Open access • Book Chapter • DOI:10.1007/978-3-319-96074-6_2

Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation — Source link

Moritz Körber

Institutions: Technische Universität München

Published on: 26 Aug 2018

Topics: Automation

Preprint of Körber, M. (in press). Theoretical considerations and development of a questionnaire to measure trust in automation. In *Proceedings 20th Triennial Congress of the IEA.* Springer.

An English or German version of the questionnaire is available at the end of the manuscript or at https://github.com/moritzkoerber/TiA_Trust_in_Automation_Questionnaire

Theoretical considerations and development of a questionnaire to measure trust in automation

Moritz Körber

Chair of Ergonomics, Technical University of Munich
Boltzmannstraße 15, 85747 Garching
Tel: +49 89 289 15376
moritz.koerber@tum.de

Abstract

The increasing number of interactions with automated systems has sparked the interest of researchers in trust in automation because it predicts not only whether but also how an operator interacts with an automation. In this work, a theoretical model of trust in automation is established and the development and evaluation of a corresponding questionnaire (*Trust in Automation*, TiA) are described.

Building on the model of organizational trust by Mayer, Davis, and Schoorman (1995) and the theoretical account by Lee and See (2004), a model for trust in automation containing six underlying dimensions was established. Following a deductive approach, an initial set of 57 items was generated. In a first online study, these items were analyzed and based on the criteria item difficulty, standard deviation, item-total correlation, internal consistency, overlap with other items in content, and response quote, 40 items were eliminated and two scales were merged, leaving six scales (*Reliability/Competence*, *Understandability/Predictability*, *Propensity to Trust, Intention of Developers*, *Familiarity*, and *Trust in Automation*) containing a total of 19 items.

The internal structure of the resulting questionnaire was analyzed in a subsequent second online study by means of an exploratory factor analysis. The results show sufficient preliminary evidence for the proposed factor structure and demonstrate that further pursuit of the model is reasonable but certain revisions may be necessary. The calculated omega coefficients indicated good to excellent reliability for all scales. The results also provide evidence for the questionnaire's criterion validity: Consistent with the expectations, an unreliable automated driving system received lower trust ratings as a reliably functioning system. In a subsequent empirical driving simulator study, trust ratings could predict reliance on an automated driving system and monitoring in form of gaze behavior. Possible steps for revisions are discussed and recommendations for the application of the questionnaire are given.

1. Introduction

It has become impossible to evade automation: Thanks to the technological progress made, many functions that were previously carried out by humans can now be fully or partially replaced by machines (Parasuraman, Sheridan, & Wickens, 2000). As a consequence, they are taking over more and more functions in work and leisure environments of all kinds in our day-to-day lives. The resulting increase in the number of interactions with automated systems has sparked the interest of human factors researchers to investigate trust in automation with the overall goal to ensure safe and efficient joint system performance in mind (Drnec, Marathe, Lukos, & Metcalfe, 2016). An empirical investigation of trust in automation necessitates a measurement of trust in automation. Trust in automation is a latent construct, which is not directly observable; thereby, researchers rely on indicators such as neuroscientific methods (Drnec et al., 2016), behavioral measures (e.g., eye tracking; Hergeth, Lorenz, Vilimek, & Krems, 2016), or questionnaires (Jian, Bisantz, & Drury, 2000; Madsen & Gregor, 2000).

Trust in automation and reliance on automation are closely related: "People tend to rely on automation they trust and tend to reject automation they do not" (Lee & See, 2004, p. 51). Yet, trust in automation and reliance on automation are at the same time distinct constructs. In their theory of reasoned action, Ajzen and Fishbein (1980) argue that behavior, such as reliance, results from an intention and that this intention is a function of attitudes, which in turn are an affective evaluation of beliefs. Trust in automation as an attitude, thus, stands between the belief about the characteristics of an automated system, such as its reliability, and the intention to rely on it. Attitude, intention, and actual behavior are not in a deterministic but in a probabilistic relationship (Ajzen & Fishbein, 1980). Whether trust translates into reliance behavior depends on a dynamic interaction of operator, automation, situational factors, and interface (Lee & See, 2004). As a result, other factors, such as the effort to engage or self-confidence, also affect the intention to rely on an automated system (Bisantz & Seong, 2001; Dzindolet, Beck, Pierce, & Dawe, 2001; Kirlik, 1993; Lee & See, 2004; Meyer, 2004). Environmental and cognitive constraints, such as time pressure, then determine whether a formed intention translates into actual reliance on automation. Even if trust is at a high level and the automated system is perceived as capable, reliance does not necessarily follow (Kirlik, 1993). That means, to measure trust as an attitude itself, a questionnaire or another similar methodology that is distinct from observable risk taking is necessary (Mayer et al., 1995). Furthermore, the conceptualizations of trust in automation refer to the construct as an attitude (Lee & See, 2004), a mainly affective response closely related to beliefs and expectations. Affective responses are not always accompanied by overt behavior. For example, students with and without math anxiety may behave the same way during a math test even though their internal state differs (McCoach, Gable, & Madura, 2013). An affective response is, thereby, probably only completely accessible through self-report (Paulhus & Vazire, 2007). A questionnaire, therefore, is an attractive method to measure trust in automation.

2. Theoretical Model

A literature review of available questionnaires on trust in automation revealed that the questionnaires comprise single-item as well as multi-item scales. Single-item scales allow a quick, uncomplicated measurement such as a dynamic assessment during an experiment. However, these instruments also have some drawbacks. Dimensions and models of trust have been extensively discussed, resulting in a variety of facets and concepts (Lee & See, 2004). It is questionable whether the broadness and depth of this construct can be captured by a single questionnaire item. In contrast, multiple, heterogeneous indicators (= questionnaire items) enhance construct validity by increasing the probability of adequately identifying the construct (Eisinga, Grotenhuis, & Pelzer, 2013). Consequently, Fuchs and Diamantopoulos (2009) do not recommend single-item scales if the construct in question is abstract. Likewise, a single item does not allow for a detailed analysis of the underlying reasons for a favorable or non-favorable trust score. Is the machine perceived as unreliable? Does a participant simply not trust a certain brand? It is not possible to give an answer with a single item scale. Using multiple items also helps to cancel out errors due to specificities inherent in single items, which lowers measurement error and, thereby, increases reliability (Diamantopoulos, Sarstedt, Fuchs, Wilczynski, & Kaiser, 2012;

Moosbrugger & Kelava, 2012; Robins, Hendin, & Trzesniewski, 2001). Single-item scales are correspondingly more susceptible to unknown biases in meaning and interpretation (Hoeppner, Kelly, Urbanoski, & Slaymaker, 2011). For a detailed discussion on the choice between single-item and multi-item scales, readers may consider Körber (2018). Because of these drawbacks, it was decided to develop a multi-item questionnaire.

The measurement of a latent construct such as trust requires the process of construct validation (Flake, Pek, & Hehman, 2017). In the substantive phase, the literature is reviewed, the construct is defined and conceptualized, and its dimensions, boundaries, and structure are identified. For this purpose, theoretical discourses on trust in automation were screened along with empirical articles and articles with a stronger focus on interpersonal trust[1]. The most widespread and most cited model of trust is the dyadic model of organizational trust by Mayer et al. (1995). Integrating previous theoretical accounts on trust, the parsimonious model differentiates trust from its contributing factors and its outcome, risk taking in a relationship. The authors argue that trust is only necessary in a risky situation or when having something invested. In this context, they define trust as

> the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party. (Mayer et al., 1995, p. 712)

According to their model, a person's trust depends on two components, a person's *individual propensity* or general willingness to trust others and the *trustworthiness* of the party to be trusted (trustee). A person's trust propensity results from different developmental experiences, personality type, and cultural background and determines how much a person trusts a trustee prior to any knowledge of that particular party being available. The second component, the perceived trustworthiness, is determined by three relevant attributes of the trustee: 1) *Ability*: The level of skills, competencies, and characteristics that the trustee possesses and that enables him to have influence within a specific domain. 2) *Benevolence*: The extent to which a trustee is perceived to want to do good to the trustor and avoids egocentric motives. 3) *Integrity*: The extent to which the trustee consistently adheres to a set of principles that the trustor finds acceptable. Risk taking is then the behavioral manifestation of the willingness to be vulnerable, i.e. the outcome of trust.

Since interpersonal trust and trust in automation exhibit fundamental differences (Körber, 2018), the model from Mayer et al. (1995) does not completely apply to trust in automation. Taking this into account, Lee and See (2004) follow the model of trust by Mayer et al. (1995) but fit their dimensions to the context of trust in automation. They argue that previously found bases for trust in automation can be summarized into three dimensions, *performance*, *process*, and *purpose*, which correspond to the dimensions of trustworthiness in the model by Mayer et al. (1995), as illustrated in Figure 1. *Performance* refers to the current and previous operation of the automated system and comprises characteristics such as reliability, competency, and ability. Performance information describes what the automated system can do reliably and matches the attribute ability in Mayer et al. (1995). *Process* describes how the automated system operates and if this modus operandi is appropriate for the

[1] In this literature review, the following work was considered: Barber (1983), Blomqvist (1997), Butler and Cantrell (1984), Butler (1991), Deutsch (1958), Deutsch (1960), Dzindolet et al. (2001), Hoff and Bashir (2015), Hoffman, Johnson, Bradshaw, and Underbrink (2013), Jian et al. (2000), Lee and Moray (1992), Lee and See (2004), Madhavan and Wiegmann (2007), Madsen and Gregor (2000), Mayer et al. (1995), McKnight and Chervany (1996), McKnight and Chervany (2001), Muir (1987), Muir (1994), Muir and Moray (1996), Rempel, Holmes, and Zanna (1985), Rotter (1971).

situation and the operator's goals. It subsumes characteristics such as understandability and matches integrity in Mayer et al. (1995). *Purpose* describes the intention in the automated system's design, the perception that the designers possess a positive orientation towards the operator, and the degree to which automation is used as intended by the designer. It corresponds to benevolence in Mayer et al. (1995). We follow the model from Lee and See (2004) but divide the three components into more detailed facets for item generation. Three underlying dimensions of trust in automation were postulated: *Reliability/Competence*, *Understandability/Predictability*, and *Intention of Developers*. Trust exhibits a stable individual component (Körber, 2018). Individuals consistently vary in their general propensity to trust, depending on their developmental experiences, personality type, and cultural backgrounds. Additionally, not objective characteristics but a person's subjective perception of a system's characteristics determines trust in automation in the end (Lee & See, 2004; Merritt & Ilgen, 2008). We, therefore, added the individual component, *Propensity to Trust*, from the model of Mayer et al. (1995) as a moderator but also as a direct determinant of trust in automation.

The model of Mayer et al. (1995) addresses interpersonal trust. While other human individuals may be perceived more or less as individuals, different driving automation systems seem to be perceived as a single technology (Schoettle & Sivak, 2014). This increases the importance of prior familiarity because trust is thereby probably not evaluated again for each driving automation system. Familiarity is assumed to have an indirect influence on trust in automation. With increasing familiarity, operators form expectations, calibrate their trust, and eventually, their confidence in the evaluation of the attributes increases (Hergeth, Lorenz, & Krems, 2017). For example, if no unexpected failures occur, the confidence in the system's reliability increases. As experience with a system grows, trust builds up until a certain level is reached (Beggiato, Pereira, Petzoldt, & Krems, 2015). Taking this into account, *Familiarity* with an automated system was included as a moderator in the theoretical model. Figure 1 illustrates the complete model structure. Based on Mayer et al. (1995), we define trust in automation as

> the attitude of a user to be willing to be vulnerable to the actions of an automated system based on the expectation that it will perform a particular action important to the user, irrespective of the ability to monitor or to intervene.

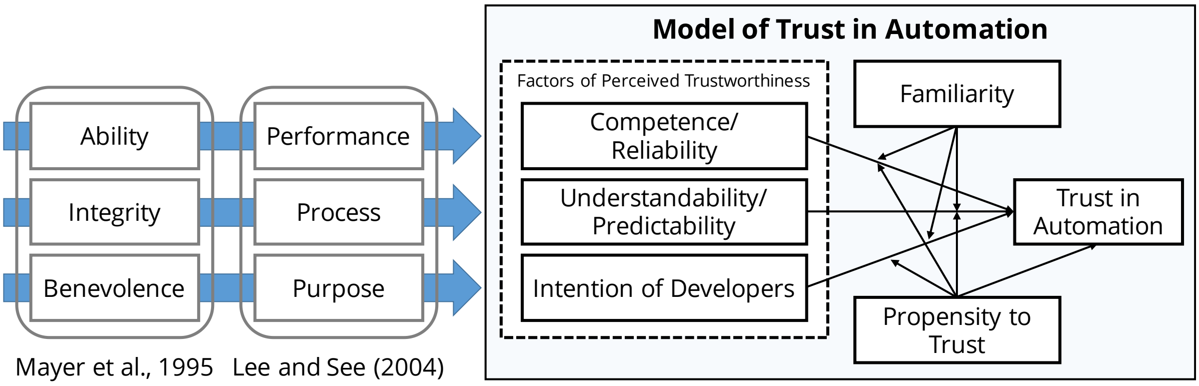

Figure 1. Model of trust in automation based on the postulated dimensions by Mayer et al. (1995) and Lee and See (2004).

3. Item generation and analysis

Likert-scales are used as means of measurement. Measurement by Likert-scales is based on summative scaling, where respondents use a ranked scale to indicate their agreement with statements. The goal is to combine the single item responses of an individual to obtain a total score that represents a reliable measurement – multiple Likert-type items form one coherent Likert scale (Hubley & Zumbo, 2013; Uebersax, 2006). A 5-point rating scale ranging between 1 (= strongly disagree) to 5 (= strongly agree) was chosen as the response format. Rating scales with a very fine-grained range, for example from 1 to 100 as in Brown and Galster (2004), offer a resolution that might be inadequate for the provided precision of the measurement, resulting in merely artificial precision. Furthermore, the self-report of trust is based on introspection. It is questionable whether the participants are able to access their trust by introspection with such a granularity as provided by the scale. Such a fine-grained scale might map an empirical structure, which does not exist in this resolution, onto numbers with limited meaning. If such scales provide no anchor points, measurement at interval scale level is also even more problematic since equidistance between the rating scale points becomes even more questionable.

We followed a deductive approach for the generation of items (Burisch, 1978) and constructed the questionnaire based on classical test theory (Moosbrugger & Kelava, 2012). An initial set of 57 items was generated. Approximately one third of the items was inversely formulated to reduce response bias (e.g., acquiescence bias) and based on Likert's notion that someone with a positive attitude about the object should also disagree with negative statements. An online survey was conducted for item analysis. In this survey, the participants watched two videos of an automated driving system (a Level 3 driving automation system; ADS). The first video gave a circa 10-minute visual and verbal explanation of the underlying technology of automated vehicles and their functionality. The second video showed an approximately 3-minute highway drive in a conditionally automated vehicle. A total of n = 94 participants completed the survey, 32 participants were female (34.00 %), 60 were male (63.80 %). The mean age was M = 35.60 years (SD = 14.60, ranging from 17 to 71 years). Based on the criteria item difficulty, standard deviation, item-total correlation, internal consistency, overlap with other items in content, and response quote, 32 items were eliminated, leaving 25 items.

The first validation was carried out in a subsequent online study. In a between-subjects design, a sample of n = 58 participants (age range 17 to 72, mean age M = 34.00 years, SD = 15.10, 58.60 % male, 37.90 % female) watched a video of a conditionally automated highway drive. Participants were randomly assigned to a *reliable* condition, where the video showed a perfectly functioning automation, or a *non-reliable* condition, where participants watched an extended version including a take-over request. As expected, participants of the reliable condition rated the ADS more reliable (t(41.32) = 3.76, p < .001, d = 1.05). Additionally, participants rated their trust directly by answering the item "I trust this system" on a 5-point rating scale ranging between 1 (= strongly disagree) to 5 (= strongly agree). All scales correlated positively with different strength with this rating (lowest: *Familiarity*: r = .33; highest: *Reliability*: r = .85). Although the total questionnaire correlated strongly with this item (r = .81), we found no significant difference between the two conditions (t(46.92) = 1.21, p = .23, d = 0.33), on the contrary for the direct question (t(45.63) = 2.58, p = .01, d = 0.71). Because of their high correlation, the scales competence and reliability were merged, leading to a reduction to 17 items. The internal consistency of the scales ranged from acceptable (α = .75; *Propensity to trust*) to excellent (α = .92; *Reliability/Competence*).

McCoach et al. (2013) recommend utilizing an exploratory factor analysis (EFA) to evaluate the structure in the very first pilot study because it allows for the highest flexibility of potential solutions. An exploratory factor analysis was conducted to assess whether the structure of the covariation among items is consistent with the proposed factor structure of the trust model. The analysis was performed in JASP (Love et al., 2015). The dataset showed a sufficient basis to conduct an initial exploratory factor analysis (KMO = .80, Bartlett-Test: χ^2(136) = 418.81, p < .001). Following the recommendations of Sakaluk and Short (2017) and McCoach et al. (2013), we chose principal axis factoring as the extraction method and oblique rotation (oblimin) to make the factor solution more interpretable. Parallel analysis by Horn (1965) as well as multiple item factor loadings > .40 on only one single factor determined the extracted factors (Figure 2). Results of the analysis provide initial support for the assumed factorial structure. The resulting pattern matrix (Table 1) shows a clear structure of four factors with high over-determination, "the degree to which each factor is clearly represented by a sufficient number of variables" (MacCallum, Widaman, Zhang, & Hong, 1999, p. 89). Each factor exhibits high pattern coefficients (> .50) by multiple variables while each of the items does not load substantially (> .35) onto other factors, a requirement for a stable solution. Medium to high communalities were observed. Table 2 and Table 3 provide further information on the resulting solution.

Table 1
Pattern matrix generated by principal axis factoring; loadings < .35 have been omitted

	Factor 1	Factor 2	Factor 3	Factor 4	Uniqueness
Familiarity 1			.81		.31
Familiarity 2			.80		.34
Intention of Developers 1		.74			.46
Intention of Developers 2		.49			.45
Propensity to Trust 1				.58	.60
Propensity to Trust 2				.55	.36
Propensity to Trust 3				.59	.55
Reliability/Competence 1	.88				.15
Reliability/Competence 2	.70				.34
Reliability/Competence 3	.79				.23
Reliability/Competence 4	.82				.30
Reliability/Competence 5	.86				.28
Reliability/Competence 6	.70				.44
Understanding/Predictability 1		.65			.36
Understanding/Predictability 2		.60			.44
Understanding/Predictability 3	.64				.24
Understanding/Predictability 4		.62			.50

Table 2
Inter-correlations matrix of the extracted factors

	Factor 1	Factor 2	Factor 3	Factor 4
Factor 1				
Factor 2	.65			
Factor 3	.25	.24		
Factor 4	.19	.31	.04	

Table 3
Fit indices of the resulting model

Chi-squared test			Additional fit indices	
Value	*df*	*p*	RMSEA	TLI
112.139	74	.003	0.09 [0.05, 0.11]	.85

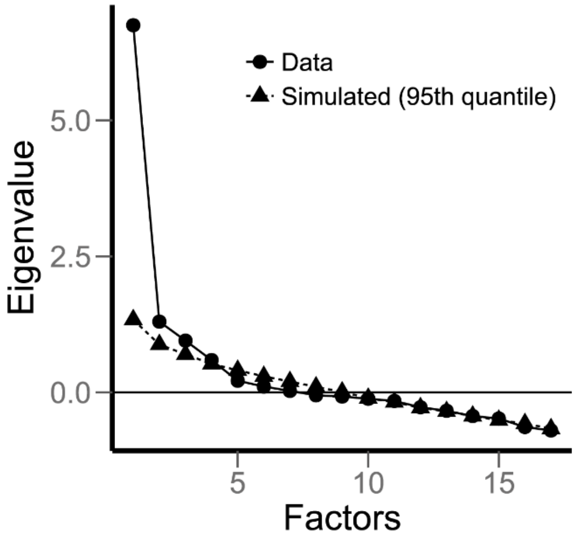

Figure 2. Scree plot of the extracted factors with a parallel analysis by Horn (1965) superimposed.

McNeish (2017) advises against using Cronbach's alpha as a reliability index because its rigid assumptions are routinely violated. He suggests using the omega coefficient, which is conceptually similar to Cronbach's alpha but makes less strict assumptions. In fact, omega total is a more general version of Cronbach's alpha: It also assumes unidimensionality, but the items are allowed to vary in how strongly they are related to the measured construct. Revelle's omega differs from omega total in its more sophisticated variance decomposition. Given that the items each implement a 5-point rating scale, relying on Pearson covariance matrices is reasonable (Rhemtulla, Brosseau-Liard, & Savalei, 2012). All scales exhibited good to excellent internal consistency (Table 4).

Table 4
Indices of the internal consistency of each scale; ^{a} since Omega total and Revelle's omega cannot be calculated for scales with fewer than three items, the Spearman-Brown coefficient according to Eisinga et al. (2013) was calculated

	Omega Total	Revelle's Omega
Familiarity	.83^a	-
Intention of Developers	.79^a	-
Propensity to Trust	.78	.77
Reliability/Competence	.92	.95
Understanding	.81	.88

The factor Reliability/Competence was the first extracted factor and, therefore, explained a very major part of the variance, which may be expected given the design of the study, i.e. automation reliability was manipulated between the conditions. However, no factor for *Intention of the Developers* could be extracted. The reason for this may lie in the domain of automated driving. A driving automation system is an expensive, highly sophisticated system whose development was motivated by the increase in safety and comfort. The developers of the system are known to be professional car manufacturers. Thus, it is hard to imagine that a driving automation system's developers did not act in a benevolent manner. A revised version of the questionnaire may eliminate this dimension, at least in the domain of automated driving. Item 3 of Understanding ("The system state was always clear to me") seems to exhibit a certain degree of multidimensionality and may also be eliminated if this again is the case in future analyses.

Although the aim was to conduct an EFA, fit indices for the model, known from confirmatory factor analysis (CFA), are also reported (Table 3). Fit indices indicate how well the empirical data of the study actually conform to the proposed model. A CFA, therefore, is a more stringent test if the pattern of relationships among the items can be explained by the proposed model/factor structure (McCoach et al., 2013). The chi-squared test evaluates the null hypothesis that the proposed model exactly reproduces the population covariance matrix implied by the data (McCoach et al., 2013). This null hypothesis has to be rejected for the four-factor model. Besides the chi-squared test is generally too liberal at small samples sizes, as in this study, the informative value of this rejection is limited by the fact that a model is always a simplification of a process in reality that never intends to exactly recreate it (McCoach, 2003). The root mean square error of approximation (RMSEA) is an index of absolute fit that compensates for the effect of model complexity (Hu & Bentler, 1999) and can be considered an estimate of the misfit of the model per degree of freedom in the population (Preacher, Zhang, Kim, & Mels, 2013). Cut-offs for small sample sizes ($N \leq 250$) are .08 for a mediocre fit whereas .10 and larger indicates a poor fit (Heene, Hilbert, Draxler, Ziegler, & Bühner, 2011; Hu & Bentler, 1999; MacCallum, Browne, & Sugawara, 1996), indicating a mediocre fit for the four-factor structure of the trust model. However, the estimate is positively biased and the amount of the bias depends on the smallness of the sample and the degrees of freedom (Kenny, Kaniskan, & McCoach, 2014). The Tucker-Lewis Index (TLI) indicates an incremental fit and also compensates for model complexity. A TLI value at or above .95 indicates a good fit, TLI values below .90 are generally considered less than satisfactory (McCoach, 2003). The four-factor model does not fulfill this criterion. However, once again, the TLI is biased in small samples, i.e. it is underestimated in samples with fewer than 100 participants. Heene et al. (2011) echoe previous critique on the application of fixed cut-off rules for model fit because of the multiple dependencies of the fit indices on the conditions (e.g., the achieved factor loadings) and on sample size. After establishing the trust model, two items for measurement of trust in automation itself ("I trust the system" and "I can rely on the system") forming the subscale *Trust in Automation* were added.

The EFA gathered sufficient preliminary evidence of the factor structure and shows that further pursuit of the model is reasonable. Nevertheless, this analysis of construct validity is certainly not sufficient. Firstly, the sample size of $n = 58$ participants results in a case/item ratio of approximately 3:1, which reflects the absolute minimum for a sensible analysis and may be too small to produce a stable solution. However, the minimum required ratio is not constant across studies but rather depends on aspects of the variables and study design (MacCallum et al., 1999). Given a clear factor structure, a high degree of over-determination and high communalities (constantly > .60, as in this study), it is nevertheless possible to reach a stable factor solution even with a sample size smaller than 100 participants (McCoach et al., 2013). Secondly, the participants did not experience driving automation themselves but watched videos of it. Thirdly, the participants only got a short, probably first impression of a driving automation system. This may promote a single-factor structure because the participants might not have had enough insight into the driving automation system to form themselves a detailed, multifaceted impression.

The results of the initial exploratory factor analysis established sufficient initial evidence for the factor structure, affirming that further work is sensible but also needed. Thus, the development process for the questionnaire has certainly not yet come to its end. Future studies have to investigate and ensure the construct validity in greater detail and need to investigate the structure in an applied setting with an adequate sample size. Future work should also follow up this analysis with a CFA to put the established structure to a more rigorous test. In a structural equation model, the claimed paths and relationships of the model can be directly tested and different models can be compared. The questionnaire's criterion validity was examined in its first use in a driving simulator study in Körber, Baseler, and Bengler (2018).

4. Predictive validity

In the study by Körber et al. (2018), the developed questionnaire to measure trust in automation was used in an applied setting for the first time. In this driving simulator study, 40 participants encountered three critical situations while driving in a conditionally automated vehicle (SAE Level 3) on a highway while being engaged in a non-driving-related task. Eye tracking was used to assess how much the participants rely on driving automation. Furthermore, the instruction for the ADS was varied between two groups with participants receiving either trust-promoting (*Trust promoted* group) or trust-lowering (*Trust lowered* group) introductory information. The trust questionnaire was administered three times: 1) after an introductory video, 2) after an introductory drive, 3) after the experimental drive. It was expected that, firstly, self-reported trust will correlate positively with reliance on automation and, secondly, that participants of the *Trust promoted* group will report higher trust than the *Trust lowered* group.

The analysis comprised the whole *Trust in Automation Questionnaire* (TiA; 19 items) as well as just the subscale *Trust in Automation* and the subscale *Competence.* Regarding the reliability of the *Trust in Automation* subscale, the drawbacks of short scales become eminent. The scale exhibits a low reliability of $\alpha = .63$ after the video and of $\alpha = .70$ after the introductory drive, while it achieved a high reliability of $\alpha = .85$ after the experimental drive. This reflects the problems mentioned earlier with single-item scales: They are more vulnerable to random measurement errors and more susceptible to unknown biases in meaning and interpretation (Emons, Sijtsma, & Meijer, 2007; Hoeppner et al., 2011). Nevertheless, the subscale *Trust in Automation* was the scale that showed the largest difference ($M_{diff} = 0.45$, $d = 0.59$, $BF_{-0} = 4.35$) between the two groups after the introductory drive. The subscale might be more sensitive than the whole questionnaire, but this does not guarantee that its predictive performance regarding trust in other systems is superior − predictive quality might vary in different situations and context. The experiment included two situations (Situation 1: overtaking maneuver; Situation 2: adapting speed to a headway vehicle) that were solved by the automated vehicle, but a take-over was a reasonable action if one does not trust automation. In both situations, participants who intervened showed lower trust than participants who did not intervene. The effect size was comparable between the full TiA questionnaire (Situation 1: $d = 0.41$, Situation 2: $d = 0.51$) and the subscale *Trust in Automation* (Situation 1: $d = 0.50$, Situation 2: $d = 0.45$). The same results were obtained for the take-over situation, where participants who crashed reported higher trust than collision-free participants (Full TiA: $d = 0.51$; subscale *Trust in Automation*: $d = 0.58$). Both scales correlated moderately with take-over time (Full TiA: $r = .27$; subscale *Trust in Automation*: $r = .33$) and minimum time-to-collision (Full TiA: $r = -.29$; subscale *Trust in Automation*: $r = -.35$). Both full questionnaire and subscale *Trust in Automation* correlated with the participants' gaze behavior with the expected sign and at approximately the same magnitude (medium effect) in all three measurement intervals.

In summary, participants with higher trust scores consistently showed stronger reliance in all behavioral measurements compared to participants with a lower trust score. Consequently, the study confirms the predictive validity of the questionnaire. Furthermore, the medium-sized correlation between the TiA questionnaire score and the affinity for technology questionnaire (Feuerberg, Bahner,

& Manzey, 2005) of *r* = .47 (BF = 18.85) shows that trust is related to affinity for technology, yet it represents a distinct construct, supporting its construct validity.

5. Is a single item enough to measure trust in automation?

The two-item subscale *Trust in Automation* showed lower reliability but was more sensitive regarding group differences and performed equally as well as the full TiA questionnaire regarding all other measures. This provokes the question of whether a single-item scale may be sufficient for a valid measurement of trust. The benefits of using single-item measures have been listed by several researchers (Fuchs & Diamantopoulos, 2009; Hoeppner et al., 2011): Single-item scales are less monotonous and time-consuming. They can also be administered during an experiment for a momentary assessment, for example while driving. The aforementioned advantages of multi-item scales are also accompanied by drawbacks, such as boredom caused by redundant items and fatigue in lengthy questionnaires (Burisch, 1984). Nevertheless, for a detailed assessment of a multidimensional construct such as trust in automation, a multi-item measure is typically necessary (Nunnally & Bernstein, 1994).

Yet, Fuchs and Diamantopoulos (2009) argue that the use of a single-item scale may still be appropriate in certain cases. For example, Sloan, Aaronson, Cappelleri, Fairclough, and Varricchio (2002), while discussing the quality of life measurement, claim that “there comes a point where the construct becomes so complex that a single question may be the best approach” (p. 481). Hence, when measuring overall job satisfaction, the best measurement may be a question like “Overall, how satisfied are you with your job?” (Fuchs & Diamantopoulos, 2009, p. 204; Scarpello & Campbell, 1983). A single item on trust in automation reflects the conceptualization of trust in automation as a mainly affective response with influences from analytic and analogical processes. Lee and See (2004) suggest that because of the complexity of automation technology, operators probably rely less on analytic calculations to guide their behavior but rather apply heuristics to accommodate the limits of the human bounded rationality (Gigerenzer & Selten, 2002). A situation might occur where operators cannot form a complete mental model of an automated system as it is too complex to perfectly predict its behavior. Emotions can then guide behavior when rules are not effective or when cognitive resources are too limited for a calculated rational choice (Damasio, 1996; Lee & See, 2004). In the validation study, 78 % of the participants have had no contact with conditionally automated driving before. Thus, it might not have been possible for the participants to rate each dimension of the trust questionnaire adequately because of a lack of knowledge or experience. Differences in the ability to accurately rate a system have been pointed out by Annett (2002) who gives the example of expert test drivers who learn by experience to identify and rate the subtle dynamic features of a vehicle. It is conceivable that the participants’ trust rating was a rather global impression or rating, which can be captured accurately by a single item. It is unclear if participants would also provide a global rating if they had more experience with an automated vehicle.

Yet, such a simplification of the construct trust in automation comes with a cost: The *Trust in Automation* scale consists of two items, one of them with the content “I can rely on the system”. It is not surprising that such a measure highly correlates with behavioral reliance measures such as eye tracking and intervention frequency. For such a narrow conceptualization of trust, the high validity may justify the use of a single-item measure (Flake et al., 2017). The construct trust in automation, which possesses a detailed underlying theory (Lee & See, 2004; Mayer et al., 1995) would then, at the same time, become one with its measure and loses any theoretical meaning beyond that measure (Bagozzi, 1982). This measurement would then be in conflict with the definition of what it intends to measure. Indeed, as already mentioned, trust is an attitude that stands between the belief about characteristics of an automated system and the intention to rely on it. Attitude, intention, and actual behavior are not in a deterministic but in a probabilistic relationship (Ajzen & Fishbein, 1980). Whether trust translates into actual reliance on an automated system is also influenced by other factors such as self-confidence or time constraints (Dzindolet et al., 2001; Lee & See, 2004; Meyer, 2004).

Besides the psychometric drawbacks mentioned in Körber (2018), the use of a single-item measure is also problematic in longitudinal studies: If the observed value changes, it is not possible to differentiate between a true change in the construct and a change caused by imperfect reliability of the measurement (Fuchs & Diamantopoulos, 2009). Here, researchers may fall back on the multi-item questionnaire. If a single-item is administered to obtain a global assessment, it has to be taken into account that the respondents each consider an individual set of aspects of trust and of the automated system, weighted by their own individual preferences, providing a tailor-made impression (Nagy, 2002). Hence, respondents may not consider the same aspects or may not even think of a relevant aspect at all. It, thereby, remains unknown how the assessment is constituted. To ensure that each participant assesses the same construct, i.e. that a common understanding of trust exists, an accurate definition of trust in automation has to be provided in this case (Fuchs & Diamantopoulos, 2009). On the other hand, multi-item scales are less individual but more comparable. A preset of aspects, formed by the questionnaire's scales, also helps and guides the participants to rate the system.

Multiple scales also provide the possibility to express the trust rating in greater detail. With a single-item scale, should the trust score turn out to be low, the researchers then have no indication for the reason. Contrarily, multiple scales may enable researchers to find the cause in a certain characteristic of the automated system. For example, it could be perceived as reliable, but participants did not understand its functioning. Thus, it is reasonable to use a multiple-item scale such as the TiA if the aim is a thorough, multi-faceted assessment.

In conclusion, if the research objective is a global assessment, an overall feeling, or impression by the participants, then a single-item may provide all the desired information (Fuchs & Diamantopoulos, 2009). It represents a useful supplement that might be sufficient for a single and quick, yet valid assessment and "can provide an acceptable balance between practical needs and psychometric concerns" (Robins et al., 2001, p. 152). This is particularly true if trust is merely used as a moderator or as a control variable (Fuchs & Diamantopoulos, 2009). If the goal is a detailed assessment of trust in automation or if a longitudinal design is implemented, then the multi-item questionnaire may be preferred.

6. Objectives for revision and further development

The questionnaire's further development certainly needs to address its psychometric qualities. The low internal consistency of the subscale *Trust in Automation* at the beginning of the study raises the question of whether a short scale of two rather direct items is sufficient as a measurement of trust itself. Mayer and Davis (1999) provide a questionnaire for their model of interpersonal trust, which includes a four-item scale to assess trust. The items are less direct than the two trust items of the TiA questionnaire and rather aim at the willingness to be vulnerable, corresponding to their definition of trust (Mayer et al., 1995). Thus, a revised version of the TiA questionnaire may adopt this approach and offer a four-item scale (besides the original scales) for trust in automation that is closer to its definition by Körber et al. (2018). Items from Mayer and Davis (1999) adapted to the domain of automation could, for example, read "I would be comfortable handing over the driving task to the driving automation system without monitoring it" or "If I had my way, I wouldn't let a driving automation system have any influence on the driving task". A single item for assessing trust in automation such as "I trust this driving automation system" then may function as the aforementioned pragmatic variant alongside the multi-item questionnaire. In addition, information on the questionnaire's discriminant validity is still missing. Also, further data on the questionnaire's predictive performance have to be gathered.

A revision may also reconsider the inclusion of the scale *Familiarity.* Familiarity itself is not an element of trust in automation but indirectly influences it as a moderator. With increasing familiarity, operators form expectations and the confidence in their evaluation of the attributes increases. If this moderating role is of no interest in a study, the scale could be eliminated to shorten the questionnaire. A core questionnaire only containing the factors that directly influence trust then may be more appropriate. Beyond this, familiarity could also induce response bias: Low familiarity with an automated system could induce a tendency towards a global evaluation of the system due to a lack of in-depth knowledge. It would, therefore, be interesting to administer the questionnaire to participants who are already very familiar with a driving automation system. This is especially of interest regarding the difference between the predictive performance of a single-item measure and the multi-item TiA questionnaire.

In closing, it has to be considered that the measurement of trust in automation by means of a questionnaire certainly has to be viewed in perspective of its position in measurement theory. There have been concerns doubting the possibility of measurement of psychological constructs and their quantitative nature in general (Michell, 1997). However, using rating scales for the measurement of psychological constructs, such as trust, does not exclusively have to be regarded as a form of measurement in the strict sense of the term, i.e. in terms of the representational theory of measurement, where a homomorphic representation of physical empirical relations is mapped to numerical relations (Annett, 2002; Krantz, Luce, Suppes, & Tversky, 2007). Instead, following a model-based account of measurement, measurement of trust can rely on an abstract model that is valid for the prediction of an individual's performance during a certain task (Tal, 2017). As Tal (2017) argues, such a model is defined by theoretical and statistical assumptions about the measured psychological construct and its relation to the measurement task. Inference from the indication of a measurement instrument (e.g., a rating scale) to the measurement outcome is non-trivially derived from the model. Measurement is then the coherent and consistent assignment of values to parameters in this model, based on instrument indications. The model defines the content of the measurement outcome, which does not have to hold a counterpart in the observable world – a construct, in the end, is a concept, model, or schematic idea (McCoach et al., 2013). As for the measurement of intelligence, the values do not represent physical properties but empirical relationships between theoretical constructs and other constructs or behavior (Annett, 2002). Trust measurement, thus, may not deliver meaningful, absolute values per se but values that are meaningful in the context of a model of trust, which is defined by theoretical and statistical assumptions such as confirmed construct validity. In this way, the measurement outcome can be used to predict and explain behavior, decisions, or performance. For this reason, it is unreasonable to apply the same standards to the measurement of trust as to measurements such as take-over time. Nevertheless, the results of Körber et al. (2018) show that the questionnaire produces meaningful measures with relation to observable and safety-relevant behavior.

		Strongly disagree	Rather disagree	Neither disagree nor agree	Rather agree	Strongly agree	No response
1	The system is capable of interpreting situations correctly.	①	②	③	④	⑤	○
2	The system state was always clear to me.	①	②	③	④	⑤	○
3	I already know similar systems.	①	②	③	④	⑤	○
4	The developers are trustworthy.	①	②	③	④	⑤	○
5	One should be careful with unfamiliar automated systems.	①	②	③	④	⑤	○
6	The system works reliably.	①	②	③	④	⑤	○
7	The system reacts unpredictably.	①	②	③	④	⑤	○
8	The developers take my well-being seriously.	①	②	③	④	⑤	○
9	I trust the system.	①	②	③	④	⑤	○
10	A system malfunction is likely.	①	②	③	④	⑤	○
11	I was able to understand why things happened.	①	②	③	④	⑤	○
12	I rather trust a system than I mistrust it.	①	②	③	④	⑤	○
13	The system is capable of taking over complicated tasks.	①	②	③	④	⑤	○
14	I can rely on the system.	①	②	③	④	⑤	○
15	The system might make sporadic errors.	①	②	③	④	⑤	○
16	It’s difficult to identify what the system will do next.	①	②	③	④	⑤	○
17	I have already used similar systems.	①	②	③	④	⑤	○
18	Automated systems generally work well.	①	②	③	④	⑤	○
19	I am confident about the system’s capabilities.	①	②	③	④	⑤	○

		Stimme gar nicht zu	Stimme eher nicht zu	Stimme weder zu noch nicht zu	Stimme eher zu	Stimme voll zu	keine Angabe
1	Das System ist imstande Situationen richtig einzuschätzen.	①	②	③	④	⑤	○
2	Mir war durchgehend klar, in welchem Zustand sich das System befindet.	①	②	③	④	⑤	○
3	Ich kenne bereits ähnliche Systeme.	①	②	③	④	⑤	○
4	Die Entwickler sind vertrauenswürdig.	①	②	③	④	⑤	○
5	Bei unbekannten automatisierten Systemen sollte man eher vorsichtig sein.	①	②	③	④	⑤	○
6	Das System arbeitet zuverlässig.	①	②	③	④	⑤	○
7	Das System reagiert unvorhersehbar.	①	②	③	④	⑤	○
8	Die Entwickler nehmen mein Wohlergehen ernst.	①	②	③	④	⑤	○
9	Ich vertraue dem System.	①	②	③	④	⑤	○
10	Ein Ausfall des Systems ist wahrscheinlich.	①	②	③	④	⑤	○
11	Ich konnte nachvollziehen, warum etwas passiert ist.	①	②	③	④	⑤	○
12	Ich vertraue einem System eher, als dass ich ihm misstraue.	①	②	③	④	⑤	○
13	Das System kann wirklich komplizierte Aufgaben übernehmen.	①	②	③	④	⑤	○
14	Ich kann mich auf das System verlassen.	①	②	③	④	⑤	○
15	Das System könnte stellenweise einen Fehler machen.	①	②	③	④	⑤	○
16	Zu erkennen, was das System als Nächstes macht, ist schwer.	①	②	③	④	⑤	○
17	Ich habe ähnliche Systeme bereits genutzt.	①	②	③	④	⑤	○
18	Automatisierte Systeme funktionieren generell gut.	①	②	③	④	⑤	○
19	Ich bin überzeugt von den Fähigkeiten des Systems.	①	②	③	④	⑤	○

Items and corresponding scales.

Item	Scale
The system is capable of interpreting situations correctly.	Reliability/Competence
The system works reliably.	Reliability/Competence
A system malfunction is likely.*	Reliability/Competence
The system is capable of taking over complicated tasks	Reliability/Competence
The system might make sporadic errors.*	Reliability/Competence
I am confident about the system's capabilities.	Reliability/Competence
The system state was always clear to me.	Understanding/Predictability
The system reacts unpredictably.*	Understanding/Predictability
I was able to understand why things happened.	Understanding/Predictability
It's difficult to identify what the system will do next.*	Understanding/Predictability
I already know similar systems.	Familiarity
I have already used similar systems.	Familiarity
The developers are trustworthy.	Intention of Developers
The developers take my well-being seriously.	Intention of Developers
One should be careful with unfamiliar automated systems.*	Propensity to Trust
I rather trust a system than I mistrust it.	Propensity to Trust
Automated systems generally work well.	Propensity to Trust
I trust the system.	Trust in Automation
I can rely on the system.	Trust in Automation

Note. * = inverse item.

References

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior.* Englewood Cliffs, NJ: Prentice Hall.

Annett, J. (2002). Subjective rating scales: Science or art? *Ergonomics*, *45*(14), 966–987. https://doi.org/10.1080/00140130210166951

Bagozzi, R. P. (1982). The role of measurement in theory construction and hypothesis testing: Toward a holistic model. In C. Fornell (Ed.), *Praeger scientific. A second generation of multivariate analysis* (pp. 5–23). New York, NY: Praeger.

Barber, B. (1983). *The logic and limits of trust.* New Brunswick, N.J.: Rutgers University Press.

Beggiato, M., Pereira, M., Petzoldt, T., & Krems, J. F. (2015). Learning and development of trust, acceptance and the mental model of ACC. A longitudinal on-road study. *Transportation Research Part F: Traffic Psychology and Behaviour*, *35*, 75–84. https://doi.org/10.1016/j.trf.2015.10.005

Bisantz, A. M., & Seong, Y. (2001). Assessment of operator trust in and utilization of automated decision-aids under different framing conditions. *International Journal of Industrial Ergonomics*, *28*(2), 85–97. https://doi.org/10.1016/S0169-8141(01)00015-4

Blomqvist, K. (1997). The many faces of trust. *Scandinavian Journal of Management*, *13*(3), 271–286. https://doi.org/10.1016/S0956-5221(97)84644-1

Brown, R. D., & Galster, S. M. (2004). Effects of reliable and unreliable automation on subjective measures of mental workload, situation awareness, trust and confidence in a dynamic flight task. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 2004* (pp. 147–151).

Burisch, M. (1978). Construction strategies for multiscale personality inventories. *Applied Psychological Measurement*, *2*(1), 97–111. https://doi.org/10.1177/014662167800200110

Burisch, M. (1984). Approaches to personality inventory construction: A comparison of merits. *American Psychologist*, *39*(3), 214–227. https://doi.org/10.1037/0003-066X.39.3.214

Butler, J. K. (1991). Toward understanding and measuring conditions of trust: Evolution of a conditions of trust inventory. *Journal of Management*, *17*(3), 643–663. https://doi.org/10.1177/014920639101700307

Butler, J. K., & Cantrell, R. S. (1984). A behavioral decision theory approach to modeling dyadic trust in superiors and subordinates. *Psychological reports*, *55*(1), 19–28. https://doi.org/10.2466/pr0.1984.55.1.19

Damasio, A. R. (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *351*(1346), 1413–1420. https://doi.org/10.1098/rstb.1996.0125

Deutsch, M. (1958). Trust and suspicion. *Journal of Conflict Resolution*, *2*(4), 265–279. https://doi.org/10.1177/002200275800200401

Deutsch, M. (1960). The effect of motivational orientation upon trust and suspicion. *Human Relations*, *13*(2), 123–139. https://doi.org/10.1177/001872676001300202

Diamantopoulos, A., Sarstedt, M., Fuchs, C., Wilczynski, P., & Kaiser, S. (2012). Guidelines for choosing between multi-item and single-item scales for construct measurement: A predictive validity perspective. *Journal of the Academy of Marketing Science*, *40*(3), 434–449. https://doi.org/10.1007/s11747-011-0300-3

Drnec, K., Marathe, A. R., Lukos, J. R., & Metcalfe, J. S. (2016). From trust in automation to decision neuroscience: Applying cognitive neuroscience methods to understand and improve interaction decisions involved in human automation interaction. *Frontiers in Human Neuroscience*, *10*, 54. https://doi.org/10.3389/fnhum.2016.00290

Dzindolet, M. T., Beck, H. P., Pierce, L. G., & Dawe, L. A. (2001). *A framework of automation use* (No. ARL-TR-2412). Aberdeen Proving Ground, MD.

Eisinga, R., Grotenhuis, M. t., & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International journal of public health*, *58*(4), 637–642. https://doi.org/10.1007/s00038-012-0416-3

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2007). On the consistency of individual classification using short scales. *Psychological Methods*, *12*(1), 105–120. https://doi.org/10.1037/1082-989X.12.1.105

Feuerberg, B. V., Bahner, J. E., & Manzey, D. (2005). Interindividuelle Unterschiede im Umgang mit Automation – Entwicklung eines Fragebogens zur Erfassung des Complacency-Potentials [Interindividual differences in the interaction with automation – Development of a questionnaire to assess potential for complacency]. In L. Urbas & C. Steffens (Eds.), *Zustandserkennung und Systemgestaltung. 6. Berliner Werkstatt Mensch-Maschine-Systeme.* (pp. 199–202). Düsseldorf: VDI-Verlag.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Fuchs, C., & Diamantopoulos, A. (2009). Using single-item measures for construct measurement in management research: Conceptual issues and application guidelines. *Die Betriebswirtschaft*, *69*(2), 195.

Gigerenzer, G., & Selten, R. (Eds.). (2002). *Bounded rationality: The adaptive toolbox*. Cambridge, Mass: MIT Press.

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: A cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods*, *16*(3), 319–336. https://doi.org/10.1037/a0024917

Hergeth, S., Lorenz, L., & Krems, J. F. (2017). Prior familiarization with takeover requests affects drivers' takeover performance and automation trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *59*(3), 457–470. https://doi.org/10.1177/0018720816678714

Hergeth, S., Lorenz, L., Vilimek, R., & Krems, J. F. (2016). Keep your scanners peeled: Gaze behavior as a measure of automation trust during highly automated driving. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *58*(3), 509–519. https://doi.org/10.1177/0018720815625744

Hoeppner, B. B., Kelly, J. F., Urbanoski, K. A., & Slaymaker, V. (2011). Comparative utility of a single-item versus multiple-item measure of self-efficacy in predicting relapse among young adults. *Journal of substance abuse treatment*, *41*(3), 305–312. https://doi.org/10.1016/j.jsat.2011.04.005

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Hoffman, R. R., Johnson, M., Bradshaw, J. M., & Underbrink, A. (2013). Trust in automation. *IEEE Intelligent Systems*, *28*(1), 84–88. https://doi.org/10.1109/MIS.2013.24

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185. https://doi.org/10.1007/BF02289447

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hubley, A. M., & Zumbo, B. D. (2013). Psychometric characteristics of assessment procedures: An overview. In K. F. Geisinger (Ed.), *APA handbooks in psychology: Vol. 1. Test theory and testing and assessment in industrial and organizational psychology* (pp. 3–20). Washington, D.C.: American Psychological Association.

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2014). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, *44*(3), 486–507. https://doi.org/10.1177/0049124114543236

Kirlik, A. (1993). Modeling strategic behavior in human-automation interaction: Why an "aid" can (and should) go unused. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *35*(2), 221–242. https://doi.org/10.1177/001872089303500203

Körber, M. (2018). Individual differences in human-automation interaction: A driver-centered perspective on the introduction of automated vehicles (Dissertation). Technical University of Munich, Munich.

Körber, M., Baseler, E., & Bengler, K. (2018). Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied Ergonomics*, *66*, 18–31. https://doi.org/10.1016/j.apergo.2017.07.006

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (2007). *Additive and polynomial representations. Foundations of measurement: Vol. 1.* Mineola, NY: Dover Publ.

Lee, J. D., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243–1270. https://doi.org/10.1080/00140139208967392

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Love, J., Selker, R., Verhagen, J., Marsman, M., Gronau, Q. F., Jamil, T.,. . . Rouder, J. N. (2015). Software to sharpen your stats. *APS Observer*, *28*(3), 27–29.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130–149. https://doi.org/10.1037/1082-989X.1.2.130

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84–99. https://doi.org/10.1037/1082-989X.4.1.84

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277–301. https://doi.org/10.1080/14639220500337708

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. In *Proceedings of the 11th Australasian Conference on Information Systems* (pp. 6–8).

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.5465/AMR.1995.9508080335

Mayer, R. C., & Davis, J. H. (1999). The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of Applied Psychology*, *84*(1), 123–136. https://doi.org/10.1037//0021-9010.84.1.123

McCoach, D. B. (2003). SEM isn't just the Schoolwide Enrichment Model anymore: Structural equation modeling (SEM) in gifted education. *Journal for the Education of the Gifted*, *27*(1), 36–61. https://doi.org/10.1177/016235320302700104

McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). *Instrument development in the affective domain: School and corporate applications* (3rd ed.). New York, NY: Springer. Retrieved from http://dx.doi.org/10.1007/978-1-4614-7135-6

McKnight, D. H., & Chervany, N. L. (1996). *The meanings of trust* (WP No. 96-04).

McKnight, D. H., & Chervany, N. L. (2001). Trust and distrust definitions: One bite at a time. In R. Falcone, M. Singh, & Y.-H. Tan (Eds.), *Lecture notes in computer science: Vol. 2246. Trust in Cyber-societies: Integrating the Human and Artificial Perspectives* (pp. 27–54). Berlin, Heidelberg: Springer.

McNeish, D. (2017). Thanks coefficient alpha, we'll take it from here. *Psychological Methods.* Advance online publication. https://doi.org/10.1037/met0000144

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(2), 194–210. https://doi.org/10.1518/001872008X288574

Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(2), 196–204.

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*(3), 355–383. https://doi.org/10.1111/j.2044-8295.1997.tb02641.x

Moosbrugger, H., & Kelava, A. (Eds.). (2012). *Testtheorie und Fragebogenkonstruktion* [Test theory and construction of questionnaires] (2nd ed.). *Springer-Lehrbuch.* Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg.

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*(11), 1905–1922. https://doi.org/10.1080/00140139408964957

Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, *39*(3), 429–460. https://doi.org/10.1080/00140139608964474

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, *27*(5-6), 527–539. https://doi.org/10.1016/S0020-7373(87)80013-5

Nagy, M. S. (2002). Using a single-item approach to measure facet job satisfaction. *Journal of Occupational and Organizational Psychology*, *75*(1), 77–86. https://doi.org/10.1348/096317902167658

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). *McGraw-Hill.* New York, NY: McGraw-Hill.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics*, *30*(3), 286–297. https://doi.org/10.1109/3468.844354

Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 224–239). New York, NY: The Guilford Press.

Preacher, K. J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: A model selection perspective. *Multivariate behavioral research*, *48*(1), 28–56. https://doi.org/10.1080/00273171.2012.710386

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology*, *49*(1), 95–112. https://doi.org/10.1037/0022-3514.49.1.95

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373. https://doi.org/10.1037/a0029315

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, *27*(2), 151–161. https://doi.org/10.1177/0146167201272002

Rotter, J. B. (1971). Generalized expectancies for interpersonal trust. *American Psychologist*, *26*(5), 443–452. https://doi.org/10.1037/h0031464

Sakaluk, J. K., & Short, S. D. (2017). A methodological review of exploratory factor analysis in sexuality research: Used practices, best practices, and data analysis resources. *Journal of sex research*, *54*(1), 1–9. https://doi.org/10.1080/00224499.2015.1137538

Scarpello, V., & Campbell, J. P. (1983). Job satisfaction: Are all the parts there? *Personnel Psychology*, *36*(3), 577–600. https://doi.org/10.1111/j.1744-6570.1983.tb02236.x

Schoettle, B., & Sivak, M. (2014). *Public opinion about self-driving vehicles in China, India, Japan, the U.S., the U.K., and Australia* (No. UMTRI-2014-30). Ann Arbor, MI.

Sloan, J. A., Aaronson, N., Cappelleri, J. C., Fairclough, D. L., & Varricchio, C. (2002). Assessing the clinical significance of single items relative to summated scores. *Mayo Clinic Proceedings*, *77*(5), 479–487. https://doi.org/10.4065/77.5.479

Tal, E. (2017). Measurement in science. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy.* Metaphysics Research Lab, Stanford University.

Uebersax, J. S. (2006). Likert scales: Dispelling the confusion. Retrieved from http://www.john-uebersax.com/stat/likert.htm