Eileen Roesler, Tobias Rieger, Dietrich Manzey

Trust towards human vs. automated agents: Using a multidimensional trust questionnaire to assess the role of performance, utility, purpose, and transparency

Trust towards Human vs. Automated Agents: Using a Multidimensional Trust Questionnaire to Assess The Role of Performance, Utility, Purpose, and Transparency

Eileen Roesler*
Technische Universität Berlin

Tobias Rieger*
Technische Universität Berlin

Dietrich Manzey
Technische Universität Berlin

In various domains, humans are supported by automated systems. Earlier research has suggested that trust in automated agents differs from trust in other humans. The present studies aimed at taking a multi-dimensional look at effects on trust towards automation and humans. To this end, we conducted two studies to empirically validate a multi-dimensional trust questionnaire to assess performance, utility, purpose, and transparency subdimensions of trust (Study 1, $N = 160$) and to study experimental effects of support agent (i.e., human vs. decision support system) and failure experience (i.e., none vs. one; Study 2, $N = 181$). The expected factor structure was confirmed. Moreover, the results showed that being supported by a human mostly impacted the performance subscale. In sum, the findings illustrate the importance to study trust not only uni-dimensionally but to consider different subdimensions, particularly as a single-item trust measurement was mostly correlated to performance and utility subscales.

For a plethora of critical decision-making tasks, such as deciding about applicants' loans or evaluating medical x-rays, humans are often assisted by support agents. These support agents previously were mainly human experts, but are increasingly replaced by automated systems. The general idea of having critical tasks supported by either another human or an automated decision support system (DSS) is to improve safety and performance. Unfortunately, the joint performance of both human-human teams (e.g., Cymek, 2018) as well as human-DSS teams (e.g., Rieger & Manzey, 2022) is often worse than ideal. Moreover, trust towards these different types of support agents has been found to differ (for a review, see Madhavan & Wiegmann, 2007). One finding regarding the differences between human-human and human-DSS trust, which has been often referred to is the so-called *perfect automation schema* by Dzindolet, Pierce, Beck, and Dawe (2002). The central proposition of the perfect automation schema is that, a priori, humans have higher trust towards automation, but after a failure has been experienced, trust decreases more strongly when interacting with a DSS than when interacting with another human. Dzindolet et al. (2002) attributed this difference to expectations of near-perfect performance towards automated systems but not towards humans.

However, some recent findings challenge this often referred effect of a perfect automation schema. More specifically, Rieger, Roesler, and Manzey (2022) compared trust in human expert support and automated decision support in different domains. In three experiments with consistent results, they even found higher trust in the human agent than the automation, suggesting something like an imperfect automation schema. Moreover, failure experience reduced trust in both the DSS and the human condition—as would be expected from most theoretical models of human-automation interaction (e.g., Hoff & Bashir, 2015). Crucially, the size of the failure effect did not differ between the DSS and the human condition, indicating that both trust towards humans as well as DSS support suffered from failure experience.

Both the findings of Rieger et al. (2022) and Dzindolet et al. (2002) rely on rather simple assessments of trust or utility, and did not consider a multi-faceted view of trust. Even though both studies found differences between humans and automation, neither can address which underlying trust facets play a role for the respective differences. For instance, a single-item trust (Rieger et al., 2022) is not capable of capturing different trust dimensions and a more detailed look might be necessary to address theory-based trust dimensions.

More specifically, most theoretical models of trust in automation (and also, of trust in humans) consider more than one single dimension. For instance, in their seminal work on trust in automation, Lee and See (2004) describe performance, purpose, and process as separate dimensions of trust. They consider performance as what the automation does, mainly in terms of reliability and the ability to help the operator (or advice-taker) to achieve one's goals. Purpose refers to why the automation was implemented, and also includes the designer's intentions and whether the system's benefit is comprehensible. Finally, process is about how the system works. More specifically, it includes aspects such as comprehensibility and familiarity of the system.

Often-used single-items of trust or uni-dimensional trust-in-technology questionnaires (e.g., Jian, Bisantz, & Drury, 2000) do not capture these different dimensions, and thus cannot contribute much to a more detailed understanding of the underlying determinants of trust and trust dynamics in interaction with different support agents. A more detailed multi-dimensional trust-in-automation questionnaire, which might also be used for investigating trust in human support agents, has been proposed by Wiczorek (2011, Multi-dimensional Trust Questionnaire; MTQ). Theoretically, it is based on the concept of Lee and See (2004). Specifically, the performance and purpose dimensions are directly assessed as separate dimensions. Further, even though the process dimension is not directly addressed, the MTQ includes a transparency subscale which is

*Shared first authorship. The hypotheses for Study 2 were preregistered. All data, analysis code, and an implementation of the questionnaire for jspsych is available at the Open Science Framework under osf.io/56cwx.

closely related to the process dimension of Lee and See (2004). Finally, for a more differentiated assessment of the performance dimension of Lee and See (2004) which also includes aspects of usefulness, the MTQ separately assesses a utility subscale which addresses how useful a system is to fulfill the task. In summary, the MTQ (Wiczorek, 2011) allows for a more fine-grained assessment of trust in order to gain a better understanding which trust dimensions are impacted from a given experimental manipulation.

However, thus far, the questionnaire was only available in German. Given this, the current research has two objectives. In Study 1, we aimed at a psychometric evaluation of an English version of this questionnaire. In Study 2, we conducted an experimental study to gain a better understanding of the trust facets impacted by receiving decision support from either a DSS or another human as well as from failure experience.

STUDY 1 – QUESTIONNAIRE

As mentioned above, the goal of Study 1 was to translate and validate the Wiczorek (2011) questionnaire. To this end, we conducted an online study where, after experiencing working together with a perfectly reliable DSS, participants evaluated the system with a single-item trust and the MTQ with its four subscales (performance, utility, purpose, transparency; with four items each). The stimuli and procedure were the same as in Rieger et al. (2022, Experiment 2; see also, e.g., Appelganc, Rieger, Roesler, & Manzey, 2022). The data was used for psychometric analyses of the factor structure of the questionnaire. Moreover, intercorrelations of the MTQ's subscales and a global single-item trust were assessed.

Method

Participants. As the reliability of factor analysis is dependent on the sample size, we aimed to have 10 times as many participants as variables (Nunnally, 1978). As the MTQ consists of 16 items, 164 participants were recruited via Prolific (4 participants failed an attention check) to achieve a final sample size of 160 participants (mean age = 33.60, SD = 5.73, 52.5% female).

Procedure and Design. Participants took part in the experiment in their own web browser and the experiment was programmed in jspsych (de Leeuw, 2015) and ran on a JATOS server (Lange, Kühn, & Filevich, 2015). After giving their informed consent, the experiment started with a general introduction on the simulated radiological x-ray screening task they would perform later on. Specifically, the task was to evaluate which percentage of simulated x-rays were brighter than a given cutoff (brighter than a grayscale value of 150). In order to enable participants performing this task, they were first shown a grayscale continuum, with the critical cutoff threshold marked, along with an example image. Participants were instructed that parts brighter than this could possibly be malignant and that later on, they would be asked to estimate the percentage of potentially malignant tissue in x-ray samples. They were also informed that this was a very cautious cutoff with no reason for any concern with percentages lower than 15%.

After this general introduction of the task, participants were shown three example scenarios with the correct solution. After these example scenarios, participants were informed that they would receive support from a highly reliable DSS but they would make the final decisions. The DSS was characterized as a well-established system which made its decisions based on prior x-ray data. Participants were also informed that the DSS's reliability was greater than 90%. Subsequent to this framing, participants were asked to answer two short attention check questions about the support agent.

Then, the main part of the experiment started. Specifically, participants were instructed that they would now be evaluating fictional personas' x-rays and were shown an example. Participants worked on a total of 10 trials where the DSS's recommendation was always correct. On each trial, the persona with the x-ray image was shown for 5 seconds along with the information that the DSS is currently evaluating the image. After 5 seconds, participants were told that they can now press the spacebar to continue to see the DSS's recommendation. Then, they filled in their final decision in an input field. After each trial, participants needed to press the spacebar to continue to the next trial.

Subsequent to these ten experimental trials, participants filled out two trust measures. That is, trust was assessed both, using a single-item trust (i.e., "how much do you trust the decision support system?" from 0 (not at all) to 100 (completely)), as well as the translated version of the MTQ (translated from Wiczorek, 2011), measured on a four-point Likert scale ("Disagree", "Somewhat disagree", "somewhat agree", "Agree"). Besides the four scales of the original MTQ, we embedded an additional attention check item in the MTQ to ensure data quality. The order of the two trust measures was counterbalanced across participants. Finally, participants filled out a short sociodemographic questionnaire and were debriefed.

Results

Item Reduction & Scale Analysis. In order to make the questionnaire more economical and reliable, items of the translated scale were removed if the respective item did not improve the reliability of the scale (4-item scales $\alpha_{PE} = .89$, $\alpha_U = .85$, $\alpha_{PU} = .50$, $\alpha_T = .85$). This led to the reduction of one item per scale, resulting in three items per subscale with acceptable to excellent internal consistencies (3-item scales $\alpha_{PE} = .92$, $\alpha_U = .87$, $\alpha_{PU} = .74$, $\alpha_T = .85$).

Factor Analysis. In order to analyze the underlying factor structure of the remaining 12 items, we performed an exploratory maximum likelihood factor analysis with oblique rotation (promax). Before conducting this factor analysis, we checked for the appropriateness of sample size and the data. Both the significant Bartlett's test of sphericity ($X^2(66) = 1371.58$, $p < .001$) and the very good Kaiser-Meyer-Olkin measure of sampling adequacy (KMO=.88) indicated that the sample size and data is adequate for the following factor analysis.

To determine the optimal number of factors for the exploratory factor analysis, a visual inspection of a screeplot (Cattell, 1966) and the Kaiser's criterion (Kaiser, 1960) of Eigenvalues were used. The plot indicated a three factorial structure. Therefore, the factor analysis was performed with three factors.

Table 1 displays the obtained pattern matrix by showing

factor loadings above .40 (Stevens, 2009). The first factor comprised all items targeting performance as well as utility subscale and accounted for 37% of the variance. The second factor included the transparency items and accounted for 18% of the variance. Factor three included all purpose items and accounted for 13% of the variance. In sum, the overall scale accounted for 68% of the variance of the data.

Table 1. *Items and their factor loadings.*

		Factors		
	Items	1	2	3
PE1	The system works safely.	0.87		
PU1	The intention of the system is positive.			0.56
PE2	The system works well.	0.86		
T1	The way the system works is clear to me.		0.49	
U1r	The system makes my work more difficult.	0.80		
PU2	The system is intended to help improve overall performance.			0.60
PE3	The system works accurately.	0.91		
T2	I am well informed how the system works.		0.91	
U2	The system is useful for my work.	0.80		
T3	I understand how the system works.		0.94	
PU3	The system was implemented to help me.			0.82
U3	I find that the system supports my work.	0.72		

Internal Consistency and Validity. After removing one item per subscale, the overall scale (α = .91) showed an excellent internal consistency. This was also the case for the combined reliability-utility scale (α = .94). Furthermore, the purpose scale (α = .74) showed an acceptable and the transparency scale (α = .85) a good internal consistency. To investigate the construct validity, the single-item trust measurement was correlated with the overall scale mean and the three single scales. The results showed that the overall scale (r = .77) as well as the performance-utility scale (r = .83), the purpose scale (r = .52), and the transparency scale (r = .39) highly correlated with single-item trust (all $ps < .001$).

Discussion

The aim of the questionnaire study was twofold. On the one hand, we wanted to validate the factorial structure of the translated questionnaire. On the other hand, we aimed to study the interconnection between the uni-dimensional single-item measurement of trust and the different dimensions of the MTQ.

First, the analysis of the trust questionnaire revealed three dimensions of trust. Whereas two of the factors directly corresponded to the purpose and transparency (process) dimension of trust according to the original conceptualization of these dimensions of trust by Wiczorek (2011), performance and utility loaded on the same factor. This latter effect could relate to a conceptual relationship of these two aspects which might even be combined to one scale resulting in a three-factorial trust concept congruent with the one of Lee and See (2004). However, the results could also be associated with the type of system (a simple DSS) and the specific paradigm (a single-task x-ray estimation) we used. More specifically, it seems reasonable to assume that utility and performance might end up as different dimensions in a much more complex task environment (e.g., supervisory control settings) with an imperfect system. For these methodological reasons, we would advise against our factor analysis and recommend to consider performance and utility as separate scales in future research before the high correlation between these two dimensions has been replicated with other systems.

Second, the correlation of the uni-dimensional item with the MTQ overall score revealed a high correlation which indicates a sufficient content validity of the questionnaire. Interestingly, the uni-dimensional measure seems to be most prominently related to performance and utility. The uni-dimensional assessment of trust thus might be primarily related to the outcome of the system reflecting its reliability and validity. However, this measure seems to be considerably less related to how the agent operates. Therefore, the MTQ enables insights to the process component of trust (Lee & See, 2004) by assessing purpose and transparency.

In summary, the results of Study 1 suggest that the MTQ is a valid instrument to assess trust multi-dimensionally. Thus, we conducted Study 2 using the MTQ in order to study why humans are trusted more than automation.

STUDY 2 - EXPERIMENTAL STUDY

Study 2 had three goals. First, we wanted to study which trust subdimensions are affected by different support agents (i.e., human vs. DSS). Second, we were interested to see which subdimensions are impacted from failure experience. Third, we were also interested in checking whether the subdimensions' correlations with the single-item trust are descriptively different for both types of agents.

Regarding the first goal, we expected to find a main effect of support agent on single-item trust as in Rieger et al. (2022), with higher trust for human than for automation support. To further disentangle this trust effect, the MTQ was used. Here, we hypothesized that the main effect of agent is only found for the purpose and transparency scales. We expected an effect for the purpose subscale because it refers to benevolence (i.e., a positive orientation) of the interaction partner (Lee & See, 2004) which humans can express but automation cannot. Moreover, we expected an effect for the transparency subscale as the lack of system transparency is often criticized for automation (e.g., Wickens, 2017) but obviously not for humans.

Regarding the second goal, we again expected a main effect for the single-item trust with lower trust after a failure experience than after experiencing perfect support. We also had subscale-specific hypotheses. That is, we expected failure experience to mainly impact the performance and utility subscales of the MTQ. We expected particularly these subscales to reflect the failure effect because performance is directly linked to true reliability and utility is closely linked to the support agent's capabilities, which are of course smaller when reliability is lower. Moreover, as Rieger et al. (2022) did not observe any interaction effects of agent and failure experience throughout three experiments, we did not expect to observe one here–neither for the single item nor for any MTQ subscale.

Finally, we had no clear hypotheses with respect to the third goal and checked the correlations in an exploratory manner.

Method

Participants. A fresh sample of 200 participants (mean age = 32.54, SD = 5.16, 51.4% female) was recruited via Prolific. One additional participant was also tested but excluded from any further analyses due to a failed attention check.

Procedure and Design. The procedure and design were very similar to Study 1. Specifically, whereas in Study 1 all participants had a DSS as their support agent and the ten trials were without failure experience, both support agent and failure experience were systematically varied across participants in Study 2. This resulted in a 2 (support agent: human vs. DSS) x 2 (failure experience: none vs. one) between-subjects design.

In the human condition, participants were told that they would be supported by an experienced colleague during the ten trials, and also that their colleague was more than 90% reliable. In the failure conditions, the respective support agent made an obvious error in trial 7, indicating a very low percentage where the true proportion of bright pixels was rather high (59% brighter than the cutoff was evaluated to be okay with 8% as the recommendation).

Based on the results of the factor analysis conducted in Study 1, we used a reduced version of the MTQ with only three items per subscale. Moreover, all questions were adapted to fit the respective condition (i.e., the wording of the questions always referred to either a "decision support system" or a "colleague"). The order of the single item trust and the MTQ was again counterbalanced across participants.

Results

First, we performed a manipulation check in the failure condition. We excluded 19 participants who exactly followed the obviously wrong advice of the support agent, as it is not certain that they detected the failure. Subsequently, the effects of the dependent variables were analyzed by 2 (support agent) x 2 (failure condition) ANOVAs.

The analysis of single-item trust revealed no significant main effect of failure or interaction effect ($ps > .105$). However, the main effect of support agent $F(1, 177) = 3.84$, $p = .052$, $\eta^2_G = .021$ just failed to reach the conventional level of significance. Participants tended to report higher trust towards the human ($M = 80.40$; $SD = 18.75$) compared to the DSS ($M = 74.33$; $SD = 19.71$).

Neither the analysis of the overall MTQ score ($ps > .191$), nor the analyses of the dimensions purpose ($ps > .283$), utility ($ps > .105$), or transparency ($ps > .395$) revealed any significant effects. Only for the performance dimension a significant main effect of the support agent was found $F(1, 177) = 9.61$, $p = .002$, $\eta^2_G = .052$, indicating higher perceived performance of the human ($M = 3.52$; $SD = 0.58$) compared to the DSS ($M = 3.23$; $SD = 0.64$). However, neither the main effect of failure nor the interaction effect were significant ($ps > .225$).

Finally, we analyzed the correlations between the different MTQ scales and the single-trust item. Pearson's product-moment correlations were calculated separately for the human support and the DSS to investigate whether different aspects of trust are important for different support agents. Figure 1 shows the respective bivariate correlations. For the human condition, the single item correlated highly with performance ($r = .76$), utility ($r = .76$), purpose ($r = .53$), and transparency ($r = .59$). For the DSS, high correlations of the single item with performance ($r = .71$) and utility ($r = .56$), as well as weak correlations with purpose ($r = .22$) and transparency ($r = .30$) were found. Whereas the correlation between the single item and the performance scale did not significantly differ between human and DSS ($z = 0.75, p = .453$), the correlations of the single item with the utility ($z = 2.41, p = .016$), purpose ($z = 2.49, p = .013$), and transparency ($z = 2.42, p = .016$) scale were significantly higher for the human compared to the DSS. We decided against separate correlations for the failure conditions as no significant differences occurred in the ANOVAs.

Discussion

The second study aimed to investigate what subdimensions of trust are affected by the reliability of the system and the type of the support.

Surprisingly, in contrast to numerous earlier findings (Dzindolet et al., 2002; Hoff & Bashir, 2015; Rieger et al., 2022), there was no significant effect of failure experience for any of the trust measures. Perhaps, the fact that the MTQ is typically measured on a four-point Likert scale might have contributed to a potential ceiling effect, hiding a true effect of failure experience. Moreover, Miller (2009) pointed out that "even when a real effect is present, some replication failures must be expected as one of the unfortunate consequences of variability" (p. 617), and the present study might be just one of those unfortunate consequences.

For the type of support, the trend of the single item trust and the significant main effect of the MTQ performance scale stand in contrast to the perfect automation schema (Dzindolet et al., 2002). In line with earlier research (Appelganc et al., 2022; Rieger et al., 2022), the results indicate that human expert support tends to be trusted more than DSS. The results further broaden the research body by showing that higher trust towards humans might be related to higher associated performance. Given that the actual reliability of the two agents was equal, this might look surprising. However, perceived and actual reliability are often not the same (Madhavan & Wiegmann, 2007; Rieger et al., 2022), and the type of agent whose reliability is perceived can also make a difference here. One possible reason why medical experts are perceived as more capable than an automated system might be the pre-existing knowledge in regard to expectation and reputation (Hoff & Bashir, 2015). The present results indicate that human expertise in prestigious domains like the medical one (Hauser & Warren, 1997) seem to exceed the prestige and expectations of expert systems. Since the support was highly reliable for all conditions (95% reliable on average), the expectations might be maintained throughout the collaboration. Future research is necessary to explicitly test this assumption by comparing the preexisting knowledge and the reputation of human and automated expert support.

With respect to the question which subdimensions might be most important for the global trust towards support agents, the second study validated the finding that uni-dimensional trust is most strongly associated with performance and utility of a sup-

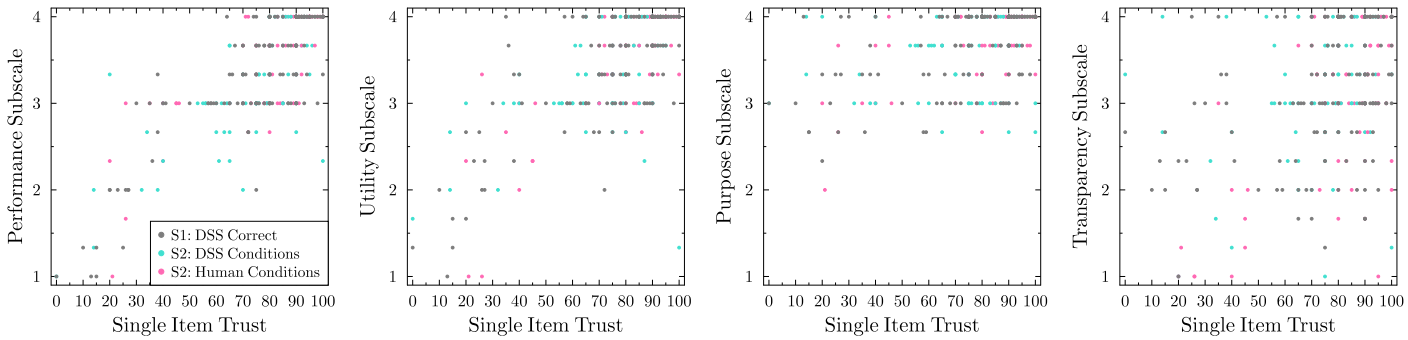

Figure 1. Bivariate correlations between the single-item trust and the performance, utility, purpose, and transparency subscales.

port for both the human and automated support. Interestingly, purpose and transparency are related to the single item trust strongly for human-human interaction and weakly for human-automation interaction. The direct comparison of the correlations showed that all scales besides the performance scale are less correlated with human-automation trust than human-human trust. The result is in line with earlier research illustrating that the reliability of an agent is the one of the most important determinants of trust in human-automation interaction (Hoff & Bashir, 2015). Whereas a performance-oriented trust concept might be sufficient for classical DSS, the recent technological trend towards self learning systems also challenges the trust dimensions which are more associated with human support (i.e., transparency, and purpose). Artificially intelligent technologies extend DSS with a higher autonomy and agency (Legaspi, He, & Toyoizumi, 2019). Consequently, those technologies are often linked to a possible lack of transparency. This is also supported by public scandals that question not only the transparency, but also the positive intention of these systems (O'Neil, 2016). Therefore, it is particularly relevant to also include trust facets that go beyond performance and utility in research which investigates application domains with real-life relevance. Thus, future research should include a multi-dimensional view at trust, particularly with novel application domains.

CONCLUSION

Overall, both studies illustrate the importance to approach trust on a multi-dimensional level. Uni-dimensional trust measures most prominently account for the performance and utility subdimensions of trust towards a support agent. However, purpose and transparency gain increasing importance as (intelligent) automated systems increasingly make decisions that affect people's lives (O'Neil, 2016). Thus, future research studying trust in agents (e.g., automated systems, robots, other humans, etc.) should not only consider performance and utility aspects, but also take issues like purpose and transparency into account.

REFERENCES

Appelganc, K., Rieger, T., Roesler, E., & Manzey, D. (2022). How much reliability is enough? a context-specific view on human interaction with (artificial) agents from different perspectives. *Journal of Cognitive Engineering and Decision Making*. doi: 10.1177/15553434221104615

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276. doi: 10.1207/s15327906mbr0102_10

Cymek, D. H. (2018). Redundant automation monitoring: Four eyes don't see more than two, if everyone turns a blind eye. *Human Factors*, *60*(7), 902–921. doi: 10.1177/0018720818781192ï¿¡

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12. doi: 10.3758/s13428-014-0458-y

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, *44*(1), 79–94. doi: 10.1518/0018720024494856

Hauser, R. M., & Warren, J. R. (1997). 4. socioeconomic indexes for occupations: A review, update, and critique. *Sociological Methodology*, *27*(1), 177–298. doi: 10.1111/1467-9531.271028

Hoff, K. A., & Bashir, M. (2015). Trust in automation. *Human Factors*, *57*(3), 407–434. doi: 10.1177/0018720814547570

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. doi: 10.1207/s15327566ijce0401_04

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*(1), 141–151. doi: 10.1177/001316446002000116

Lange, K., Kühn, S., & Filevich, E. (2015). "Just Another Tool for Online Studies" (JATOS): An easy solution for setup and management of web servers supporting online studies. *PLOS ONE*, *10*(6), e0130834. doi: 10.1371/journal.pone.0130834

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. doi: 10.1518/hfes.46.1.50_30392

Legaspi, R., He, Z., & Toyoizumi, T. (2019). Synthetic agency: sense of agency in artificial intelligence. *Current Opinion in Behavioral Sciences*, *29*, 84–90. doi: 10.1016/j.cobeha.2019.04.004

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. doi: 10.1080/14639220500337708

Miller, J. (2009). What is the probability of replicating a statistically significant effect? *Psychonomic Bulletin & Review*, *16*(4), 617–640. doi: 10.3758/pbr.16.4.617

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Rieger, T., & Manzey, D. (2022). Human performance consequences of automated decision aids: The impact of time pressure. *Human Factors*, *64*(4), 617–634. doi: 10.1177/0018720820965019

Rieger, T., Roesler, E., & Manzey, D. (2022). Challenging presumed technological superiority when working with (artificial) colleagues. *Scientific Reports*, *12*(1). doi: 10.1038/s41598-022-07808-x

Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences*. New York: Routledge.

Wickens, C. D. (2017, oct). Automation stages & levels, 20 years after. *Journal of Cognitive Engineering and Decision Making*, *12*(1), 35–41. doi: 10.1177/1555343417727438

Wiczorek, R. (2011). Entwicklung und Evaluation eines mehrdimensionalen Fragebogens zur Messung von Vertrauen in technische Systeme. In *Reflexionen und Visionen der Mensch-Maschine-Interaktion–Aus der Vergangenheit lernen, Zukunft gestalten* (Vol. 9, pp. 621–626).