Human and AI Trust: Trust Attitude Measurement Instrument Development

RETNO LARASATI, Knowledge Media Institute, The Open University, United Kingdom
ANNA DE LIDDO, Knowledge Media Institute, The Open University, United Kingdom
ENRICO MOTTA, Knowledge Media Institute, The Open University, United Kingdom

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, in the context of AI medical support systems (specifically cancer/health prediction). 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.

CCS Concepts: • **Human-centered computing** → *Human Computer Interaction*; **Human computer interaction (HCI)**.

Additional Key Words and Phrases: Human-AI Trust, Human-AI Interaction, Trust Factors, Trust Measurement

ACM Reference Format:
Retno Larasati, Anna De Liddo, and Enrico Motta. 2023. Human and AI Trust: Trust Attitude Measurement Instrument Development. 1, 1 (April 2023), 14 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The call for trustworthy AI was made by formal national institutions around the world (Europe [1], USA [9], China [18]), and trustworthiness in AI systems is increasingly becoming an ethical and societal need. Trust is a crucial factor in all kinds of relationships, and considered as humans' primary reason for acceptance [28]. 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 [40]. In 2018, a government-backed AI healthcare application, Babylon, also received criticism for the inaccuracies in diagnosis, [25] 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 [70].

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 [50]. As AI is increasingly embedded in all sorts of largely adopted systems, research evidence indicates that users tend to over-trust and continue to rely on a system even when it malfunctions [17]. 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' [30, 61]. People often neglect automation bias and tend to trust a system when they think the answer came from an algorithm

rather than another person [50]. This misplaced trust (distrust or over-trust) has motivated research on trustworthy AI, whether via developing new forms of explainable AI or AI transparency [3, 69, 87, 94].

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 [4, 11, 76], reliance [24, 93], and also straightforward trust rating [11, 12, 91] amongst other evaluation methods. Aside from the context specific nature of trust [47], 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 [1] [77], human-automation trust [41, 57, 63], and human-technology trust [55], which have been used interchangeably between sub-fields. Combination of existing questionnaires was also implemented to achieve usable measurement tool [14, 29, 92]. However, most of the research has not conducted thorough validation test [14, 29, 85], to their adapted questionnaires. This lack of measurement 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 [5, 10, 13, 22, 37, 64, 65, 71, 72]. 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).

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 [53]. 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 [45][78], trust in information systems [54][48], and trust in robots[66][34].

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 [58], 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" [67], it is said to be externally non-observable [53, 90]. Psychological constructs are determined by psychological factors and can therefore be measured using self-reports, attitude scales, or questionnaires, for example, the Likert Scale [49]. The Likert scale is one of the scales that has been widely used and supported by the attitude

[1]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.

measurement literature [8, 33, 65]. 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) [2]. Several disciplines, such as psychology [73] and management [53], have been looking at human trust in technology. In particular, much work has been done investigating trust in human-automation interaction [44–46, 59, 78] and human other technologies interaction [48, 55]. However, only some of these studies have included measurement scales [15, 23, 41, 77]. Several trust measurement scales have become recurring trust scales used in human-AI research. One such scale, Jian et al.'s [41], is reported to be the most cited trust scale in human factors research. The scale by Jian et al. was developed for human-automated systems trust [41], the survey questions comprising the scale are very generic, which can be one of the main reasons for its re-usability.

Madsen and Gregor [52] 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. Through the scale validation process, high internal consistency (Cronbach's alpha > 0.94) and construct validity were established, with poor criterion-related validity. McKnight developed another trust measurement instrument to capture the trust relationship between users and specific technologies [54]. This scale was developed based on an understanding of trust in the broader context of society and previous research on human-human trust. Through evaluation, this instrument show good reliability (Cronbach's alpha > 0.9), construct validity, and criterion-related validity. 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.

MEASUREMENT INSTRUMENT DEVELOPMENT PROCESS

To develop a sound human trust measurement instrument, we followed recommendations by previous research in psychometric [10, 37]. 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 [5]. 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.

3 ITEM DEVELOPMENT

3.1 Methods

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 factors (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

[2]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

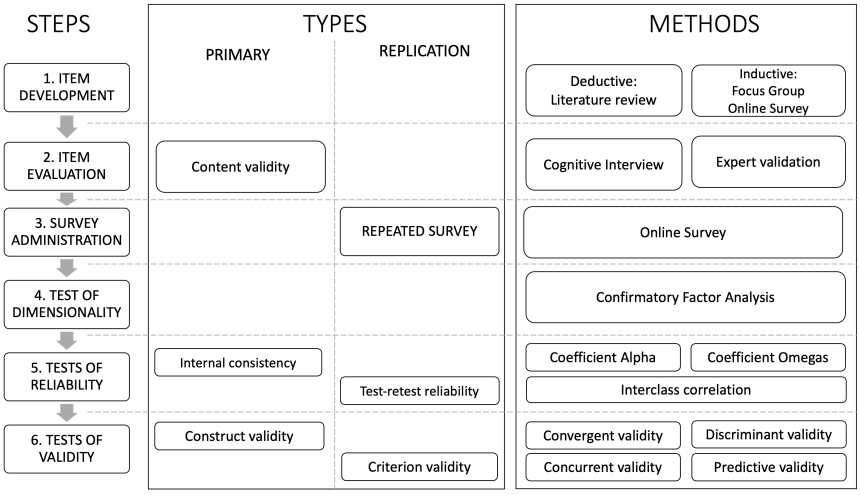

Fig. 1. Six Steps Process To Develop a Sound Human Trust Measurement Instrument.

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).

3.2 Results: Trust Factors and Items Developed

3.2.1 Deductive Method. We have 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 [45][38][78]. 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[34, 66]. In the context of information systems, trust consists of basic personality trust, basic institutional trust, and basic system trust. [48, 54]. 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 [60]. 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

Trust Factors	Description
perceived technical competence	system is perceived to perform the tasks accurately and correctly based on the information that is input. [4, 23, 38, 45, 52, 53, 62, 78]
perceived reliability	system is perceived to be, in the usual sense of repeated, consistent functioning. [27, 38, 43, 45, 56, 62, 73, 78].
perceived understandability	user can form a mental model and predict future system behaviour. [35, 36, 45, 62, 78, 82]
personal attachment	user finds using the system agreeable, preferable, suits their personal taste. [42, 52, 78, 80, 84].
faith	user has faith in the future ability of the system to perform even in situations in which it is untried. [43, 52, 73, 78].
perceived helpfulness	system is perceived to provide adequate, effective, and responsive help. [45, 53, 54, 73, 78]

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

[52] and in human-automation [78] 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 [74]. Therefore, we looked at the literature on trust factors while not overly considering discipline-specific categorisations.

Table 1 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.

3.2.2 Inductive Method

. **Online Survey**

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 [6].

We 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 (from deductive method) 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 [79]. 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 importance of following statement: "The application would use appropriate methods to get results based on the information I input."

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 2). 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 to use*". Perceived reliability, perceived technical competence, and perceived helpfulness domains demonstrated excellent internal

consistency, with their alpha coefficient > 0.8 [20, 65]. 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.

	reliability			technical competence			understandability			personal attachment			helpfulness			faith		
	r1	r2	r3	tc1	tc2	tc3	u1	u2	u3	p1	p2	p3	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 2. Importance Rating: Initial Measurement Scale with Six Domains and Three Statement Items Each

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 five 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. "-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 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*. For *Institutional Credibility*, participants meant the users' trust is placed in the institution which regulates the certification of the AI system. *User Autonomy* is also deemed as important by participants. 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 [19, 75].

4 ITEM EVALUATION

4.1 Methods

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.

4.2 Results

4.2.1 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, Content Validity Index (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 [21, 68, 81] and "For Revision" for CVI > 0.7 [2]. 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 κ >=0.74, "Good" for 0.74>κ>=0.6, and "Fair" for 0.59> κ>=0.40 [16]. Since it is recommended to create a large number of items in the early stage of item development [37], we created five item statements for each domain, including two new added domains, as our second version of measurement instrument, making it 40 instrument items in total.

We selected two or more items passed the CVI and κ "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 κ of the Clarity ratings from the rest of 20 items, to decide if the item needs revision or not. 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.

4.2.2 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. 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 [88][7], 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 [88]. 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 attachment domains. Based on these results, none of the item statements were dropped and some modifications on the word choice were applied.

5 SURVEY ADMINISTRATION

At this stage, we then administered the survey as the main measurement instrument test. 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 [89]. 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 are two demographic items and two trust propensity questions. The second set of questions contains 16 statements from the trust 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 7-point 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. The Mechanical Turk tasks were up for two weeks and data from 300 participants were collected.

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 results.
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 responsively. 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 confident 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.

Table 3. 16 Survey Questions/Statements (Two Statements for Each Domain

MEASUREMENT INSTRUMENT EVALUATION

6 TEST OF DIMENSIONALITY

To test the dimension of the proposed measurement instrument, we used Confirmatory Factor Analysis (CFA). CFA investigates how well the hypothesised factor structure (model) fits with the data [51], which is trust that consist of

eight trust factors. 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 [10, 37, 39].

Based on the root mean square error of approximation (RMSEA = 0.046), standardized root mean square residual (SRMR = 0.023), comparative fit index (CFI = 0.987), and Tucker–Lewis index (TLI = 0.979), the hypothesised model fits well and does not need additional alteration.

7 TEST OF RELIABILITY

7.1 Internal Consistency

The internal consistency was assessed with alpha and omega coefficients calculated in R. Table 4 depicts that all alphas and omegas values are above 0.7, indicating internal consistency in all dimensions and overall measurement[65]. 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 4. Reliability tests for the measurement instrument.

7.2 Test-retest reliability from 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 Mechanical Turk tasks were up for one month and data from 304 participants were collected. 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.

8 TEST OF VALIDITY

8.1 Methods

Validity is the extent to which "evidence and theory support the interpretation of test scores required by the proposed use of the test" [5]. Validity assessment can be summarised in three main forms:

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

We have evaluated the content validity and face validity in the Item Evaluation stage, and in the next section we described the construct and criterion validity tests results.

8.2 Results

Construct Validity: Convergent Validity. A measurement can established convergent validity when the measurement domain/construct correlates highly with each other. All items composite reliability (CR) values are above 0.7 [33], all average variance extracted (AVE) are above 0.5, suggesting the convergent validity of measurement instrument. This result suggest that our measurement instrument could examines or measures trust in different ways while still yields similar results.

Construct Validity: Discriminant Validity. To evaluate the discriminant validity of our trust measurement, we assessed the relation between our measurement and participants' trust propensity level. The discriminant validity was assessed with Fornell and Larcker criterion [26] and Heterotrait-monotrait (HTMT) criterion [31, 86]. The correlation coefficients suggested low similarity between our trust factors and trust propensity. These results supported the discriminant validity of our trust 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. Based on the correlation coefficient, 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 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. The p-values indicated two faith items (.001) and one institution credibility item (.005) 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.

9 LIMITATION

This study is not without limitation. Future longitudinal research utilising our measurement instrument would be appropriate to examine the test-retest reliability for a longer time period. Replication with different types of AI systems, or possibly outside of healthcare application, would be beneficial to developing the trust model and also solidify the generality of our measurement instrument. 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. [83] was one of the reasons why Jian et al. [41] trust scale became the most well-cited trust scale in the human factors literature [32]. 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.

10 CONCLUSION

This study 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.

REFERENCES

[1] European Commission. 2020. 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.

[2] E Abdollahpour, S Nejat, M Nourozian, and R Majdzadeh. 2010. The process of content validity in instrument development. *Iranian Epidemiology* 6, 4 (2010), 66–74.

[3] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. 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, 582.

[4] Stavros Antifakos, Nicky Kern, Bernt Schiele, and Adrian Schwaninger. 2005. 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, 9–14.

[5] American Educational Research Association, American Psychological Association, National Council on Measurement in Education, et al. 1999. *Standards for educational and psychological testing*. American Educational Research Association.

[6] Christine Barter and Emma Renold. 1999. The use of vignettes in qualitative research. (1999).

[7] Paul C Beatty and Gordon B Willis. 2007. Research synthesis: The practice of cognitive interviewing. *Public opinion quarterly* 71, 2 (2007), 287–311.

[8] Michele Biasutti and Sara Frate. 2017. A validity and reliability study of the attitudes toward sustainable development scale. *Environmental Education Research* 23, 2 (2017), 214–230.

[9] US Defense Innovation Board. 2019. 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.

[10] Godfred O Boateng, Torsten B Neilands, Edward A Frongillo, Hugo R Melgar-Quiñonez, and Sera L Young. 2018. Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Frontiers in public health* 6 (2018), 149.

[11] Tom Bridgwater, Manuel Giuliani, Anouk van Maris, Greg Baker, Alan Winfield, and Tony Pipe. 2020. 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, 23–31.

[12] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*. IEEE, 160–169.

[13] Donald T Campbell and Donald W Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin* 56, 2 (1959), 81.

[14] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. 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*. 1–12.

[15] Shih-Yi Chien, Michael Lewis, Katia Sycara, Jyi-Shane Liu, and Asiye Kumru. 2018. The effect of culture on trust in automation: reliability and workload. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 4 (2018), 1–31.

[16] Domenic V Cicchetti and Sara A Sparrow. 1981. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *American journal of mental deficiency* (1981).

[17] Michael R Cohen and Judy L Smetzer. 2017. 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.

[18] National Governance Committee. 2021. 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/.

[19] Joanne E Croker, Dawn R Swancutt, Martin J Roberts, Gary A Abel, Martin Roland, and John L Campbell. 2013. Factors affecting patients' trust and confidence in GPs: evidence from the English national GP patient survey. *BMJ open* 3, 5 (2013), e002762.

[20] Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika* 16, 3 (1951), 297–334.

[21] Linda Lindsey Davis. 1992. Instrument review: Getting the most from a panel of experts. *Applied nursing research* 5, 4 (1992), 194–197.

[22] Robert F DeVellis and Carolyn T Thorpe. 2021. *Scale development: Theory and applications*. Sage publications.

[23] Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. 2003. The role of trust in automation reliance. *International journal of human-computer studies* 58, 6 (2003), 697–718.

[24] Shi Feng and Jordan Boyd-Graber. 2019. What can ai do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 229–239.

[25] Forbes. 2018. This Health Startup Won Big Government 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.

[26] Claes Fornell and David F Larcker. 1981. Evaluating structural equation models with unobservable variables and measurement error. *Journal of marketing research* 18, 1 (1981), 39–50.

[27] David Gefen. 2000. E-commerce: the role of familiarity and trust. *Omega* 28, 6 (2000), 725–737.

[28] David Gefen, Elena Karahanna, and Detmar W Straub. 2003. Trust and TAM in online shopping: an integrated model. *MIS quarterly* 27, 1 (2003), 51–90.

[29] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. 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.

[30] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2011. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association* 19, 1 (2011), 121–127.

[31] Andrew H Gold, Arvind Malhotra, and Albert H Segars. 2001. Knowledge management: An organizational capabilities perspective. *Journal of management information systems* 18, 1 (2001), 185–214.

[32] Robert S Gutzwiller, Erin K Chiou, Scotty D Craig, Christina M Lewis, Glenn J Lematta, and Chi-Ping Hsiung. 2019. 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, 217–221.

[33] Joseph F Hair Jr, Marcelo LDS Gabriel, Dirceu da Silva, and Sergio Braga Junior. 2019. Development and validation of attitudes measurement scales: fundamental and practical aspects. *RAUSP Management Journal* 54, 4 (2019), 490–507.

[34] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors* 53, 5 (2011), 517–527.

[35] Monika Hengstler, Ellen Enkel, and Selina Duelli. 2016. Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change* 105 (2016), 105–120.

[36] Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 241–250.

[37] Timothy R Hinkin. 1998. A brief tutorial on the development of measures for use in survey questionnaires. *Organizational research methods* 1, 1 (1998), 104–121.

[38] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* 57, 3 (2015), 407–434.

[39] Li-tze Hu and Peter M Bentler. 1999. 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.

[40] Intel. 2018. U.S. Healthcare Leaders Expect Widespread Adoption of Artificial Intelligence by 2023 | Intel Newsroom. (Accessed on 02/10/2019).

[41] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics* 4, 1 (2000), 53–71.

[42] Devon Johnson and Kent Grayson. 2005. Cognitive and affective trust in service relationships. *Journal of Business research* 58, 4 (2005), 500–507.

[43] Cynthia Johnson-George and Walter C Swap. 1982. 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.

[44] John D Lee and Neville Moray. 1994. Trust, self-confidence, and operators' adaptation to automation. *International journal of human-computer studies* 40, 1 (1994), 153–184.

[45] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[46] Stephan Lewandowsky, Michael Mundy, and Gerard Tan. 2000. The dynamics of trust: comparing humans to automation. *Journal of Experimental Psychology: Applied* 6, 2 (2000), 104.

[47] Mengyao Li, Brittany E Holthausen, Rachel E Stuck, and Bruce N Walker. 2019. 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*. 177–185.

[48] Xin Li, Traci J Hess, and Joseph S Valacich. 2008. 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.

[49] Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology* (1932).

[50] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.

[51] Scott B MacKenzie, Philip M Podsakoff, and Richard Fetter. 1991. 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.

[52] Maria Madsen and Shirley Gregor. 2000. Measuring human-computer trust. In *11th australasian conference on information systems*, Vol. 53. Citeseer, 6–8.

[53] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.

[54] D Harrison Mcknight, Michelle Carter, Jason Bennett Thatcher, and Paul F Clay. 2011. Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems (TMIS)* 2, 2 (2011), 12.

[55] D Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. 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.

[56] D Harrison McKnight, Larry L Cummings, and Norman L Chervany. 1998. Initial trust formation in new organizational relationships. *Academy of Management review* 23, 3 (1998), 473–490.

[57] Stephanie M Merritt. 2011. Affective processes in human–automation interactions. *Human Factors* 53, 4 (2011), 356–370.

[58] Joachim Meyer and John D Lee. 2013. Trust, reliance, and compliance. (2013).

[59] Neville Moray, Toshiyuki Inagaki, and Makoto Itoh. 2000. Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *Journal of experimental psychology: Applied* 6, 1 (2000), 44.

[60] JL Morrow Jr, Mark H Hansen, and Allison W Pearson. 2004. The cognitive and affective antecedents of general trust within cooperative organizations. *Journal of managerial issues* (2004), 48–64.

[61] Kathleen L Mosier and Linda J Skitka. 2018. Human decision makers and automated decision aids: Made for each other? In *Automation and human performance*. Routledge, 201–220.

[62] Bonnie M Muir. 1987. Trust between humans and machines, and the design of decision aids. *International journal of man-machine studies* 27, 5-6 (1987), 527–539.

[63] Bonnie Marlene Muir. 2002. Operators' trust in and use of automatic controllers in a supervisory process control task. (2002).

[64] Kevin R Murphy and Charles O Davidshofer. 1988. Psychological testing. *Principles, and Applications, Englewood Cliffs* 18 (1988).

[65] Jum C Nunnally. 1994. *Psychometric theory 3E*. Tata McGraw-hill education.

[66] Kristin E Oleson, Deborah R Billings, Vivien Kocsis, Jessie YC Chen, and Peter A Hancock. 2011. 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, 175–178.

[67] Richard M Perloff. 1993. *The dynamics of persuasion: Communication and attitudes in the 21st century*. Routledge.

[68] Denise F Polit and Cheryl Tatano Beck. 2006. 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.

[69] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810* (2018).

[70] PwC. 2016. Survey results: Why AI and robotics will define New Health: Publications: Healthcare: Industries: PwC. (Accessed on 02/10/2019).

[71] Tenko Raykov. 1997. Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate behavioral research* 32, 4 (1997), 329–353.

[72] Tenko Raykov and George A Marcoulides. 2011. *Introduction to psychometric theory*. Routledge.

[73] John K Rempel, John G Holmes, and Mark P Zanna. 1985. Trust in close relationships. *Journal of personality and social psychology* 49, 1 (1985), 95.

[74] Denise M Rousseau, Sim B Sitkin, Ronald S Burt, and Colin Camerer. 1998. Not so different after all: A cross-discipline view of trust. *Academy of management review* 23, 3 (1998), 393–404.

[75] Rosemary Rowe and Michael Calnan. 2006. Trust relations in health care—the new agenda. *The European Journal of Public Health* 16, 1 (2006), 4–6.

[76] Nicole Salomons, Michael Van Der Linden, Sarah Strokhorb Sebo, and Brian Scassellati. 2018. Humans conform to robots: Disambiguating trust, truth, and conformity. In *2018 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 187–195.

[77] Kristin Schaefer. 2013. The perception and measurement of human-robot trust. (2013).

[78] Kristin E Schaefer, Jessie YC Chen, James L Szalma, and Peter A Hancock. 2016. 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.

[79] Kenneth S Schultz and David J Whitney. 2005. Measurement theory in action. *Thousand Oaks* (2005).

[80] P Wesley Schultz. 2002. Environmental attitudes and behaviors across cultures. *Online readings in psychology and culture* 8, 1 (2002), 2307–0919.

[81] A Seif. 2004. Educational measurement, assessment and evaluation. *Tehran: Doran Publications* 128 (2004).

[82] Rashmi Sinha and Kirsten Swearingen. 2002. The role of transparency in recommender systems. In *CHI'02 extended abstracts on Human factors in computing systems*. ACM, 830–831.

[83] Randall D Spain, Ernesto A Bustamante, and James P Bliss. 2008. 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, 1335–1339.

[84] Frank MF Verberne, Jaap Ham, and Cees JH Midden. 2012. 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.

[85] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. 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.

[86] Clay M Voorhees, Michael K Brady, Roger Calantone, and Edward Ramirez. 2016. 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.

[87] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GPDR. *Harv. JL & Tech.* 31 (2017), 841.

[88] Gordon B Willis. 2004. *Cognitive interviewing: A tool for improving questionnaire design*. sage publications.

[89] Kevin B Wright. 2005. 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.

[90] Yaqi Xie, Indu P Bodala, Desmond C Ong, David Hsu, and Harold Soh. 2019. Robot capability and intention in trust-based decisions across tasks. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 39–47.

[91] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[92] Bowen Yu, Ye Yuan, Loren Terveen, Zhiwei Steven Wu, Jodi Forlizzi, and Haiyi Zhu. 2020. Keeping designers in the loop: Communicating inherent algorithmic trade-offs across multiple objectives. In *Proceedings of the 2020 ACM designing interactive systems conference*. 1245–1257.

[93] Beste F Yuksel, Penny Collisson, and Mary Czerwinski. 2017. Brains or beauty: How to engender trust in user-agent interactions. *ACM Transactions on Internet Technology (TOIT)* 17, 1 (2017), 1–20.

[94] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. 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*. 295–305.