level. As shown in Table 15, each domain (trust factor) of our measurement is positively correlated to trust level and the relationships are all significant (p; .05). These results suggested that our

Domains	Correlation	p-value	lower CI	upper CI
reliability	0.786	0.000	0.708	0.857
technical competence	0.829	0.000	0.764	0.895
understandability	0.581	0.000	0.438	0.722
personal attachment	0.814	0.000	0.756	0.867
helpfulness	0.813	0.000	0.755	0.871
faith	0.846	0.000	0.797	0.890
user autonomy	0.782	0.000	0.708	0.855
institution credibility	0.779	0.000	0.713	0.841

Table 15: Correlation Table: Domains (Trust Factors) with Trust. Correlation Coefficient, p-value, lower Confidence Interval, upper Confidence Interval.

trust measurement hold concurrent validity towards subjective single-question trust measurement.

# Criterion Validity: Predictive Validity

For predictive validity, first, we regressed trust level data on all items rating, with linear regression model: trust equal all trust factors items. As shown in Table 16, two faith items and one institution credibility item were significantly predictive of trust level (p < .05). Based on these results, the relationships between trust and all trust factors, except Faith, were not linear. Thus, trust level could not be predicted with 16 items and the predictive validity of our measurement was not supported.

### 10. Discussion

The main purpose of this study was to develop and validate a scale to measure the trust attitude of lay people towards AI systems. Eight domain factors and 16 items were developed through deductive and inductive methods and have been thoroughly evaluated. The proposed measurement instrument was administered as an online survey. The results have shown that the measure-

Items	Reg Coeff	p-value	lower CI	upper CI	$R^2$
reliability1	-0.014	0.813	-0.128	0.100	0.713
reliability2	0.057	0.259	-0.042	0.155	0.612
technicalcompetence1	0.113	0.057	-0.004	0.229	0.656
technical competence 2	0.113	0.106	-0.024	0.250	0.703
under standability 1	-0.027	0.638	-0.137	0.084	0.698
understandability 2	0.016	0.764	-0.090	0.123	0.652
personalattachment1	0.074	0.157	-0.029	0.177	0.698
personal attachment 2	0.024	0.672	-0.087	0.135	0.808
helpful1	0.027	0.671	-0.096	0.149	0.747
helpful2	0.092	0.154	-0.035	0.220	0.797
faith1	0.230	0.000	0.109	0.351	0.791
faith2	0.214	0.001	0.085	0.342	0.830
userautonomy1	-0.069	0.223	-0.181	0.042	0.705
userautonomy2	0.078	0.142	-0.026	0.182	0.702
institution credibility 1	0.167	0.005	0.050	0.285	0.800
institutioncredibility2	-0.118	0.084	-0.251	0.016	0.841

Table 16: Regression Table: Measurement Items with Trust. Regression Coefficient, p-value, lower Confidence Interval, upper Confidence Interval, and Coefficient of Determination  $(\mathbb{R}^2)$ 

ment instrument is internally consistent and stable; with established content and construct validity.

Guidelines from APA [38] and literature [34, 35] have emphasised the importance of content validity, criterion-related validity, and internal consistency of measurement instruments. However, to date, not all studies that use psychological construct instruments have demonstrated both their validity and reliability. The internal consistency of our instrument was established for all measured domains with alpha and omega coefficients. In recent years, research has encouraged the use of the omega coefficient and considers it a better choice than the alpha coefficient for assessing [117] reliability. However, the alpha coefficient is still a popular choice among trust measurement instruments [62, 79, 26]. A possible explanation is that alpha is considered more familiar than omega, and the difference between alpha and omega is also believed to be small [141].

In addition to internal consistency, stability over time was assessed by testretest reliability and resulted in a high correlation (ICC) for the 1-hour time
between survey administrations. It should be noted that test-retest reliability
requires a short duration between administrations to allow changes to occur, but
still long enough to prevent fatigue and preserve memory [85]. We repeated the
survey with a different group of participants, as it is advisable to use different
samples [39] while helping to increase the generalisability of the measurement
[35]. Stability over time indicates repeatability of the measurement, and a
measurement cannot be valid if it cannot be repeated.

Although reliability is necessary and should be clearly reported for all measures in a study, it is not a sufficient condition for measurement validity. Studies report that most research on trust in the Human Interaction-AI uses readily available trust questionnaires derived from Jian et al. [44], Chien et al. [60], Merritt [28], and Muir [74]; which are often not assessed for more than internal consistency reliability [142]. In addition, half of the human-AI trust papers modified the original questionnaire intended for a different system (automation) without providing any validation tests.

A measurement is considered valid when it measures what it is intended to

measure, with two main types of validity: content validity and construct validity [36]. Content validity refers to evidence of the representativeness of the content and technical quality of the instrument items, which in our study was assessed quantitatively using the content validity index (CVI) and Cohen's kappa coefficient  $(\kappa)$ , and qualitatively using cognitive interviews. Most popular studies have failed to clearly indicate the content validity assessment of their questionnaires [26, 60, 28], which can cause problems in studies that use questionnaires even without changes or modifications. We further validated the measurement qualitatively with cognitive interviews. For questionnaires aimed at specific groups of people, it is very important to check whether the designed instrument is interpreted correctly. In the cognitive interview, we found that participants were able to describe their mental processes when answering questions from our measurement instrument. As none of the participants had experience in AI medical support systems, their reflection processes were related to more common AI applications, such as, Google Maps, Google Translate, Instagram Recommendations, etc. This implies that the measurement instrument was designed for a specific group of people. This implies that our trust measurement instrument might be used to evaluate AI systems in general and not limited to AI medical support systems. This implication may be somewhat limited as a study has shown that trust in technology and trust in medical technology have the same attributes but are considered to be different constructs [143].

The domains included in our measurement were also tested to check if our hypothetical structure of the eight factors model fits the items. The dimensions are perceived reliability, perceived technical competence, perceived understandability, faith, personal attachment, perceived helpfulness, user autonomy, and institution credibility; with two items on each dimension. Confirmatory factor analysis (CFA) was performed, and the results have shown that the structural validity of the instrument was established. Based on the CFA model, construct validity was analysed, which refers to the evidence of instrument capability to measure a construct that is not directly observable. In our study, we assessed the construct validity: convergent validity and discriminant validity, of our mea-

surement instrument using Fornell-Larcker's criterion [132]. Convergent validity was established between all domain items, with all composite reliability (CR) values and AVE values passing the minimum threshold, proving that all domains explain a substantial amount of variance in each indicator. As mentioned in the previous section, the Fornell-Larcker criterion requires the AVE to be greater than 0.5, and research by Hair et al. suggested that not only AVE should be greater than 0.5 but CR should also be greater than 0.7 [55].

Discriminant validity was tested for the overall measurement and also for between domains. Based on Fornell-Larcker's and HTMT ratio criterion, the discriminant validity between trust propensity and our trust measurement was established. This finding supports the work of other studies in trust theory that distinctly defined trust and trust propensity as conceptually different [43, 73, 46, 144]. In contrast, discriminant validity between all domains was not established. Based on Fornell-Larcker's and HTMT ratio criterion, six out of eight domains were not demonstrating discriminant validity: reliability-technical competence, reliability-helpfulness, helpfulness-technical competence, faith-personal attachment, and institution credibility-faith. The high correlation between domains indicates a potential problem in some of the domain differentiation. Even though discriminant validity is a matter of degree instead of binary [140], and only five out of 28 correlations demonstrated low discriminant validity, the possible cause should be analysed further. There is a high chance for factors affecting trust, affect each other as well.

Criterion-related validity was assessed for both concurrent validity and predictive validity. The "criteria" we used was trust level, and concurrent validity was established. According to Raykov & Marcoulides [41], concurrent validity is often omitted from validation study because it has two major constraints which are: the availability of appropriate criterion variables or "gold-standards" and the large sampling errors for small sample size. Even though our sample size was adequate, the assessment was ran on the assumption that single-question trust level is the gold-standard of trust measurement. This proposed a question for future studies on how to appropriately assess concurrent validity on a

trust measurement instrument. It is argued that trust is context heavy, where different researchers could define trust in a widely different way, and currently, a widely accepted and used measurement instrument for trust from an empirical standpoint [145] has not emerged to be constituted as the gold standard. Another important point to note, concurrent validity is a part of criterion validity, which is the validity of decision made by measurement. Taken together, a reasonable approach to assess concurrent validity of a trust measurement instrument is with trust-related behaviour (decision) measurement, such as reliance. Same with predictive validity, a more appropriate approach to assess predictive validity of a trust measurement instrument is by means of a trust-related behaviour (decision) measurement. When a trust measurement instrument established its predictive validity, it means the measurement scores can determine future outcomes. Even though predictive validity was not demonstrated for our measurement instrument to predict trust level, this did not diminish the other types of validity of our measurement instrument. There is also the possibility of a non-linear relation between trust factors and trust level, and further tests need to be conducted to explore these relations better.

To analysed the content validity, convergent validity, discriminant validity, concurrent validity, and predictive validity, different tests and assessments were performed. As described in the result subsections for each validity test, there are discussions in psychometric field on the better method and thresholds recommended, such as HTMT ratio criterion as alternative method from Fornell-Lercker criterion criticism [133]. Another example of discussion was made to both HTMT and Fornell-Lercker criterion, wich were argued to be ineffective in evaluating convergent validity and discriminant validity [146]. However, applying Cheung and Wang [146] recommendation to conclude convergent validity (AVE is not significantly smaller than 0.5 and the standardised item factor loading item is not significantly less than 0) and discriminant validity (correlation less than 0.7), the previous results of our measurement instrument still hold. Criticism on methods is outside of our research scope, therefore, we applied various methods to empirically test our measurement instrument. The

results from all validity and reliability tests increased the confidence that the instrument correctly measured the construct intended to measure.

#### 11. Limitations

Several limitations of the study and possible future works can be highlighted on the three main stages: item development, measurement instrument development, and measurement instrument evaluation. In the item development stage, we looked at the literature from different fields, not only on trust scales but also on trust factors and models. Even though we proposed eight factors that could affect trust to form our trust measurement instrument, we have not analysed closely the relation between factors, and no actual trust model for human-AI system interaction has been proposed. Future analysis and, if required, data collection should be conducted to develop the trust model. In the measurement instrument development stage, we did not include detailed demography questions in the survey administered. Additional demography questions could help understand trust based on the demographics better, and also could help development of the trust model.

Repeatability is an important part of a measurement instrument, and we conducted an additional survey to assess the test-retest reliability of our measurement. Since theory suggests that trust is dynamic and may vary over time, future longitudinal research utilising our measurement instrument would be appropriate to examine the test-retest reliability for a longer time period. Instead of just an hour, repeated survey with 1-7 days in between is recommended. Additionally, our repeated survey was conducted with the same topic: breast cancer detection applications. Replication with different types of AI systems, possibly outside of healthcare application, would be beneficial to developing the trust model and also solidify the generality of our measurement instrument. In the discussion section previously, we mentioned that further evaluation will be necessary to understand trust and develop the trust theory/model. A future study that involves the measurement of trust as an attitude and how it relates

to trust-related behaviour could also provide a valuable contribution.

In summary, replication and adaptation of our proposed trust measurement instrument are highly encouraged. The trust scale replication and validation by Spain et al. [142] was one of the reasons why Jian et al. [26] trust scale became the most well-cited trust scale in the human factors literature [147]. Replication and adaptation will not only to further prove the validity and generality of the measurement but also will help to understand trust and trust model in human-AI system interaction.

#### 12. Conclusion

Trust is an important concept, and trust in Artificial Intelligence is currently widely explored in various research fields. This study makes theoretical and practical contributions by developing and validating a scale to measure trust attitudes towards AI systems for laypeople. We proposed a trust measurement comprised of eight trust factors as the domains and two statement items for each domain, which make a total of 16 items. The reliability and validity of our measurement instrument were established and are expected to be used and adapted by future researchers to evaluate their AI systems.

The methodological approach to develop and evaluate trust measurement for human-AI interaction has been described and demonstrated. Carefully designed trust measurement instruments are not only fundamental to our understanding of trust but also ensure accurate measurement of trust, which is known to be a complex construct. By making the development of measurement instruments more approachable and transparent, we hope this paper can facilitate the advancement of our understanding of trust and trustworthiness in AI systems while complementing existing trust models.

## References

[1] E.C. 2020, On artificial intelligence - a european approach to excellence and trust, https://ec.europa.eu/info/sites/default/files/

- commission-white-paper-artificial-intelligence-feb2020\_en. pdf (2020).
- [2] U.D.I. Board, AI principles: Recommendations on the ethical use of artificial intelligence by the department of defense, https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB\$\_\$AI\$\_\$PRINCIPLES\$\_\$PRIMARY\$\_\$DOCUMENT.PDF (2019).
- [3] N.G. Committee, The ethical norms for the new generation artificial intelligence, china international research center for ai ethics and governance, https://ai-ethics-and-governance.institute/2021/09/27/the-ethical-norms-for-the-new-generation-artificial-intelligence-china/(2021).
- [4] D. Gefen, E. Karahanna, D. W. Straub, Trust and tam in online shopping: an integrated model, MIS quarterly 27 (1) (2003) 51–90.
- [5] Intel, U.s. healthcare leaders expect widespread adoption of artificial intelligence by 2023 intel newsroom, (Accessed on 02/10/2019) (2018).
- [6] J. Powles, H. Hodson, Google deepmind and healthcare in an age of algorithms, Health and technology 7 (4) (2017) 351–367.
- This [7] Forbes, health startup big government won deals—but inside. doctors flagged problems, https: //www.forbes.com/sites/parmyolson/2018/12/17/ this-health-startup-won-big-government-dealsbut-inside-doctors-flagged-problems/  $?sh=787a6355eabb\ (2018).$
- [8] PwC, Survey results: Why ai and robotics will define new health: Publications: Healthcare: Industries: Pwc, (Accessed on 02/10/2019) (2016).
- [9] J. M. Logg, J. A. Minson, D. A. Moore, Algorithm appreciation: People prefer algorithmic to human judgment, Organizational Behavior and Human Decision Processes 151 (2019) 90–103.

- [10] M. R. Cohen, J. L. Smetzer, Ismp medication error report analysis: Understanding human over-reliance on technology it's exelan, not exelon crash cart drug mix-up risk with entering a "test order", Hospital pharmacy 52 (1) (2017) 7.
- [11] K. Goddard, A. Roudsari, J. C. Wyatt, Automation bias: a systematic review of frequency, effect mediators, and mitigators, Journal of the American Medical Informatics Association 19 (1) (2011) 121–127.
- [12] K. L. Mosier, L. J. Skitka, Human decision makers and automated decision aids: Made for each other?, in: Automation and human performance, Routledge, 2018, pp. 201–220.
- [13] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gpdr, Harv. JL & Tech. 31 (2017) 841.
- [14] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Vaughan, H. Wallach, Manipulating and measuring model interpretability, arXiv preprint arXiv:1802.07810 (2018).
- [15] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, M. Kankanhalli, Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, ACM, 2018, p. 582.
- [16] Y. Zhang, Q. V. Liao, R. K. Bellamy, Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 295–305.
- [17] N. Salomons, M. Van Der Linden, S. S. Sebo, B. Scassellati, Humans conform to robots: Disambiguating trust, truth, and conformity, in: 2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE, 2018, pp. 187–195.

- [18] T. Bridgwater, M. Giuliani, A. van Maris, G. Baker, A. Winfield, T. Pipe, Examining profiles for robotic risk assessment: does a robot's approach to risk affect user trust?, in: 2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE, 2020, pp. 23–31.
- [19] S. Antifakos, N. Kern, B. Schiele, A. Schwaninger, Towards improving trust in context-aware systems by displaying system confidence, in: Proceedings of the 7th international conference on Human computer interaction with mobile devices & services, ACM, 2005, pp. 9–14.
- [20] B. F. Yuksel, P. Collisson, M. Czerwinski, Brains or beauty: How to engender trust in user-agent interactions, ACM Transactions on Internet Technology (TOIT) 17 (1) (2017) 1–20.
- [21] S. Feng, J. Boyd-Graber, What can ai do for me? evaluating machine learning interpretations in cooperative play, in: Proceedings of the 24th International Conference on Intelligent User Interfaces, 2019, pp. 229–239.
- [22] M. Yin, J. Wortman Vaughan, H. Wallach, Understanding the effect of accuracy on trust in machine learning models, in: Proceedings of the 2019 chi conference on human factors in computing systems, 2019, pp. 1–12.
- [23] A. Bussone, S. Stumpf, D. O'Sullivan, The role of explanations on trust and reliance in clinical decision support systems, in: 2015 International Conference on Healthcare Informatics, IEEE, 2015, pp. 160–169.
- [24] M. Li, B. E. Holthausen, R. E. Stuck, B. N. Walker, No risk no trust: Investigating perceived risk in highly automated driving, in: Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, 2019, pp. 177–185.
- [25] K. Schaefer, The perception and measurement of human-robot trust (2013).

- [26] J.-Y. Jian, A. M. Bisantz, C. G. Drury, Foundations for an empirically determined scale of trust in automated systems, International journal of cognitive ergonomics 4 (1) (2000) 53–71.
- [27] B. M. Muir, Operators' trust in and use of automatic controllers in a supervisory process control task. (2002).
- [28] S. M. Merritt, Affective processes in human–automation interactions, Human Factors 53 (4) (2011) 356–370.
- [29] D. H. McKnight, V. Choudhury, C. Kacmar, The impact of initial consumer trust on intentions to transact with a web site: a trust building model, The journal of strategic information systems 11 (3-4) (2002) 297–323.
- [30] B. Ghai, Q. V. Liao, Y. Zhang, R. Bellamy, K. Mueller, Explainable active learning (xal) toward ai explanations as interfaces for machine teachers, Proceedings of the ACM on Human-Computer Interaction 4 (CSCW3) (2021) 1–28.
- [31] B. Yu, Y. Yuan, L. Terveen, Z. S. Wu, J. Forlizzi, H. Zhu, Keeping designers in the loop: Communicating inherent algorithmic trade-offs across multiple objectives, in: Proceedings of the 2020 ACM designing interactive systems conference, 2020, pp. 1245–1257.
- [32] H.-F. Cheng, R. Wang, Z. Zhang, F. O'Connell, T. Gray, F. M. Harper, H. Zhu, Explaining decision-making algorithms through ui: Strategies to help non-expert stakeholders, in: Proceedings of the 2019 chi conference on human factors in computing systems, 2019, pp. 1–12.
- [33] O. Vereschak, G. Bailly, B. Caramiaux, How to evaluate trust in ai-assisted decision making? a survey of empirical methodologies, Proceedings of the ACM on Human-Computer Interaction 5 (CSCW2) (2021) 1–39.
- [34] G. O. Boateng, T. B. Neilands, E. A. Frongillo, H. R. Melgar-Quiñonez, S. L. Young, Best practices for developing and validating scales for health,

- social, and behavioral research: a primer, Frontiers in public health 6 (2018) 149.
- [35] T. R. Hinkin, A brief tutorial on the development of measures for use in survey questionnaires, Organizational research methods 1 (1) (1998) 104–121.
- [36] K. R. Murphy, C. O. Davidshofer, Psychological testing, Principles, and Applications, Englewood Cliffs 18 (1988).
- [37] T. Raykov, Scale reliability, cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components, Multivariate behavioral research 32 (4) (1997) 329–353.
- [38] A. E. R. Association, A. P. Association, N. C. on Measurement in Education, et al., Standards for educational and psychological testing, American Educational Research Association, 1999.
- [39] D. T. Campbell, D. W. Fiske, Convergent and discriminant validation by the multitrait-multimethod matrix., Psychological bulletin 56 (2) (1959) 81.
- [40] J. C. Nunnally, Psychometric theory 3E, Tata McGraw-hill education, 1994.
- [41] T. Raykov, G. A. Marcoulides, Introduction to psychometric theory, Routledge, 2011.
- [42] R. F. DeVellis, C. T. Thorpe, Scale development: Theory and applications, Sage publications, 2021.
- [43] R. C. Mayer, J. H. Davis, F. D. Schoorman, An integrative model of organizational trust, Academy of management review 20 (3) (1995) 709– 734.
- [44] J. D. Lee, K. A. See, Trust in automation: Designing for appropriate reliance, Human factors 46 (1) (2004) 50–80.

- [45] K. E. Schaefer, J. Y. Chen, J. L. Szalma, P. A. Hancock, A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems, Human factors 58 (3) (2016) 377–400.
- [46] D. H. Mcknight, M. Carter, J. B. Thatcher, P. F. Clay, Trust in a specific technology: An investigation of its components and measures, ACM Transactions on Management Information Systems (TMIS) 2 (2) (2011) 12.
- [47] X. Li, T. J. Hess, J. S. Valacich, Why do we trust new technology? a study of initial trust formation with organizational information systems, The Journal of Strategic Information Systems 17 (1) (2008) 39–71.
- [48] K. E. Oleson, D. R. Billings, V. Kocsis, J. Y. Chen, P. A. Hancock, Antecedents of trust in human-robot collaborations, in: 2011 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), IEEE, 2011, pp. 175–178.
- [49] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, R. Parasuraman, A meta-analysis of factors affecting trust in human-robot interaction, Human factors 53 (5) (2011) 517–527.
- [50] J. Meyer, J. D. Lee, Trust, reliance, and compliance. (2013).
- [51] R. M. Perloff, The dynamics of persuasion: Communication and attitudes in the 21st century, Routledge, 1993.
- [52] Y. Xie, I. P. Bodala, D. C. Ong, D. Hsu, H. Soh, Robot capability and intention in trust-based decisions across tasks, in: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE, 2019, pp. 39–47.
- [53] R. Likert, A technique for the measurement of attitudes., Archives of psychology (1932).

- [54] M. Biasutti, S. Frate, A validity and reliability study of the attitudes toward sustainable development scale, Environmental Education Research 23 (2) (2017) 214–230.
- [55] J. F. Hair Jr, M. LDS Gabriel, D. d. Silva, S. Braga Junior, Development and validation of attitudes measurement scales: fundamental and practical aspects, RAUSP Management Journal 54 (4) (2019) 490–507.
- [56] J. K. Rempel, J. G. Holmes, M. P. Zanna, Trust in close relationships., Journal of personality and social psychology 49 (1) (1985) 95.
- [57] J. D. Lee, N. Moray, Trust, self-confidence, and operators' adaptation to automation, International journal of human-computer studies 40 (1) (1994) 153–184.
- [58] S. Lewandowsky, M. Mundy, G. Tan, The dynamics of trust: comparing humans to automation., Journal of Experimental Psychology: Applied 6 (2) (2000) 104.
- [59] N. Moray, T. Inagaki, M. Itoh, Adaptive automation, trust, and selfconfidence in fault management of time-critical tasks., Journal of experimental psychology: Applied 6 (1) (2000) 44.
- [60] S.-Y. Chien, M. Lewis, K. Sycara, J.-S. Liu, A. Kumru, The effect of culture on trust in automation: reliability and workload, ACM Transactions on Interactive Intelligent Systems (TiiS) 8 (4) (2018) 1–31.
- [61] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, H. P. Beck, The role of trust in automation reliance, International journal of human-computer studies 58 (6) (2003) 697–718.
- [62] M. Madsen, S. Gregor, Measuring human-computer trust, in: 11th australasian conference on information systems, Vol. 53, Citeseer, 2000, pp. 6–8.

- [63] D. A. Cook, T. J. Beckman, Current concepts in validity and reliability for psychometric instruments: theory and application, The American journal of medicine 119 (2) (2006) 166–e7.
- [64] S. Easterbrook, J. Singer, M.-A. Storey, D. Damian, Selecting empirical methods for software engineering research, in: Guide to advanced empirical software engineering, Springer, 2008, pp. 285–311.
- [65] K. B. Wright, Researching internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services, Journal of computer-mediated communication 10 (3) (2005) JCMC1034.
- [66] I. Mansell, G. Bennett, R. Northway, D. Mead, L. Moseley, The learning curve: the advantages and disadvantages in the use of focus groups as a method of data collection, Nurse Researcher 11 (4) (2004).
- [67] C. Barter, E. Renold, The use of vignettes in qualitative research (1999).
- [68] J. Finch, The vignette technique in survey research, Sociology 21 (1) (1987) 105–114.
- [69] D. H. Thom, B. Campbell, fotlg inalresearch patient-physician trust: An exploratory study, The Journal of family practice 44 (2) (1997) 169.
- [70] W. J. Winkelman, K. J. Leonard, P. G. Rossos, Patient-perceived usefulness of online electronic medical records: employing grounded theory in the development of information and communication technologies for use by patients living with chronic illness, Journal of the American Medical Informatics Association 12 (3) (2005) 306–314.
- [71] K. A. Hoff, M. Bashir, Trust in automation: Integrating empirical evidence on factors that influence trust, Human factors 57 (3) (2015) 407–434.
- [72] J. Morrow Jr, M. H. Hansen, A. W. Pearson, The cognitive and affective antecedents of general trust within cooperative organizations, Journal of managerial issues (2004) 48–64.

- [73] D. M. Rousseau, S. B. Sitkin, R. S. Burt, C. Camerer, Not so different after all: A cross-discipline view of trust, Academy of management review 23 (3) (1998) 393–404.
- [74] B. M. Muir, Trust between humans and machines, and the design of decision aids, International journal of man-machine studies 27 (5-6) (1987) 527–539.
- [75] R. Sinha, K. Swearingen, The role of transparency in recommender systems, in: CHI'02 extended abstracts on Human factors in computing systems, ACM, 2002, pp. 830–831.
- [76] J. L. Herlocker, J. A. Konstan, J. Riedl, Explaining collaborative filtering recommendations, in: Proceedings of the 2000 ACM conference on Computer supported cooperative work, ACM, 2000, pp. 241–250.
- [77] M. Hengstler, E. Enkel, S. Duelli, Applied artificial intelligence and trust—the case of autonomous vehicles and medical assistance devices, Technological Forecasting and Social Change 105 (2016) 105–120.
- [78] D. Gefen, E-commerce: the role of familiarity and trust, Omega 28 (6) (2000) 725–737.
- [79] D. H. McKnight, L. L. Cummings, N. L. Chervany, Initial trust formation in new organizational relationships, Academy of Management review 23 (3) (1998) 473–490.
- [80] C. Johnson-George, W. C. Swap, Measurement of specific interpersonal trust: Construction and validation of a scale to assess trust in a specific other., Journal of personality and social psychology 43 (6) (1982) 1306.
- [81] D. Johnson, K. Grayson, Cognitive and affective trust in service relationships, Journal of Business research 58 (4) (2005) 500–507.
- [82] F. M. Verberne, J. Ham, C. J. Midden, Trust in smart systems: Sharing driving goals and giving information to increase trustworthiness and acceptability of smart systems in cars, Human factors 54 (5) (2012) 799–810.

- [83] P. W. Schultz, Environmental attitudes and behaviors across cultures, Online readings in psychology and culture 8 (1) (2002) 2307–0919.
- [84] A. O'Cathain, K. J. Thomas, "any other comments?" open questions on questionnaires—a bane or a bonus to research?, BMC medical research methodology 4 (1) (2004) 1–7.
- [85] K. S. Schultz, D. J. Whitney, Measurement theory in action, Thousand Oaks (2005).
- [86] L. J. Cronbach, Coefficient alpha and the internal structure of tests, psychometrika 16 (3) (1951) 297–334.
- [87] A. R. Jonsen, M. Siegler, W. J. Winslade, Clinical ethics a practical approach to ethical decisions in clinical medicine (1982).
- [88] R. Rowe, M. Calnan, Trust relations in health care—the new agenda, The European Journal of Public Health 16 (1) (2006) 4–6.
- [89] J. E. Croker, D. R. Swancutt, M. J. Roberts, G. A. Abel, M. Roland, J. L. Campbell, Factors affecting patients' trust and confidence in gps: evidence from the english national gp patient survey, BMJ open 3 (5) (2013) e002762.
- [90] S. Safikhani, M. Sundaram, Y. Bao, P. Mulani, D. A. Revicki, Qualitative assessment of the content validity of the dermatology life quality index in patients with moderate to severe psoriasis, Journal of dermatological treatment 24 (1) (2013) 50–59.
- [91] V. Quiroz, D. Reinero, P. Hernández, J. Contreras, R. Vernal, P. Carvajal, Development of a self-report questionnaire designed for population-based surveillance of gingivitis in adolescents: assessment of content validity and reliability, Journal of Applied Oral Science 25 (4) (2017) 404–411.
- [92] R. M. Guion, Content validity—the source of my discontent, Applied Psychological Measurement 1 (1) (1977) 1–10.

- [93] F. F. Morgado, J. F. Meireles, C. M. Neves, A. Amaral, M. E. Ferreira, Scale development: ten main limitations and recommendations to improve future research practices, Psicologia: Reflexão e Crítica 30 (2017).
- [94] L. L. Davis, Instrument review: Getting the most from a panel of experts, Applied nursing research 5 (4) (1992) 194–197.
- [95] S. N. Haynes, D. Richard, E. S. Kubany, Content validity in psychological assessment: A functional approach to concepts and methods., Psychological assessment 7 (3) (1995) 238.
- [96] D. F. Polit, C. T. Beck, The content validity index: are you sure you know what's being reported? critique and recommendations, Research in nursing & health 29 (5) (2006) 489–497.
- [97] M. R. Lynn, Determination and quantification of content validity., Nursing research (1986).
- [98] C. A. Wynd, B. Schmidt, M. A. Schaefer, Two quantitative approaches for estimating content validity, Western journal of nursing research 25 (5) (2003) 508–518.
- [99] J. Cohen, A coefficient of agreement for nominal scales, Educational and psychological measurement 20 (1) (1960) 37–46.
- [100] P. C. Beatty, G. B. Willis, Research synthesis: The practice of cognitive interviewing, Public opinion quarterly 71 (2) (2007) 287–311.
- [101] G. B. Willis, Cognitive interviewing: A tool for improving questionnaire design, sage publications, 2004.
- [102] R. Tourangeau, Cognitive aspects of survey measurement and mismeasurement, International Journal of Public Opinion Research 15 (1) (2003) 3–7.
- [103] A. Seif, Educational measurement, assessment and evaluation, Tehran: Doran Publications 128 (2004).

- [104] E. Abdollahpour, S. Nejat, M. Nourozian, R. Majdzadeh, The process of content validity in instrument development, Iranian Epidemiology 6 (4) (2010) 66–74.
- [105] D. V. Cicchetti, S. A. Sparrow, Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior., American journal of mental deficiency (1981).
- [106] R. L. Gorsuch, Exploratory factor analysis, in: Handbook of multivariate experimental psychology, Springer, 1988, pp. 231–258.
- [107] J. C. Nunnally, Psychometric theory. (1967).
- [108] E. Guadagnoli, W. F. Velicer, Relation of sample size to the stability of component patterns., Psychological bulletin 103 (2) (1988) 265.
- [109] A. L. Comrey, Factor-analytic methods of scale development in personality and clinical psychology., Journal of consulting and clinical psychology 56 (5) (1988) 754.
- [110] L. A. Clark, D. Watson, Constructing validity: Basic issues in objective scale development., Psychological assessment 7 (3) (1995) 309.
- [111] H. F. Kaiser, The varimax criterion for analytic rotation in factor analysis, Psychometrika 23 (3) (1958) 187–200.
- [112] M. S. Bartlett, A note on the multiplying factors for various  $\chi$  2 approximations, Journal of the Royal Statistical Society. Series B (Methodological) (1954) 296–298.
- [113] S. B. MacKenzie, P. M. Podsakoff, R. Fetter, Organizational citizenship behavior and objective productivity as determinants of managerial evaluations of salespersons' performance, Organizational behavior and human decision processes 50 (1) (1991) 123–150.

- [114] L.-t. Hu, P. M. Bentler, Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, Structural equation modeling: a multidisciplinary journal 6 (1) (1999) 1–55.
- [115] M. Porta, A dictionary of epidemiology, Oxford university press, 2014.
- [116] T. Raykov, G. A. Marcoulides, A direct latent variable modeling based method for point and interval estimation of coefficient alpha, Educational and Psychological Measurement 75 (1) (2015) 146–156.
- [117] E. Cho, S. Kim, Cronbach's coefficient alpha: Well known but poorly understood, Organizational research methods 18 (2) (2015) 207–230.
- [118] S. B. Green, Y. Yang, Commentary on coefficient alpha: A cautionary tale, Psychometrika 74 (1) (2009) 121–135.
- [119] R. P. McDonald, Test theory: A unified treatment, psychology press, 2013.
- [120] A. Maydeu-Olivares, D. L. Coffman, W. M. Hartmann, Asymptotically distribution-free (adf) interval estimation of coefficient alpha., Psychological methods 12 (2) (2007) 157.
- [121] P. M. Bentler, Alpha, dimension-free, and model-based internal consistency reliability, Psychometrika 74 (1) (2009) 137–143.
- [122] D. L. Streiner, G. R. Norman, J. Cairney, Health measurement scales: a practical guide to their development and use, Oxford University Press, USA, 2015.
- [123] K. Pearson, Vii. mathematical contributions to the theory of evolution.—iii. regression, heredity, and panmixia, Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character (187) (1896) 253–318.
- [124] J. P. Weir, Quantifying test-retest reliability using the intraclass correlation coefficient and the sem, The Journal of Strength & Conditioning Research 19 (1) (2005) 231–240.

- [125] W. Kroll, A note on the coefficient of intraclass correlation as an estimate of reliability, Research Quarterly. American Association for Health, Physical Education and Recreation 33 (2) (1962) 313–316.
- [126] J. Ludbrook, Statistical techniques for comparing measurers and methods of measurement: a critical review, Clinical and Experimental Pharmacology and Physiology 29 (7) (2002) 527–536.
- [127] W. W. Chin, How to write up and report pls analyses, in: Handbook of partial least squares, Springer, 2010, pp. 655–690.
- [128] J. Henseler, C. M. Ringle, R. R. Sinkovics, The use of partial least squares path modeling in international marketing, in: New challenges to international marketing, Emerald Group Publishing Limited, 2009.
- [129] R. P. Bagozzi, Y. Yi, On the evaluation of structural equation models, Journal of the academy of marketing science 16 (1) (1988) 74–94.
- [130] L. Mâță, O. Clipa, K. Tzafilkou, The development and validation of a scale to measure university teachers' attitude towards ethical use of information technology for a sustainable education, Sustainability 12 (15) (2020) 6268.
- [131] J. Hu, R. C. Liden, Making a difference in the teamwork: Linking team prosocial motivation to team processes and effectiveness, Academy of Management Journal 58 (4) (2015) 1102–1127.
- [132] C. Fornell, D. F. Larcker, Evaluating structural equation models with unobservable variables and measurement error, Journal of marketing research 18 (1) (1981) 39–50.
- [133] J. Henseler, C. M. Ringle, M. Sarstedt, A new criterion for assessing discriminant validity in variance-based structural equation modeling, Journal of the academy of marketing science 43 (1) (2015) 115–135.
- [134] A. H. Gold, A. Malhotra, A. H. Segars, Knowledge management: An organizational capabilities perspective, Journal of management information systems 18 (1) (2001) 185–214.

- [135] C. M. Voorhees, M. K. Brady, R. Calantone, E. Ramirez, Discriminant validity testing in marketing: an analysis, causes for concern, and proposed remedies, Journal of the academy of marketing science 44 (1) (2016) 119–134.
- [136] H. Le, F. L. Schmidt, D. J. Putka, The multifaceted nature of measurement artifacts and its implications for estimating construct-level relationships, Organizational Research Methods 12 (1) (2009) 165–200.
- [137] D. J. Woehr, D. J. Putka, M. C. Bowler, An examination of g-theory methods for modeling multitrait—multimethod data: Clarifying links to construct validity and confirmatory factor analysis, Organizational Research Methods 15 (1) (2012) 134–161.
- [138] J. Loevinger, Objective tests as instruments of psychological theory, Psychological reports 3 (3) (1957) 635–694.
- [139] M. Ab Hamid, W. Sami, M. M. Sidek, Discriminant validity assessment: Use of fornell & larcker criterion versus htmt criterion, in: Journal of Physics: Conference Series, Vol. 890, IOP Publishing, 2017, p. 012163.
- [140] M. Rönkkö, E. Cho, An updated guideline for assessing discriminant validity, Organizational Research Methods 25 (1) (2022) 6–14.
- [141] L. Deng, W. Chan, Testing the difference between reliability coefficients alpha and omega, Educational and psychological measurement 77 (2) (2017) 185–203.
- [142] R. D. Spain, E. A. Bustamante, J. P. Bliss, Towards an empirically developed scale for system trust: Take two, in: Proceedings of the human factors and ergonomics society annual meeting, Vol. 52, SAGE Publications Sage CA: Los Angeles, CA, 2008, pp. 1335–1339.
- [143] E. N. Montague, B. M. Kleiner, W. W. Winchester III, Empirically understanding trust in medical technology, International Journal of Industrial Ergonomics 39 (4) (2009) 628–634.

- [144] J. A. Colquitt, B. A. Scott, J. A. LePine, Trust, trustworthiness, and trust propensity: a meta-analytic test of their unique relationships with risk taking and job performance., Journal of applied psychology 92 (4) (2007) 909.
- [145] M. L. Watson, Can there be just one trust? a cross-disciplinary identification of trust definitions and measurement, The Institute for Public Relations (2005) 1–25.
- [146] G. W. Cheung, C. Wang, Current approaches for assessing convergent and discriminant validity with sem: Issues and solutions, in: Academy of management proceedings, Vol. 2017, Academy of Management Briarcliff Manor, NY 10510, 2017, p. 12706.
- [147] R. S. Gutzwiller, E. K. Chiou, S. D. Craig, C. M. Lewis, G. J. Lematta, C.-P. Hsiung, Positive bias in the 'trust in automated systems survey'? an examination of the jian et al.(2000) scale, in: Proceedings of the human factors and ergonomics society annual meeting, Vol. 63, SAGE Publications Sage CA: Los Angeles, CA, 2019, pp. 217–221.