Cross-Country Validation of a Cultural Scale in Measuring Trust in Automation

Shih-Yi Chien, Michael Lewis
School of Information Sciences
University of Pittsburgh
Pittsburgh, PA, 15260 U.S.A.
shc56@pitt.edu, ml@sis.pitt.edu

Sebastian Hergeth
BMW Group
Research & Technology
Munich, Germany
Sebastian.Hergeth@bmw.de

Zhaleh Semnani-Azad
Department of Psychology
University of Waterloo
Waterloo, Ontario, Canada
zsemnani@uwaterloo.ca

Katia Sycara
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213 U.S.A.
katia@cs.cmu.edu

Abstract— Human automation interaction is a complex process. How autonomous assistance impacts trust in automation as well as how trust affects human calibration and use of automation has been investigated for both dynamic contexts, including the internal variables (e.g., cultural characteristics) and external factors (e.g., system settings). Having standardized measures to capture trust and its antecedents is particularly critical to understanding how factors associated with the human operators and autonomous applications affects the way they are used. This paper reports the development of a trust instrument and several rounds of cross-country validation, including U.S., German, Taiwanese, and Turkish populations. The results confirm that the instrument which was developed reliably measured human trust in automation across cultures.

INTRODUCTION

The use of autonomous systems in different application domains to reduce the situations of higher risk and improve task performance is rapidly increasing. The intersection of automation and trust for humans interacting with multi-agent systems involves a variety of factors and influences the automation uses and resulting behaviors. How automated assistance can be adopted to increase uses and trust in automation as well as how trust in the systems impacts system robustness have been longstanding questions. Researches (Parasuraman & Riley, 1997; Lee & See, 2004) have shown inappropriate trust (e.g., over-reliance vs. under-reliance) in automated systems engenders misuses of automation which potentially lead to catastrophic consequences, such as the loss of cooling at Three Mile Island. Trust in automation refers to a cognitive state or attitude yet it has most often been studied indirectly through its purported influence on behavior often without any direct cognitive measure. The nature and complexity of the tasks and failures studies have varied greatly, ranging from simple automatic target recognition classification (Dzindolet, Pierce, Beck, & Dawe, 2002) to erratic responses of a controller embedded within a complex automated system (Lee & Moray, 1992). The variety of reported effects (automation bias, complacency, reliance, compliance, etc.) mirrors these differences in tasks and scenarios.

The effects of robust autonomous intelligence and human trust across a variety of contexts largely impact trust in automation (Lee & See, 2004). Cultural contexts significantly influence trust attitude and behaviors with respect to automation. Individuals from different cultures may exhibit varying reliance on automation. Li, Rau and Li (2010) examined the hypothesis that when robots behave in more culturally appropriate ways, humans would accept their recommendations more readily. In particular, Li, Rau and Li (2010) studied the effects of culture-sensitive communication style on the willingness of people to accept advice from a robot. They found that Chinese participants preferred an implicit communication style on the part of the robots whereas Germans preferred an explicit communication style. These findings are congruent with the culturally preferred interpersonal styles in these cultures. Additional results of the same study found that the Chinese evaluated the robot as being more likable and trustworthy and were more likely to accept its implicit recommendation than the Germans.

The possibility that human's trust is influenced by other factors and context beyond automation performance has not been adequately considered in the trust calibration literature. Thus, factors (individual and social) unrelated to automation performance could be crucial to understanding the calibration of trust in automation for operators from different cultures that are interacting with the same automation. To examine the interrelations and commonalities of concepts involving trust in automation, empirical research is needed to integrate divergent manifestations of trust within a single task/test population so that common and comparable measures can be developed. The three dimensional (performance, process and purpose) model proposed by Lee and See (2004) presumes that trust is influenced by a person's knowledge of what the automation is supposed to do (purpose) and how it functions (process) as well as the actual performance. This three dimensional structure fits nicely with Mayer and Davis' (1999) Ability-Benevolence-Integrity (ABI) definition that has been widely adopted in social psychological studies of trust and suggests

that candidate items for an instrument measuring trust in automation should contain at least these dimensions. While such models seem plausible, support for the contribution of factors has typically been limited to correlation between questionnaire responses and automation use. Wasti, Tan, Brower, and Onder (2007), for example, found the widely used ABI trust scale to have "poor psychometric properties across the board" when attempting to assess measurement invariance across samples from the U.S., Turkey and Singapore. The problem of studying cross-cultural effects of trust is exacerbated by the lack of standardized measures. Principles cross cultural studies of trust in automation will require developing reliable and valid measures of trust that can allow accurate comparisons across cultures. Most of the existing works on cultural effects on trust in automation is abstract and suggestive and needs empirical validation. This is the gap that the present research proposes to fill.

CULTURAL FACTORS

Culture is an important modulator in findings on trust in interpersonal relations (Yamagishi & Yamagishi, 1994). Cultural values and norms can greatly influence an individual's trust and reliance on the automation as well as the formation, dissolution and restoration of trust. Prior work examining cultural influence employed Hofstede's cultural dimension (1991) to predict psychological and behavioral outcomes of individuals from different cultures. Although the use of Hofstede's dimensions is very useful to understand how people of different cultures behave generally, it has recently been observed (Leung & Cohen, 2011) that Hofstead's cultural metrics ignore how an individual's behaviors in terms of adherence to cultural norms interact with situations and consequently influence the values by a particular member. Thus we also examined "cultural syndromes" (Triandis, 1996) encompassing cultures of Dignity, cultures of Honor, and cultures of Face. These two cultural mechanisms guide the formation of our research and further empirical studies of how culture may affect antecedents and components of trust.

Hofstede cultural dimensions

To examine the cultural effects on trust in automation, the three of the cultural dimensions proposed by Hofstede (1991), which have been the most studied in the literature were adopted for our studies.

Power Distance (PD) is defined as "the extent to which the less powerful accept and expect that power is distributed unequally (Hofstede, 1991)." People in high PD societies expect authority figures to be benign, competent, and of high integrity (House, Javidan, & Hanges, 2002). Thus in large PD cultures people will engage in less vigilance and monitoring for possible violations by authority figures. To the extent then that people of high power distance cultures perceive the automation as authoritative, they should be quick to form trust.

Individualism/Collectivism (IDV) is "the degree of interdependence a society maintains among its members (Hofstede, 1991)." In individualistic societies people tend to take care of themselves and direct family members only, instead of protecting people in exchange for unquestioning loyalty (Hofstede, 1991). In studying trust dynamics, Fulmer (2010) found the "black sheep" effect, in which collectivists became less trusting after experiencing violations from in-group rather than out-group members.

Uncertainty Avoidance (UA) is defined as "the extent to which the members of a culture feel threatened by uncertain or unknown situations (Hofstede, 1991)." People in greater UA cultures tend to shun ambiguous situations and look for structure in their organizations, institutions, and relationships, which makes events clearly interpretable and predictable.

Cultural syndromes

Cultural syndromes (Triandes 1996) provide an alternative way to characterize cultural differences.

Dignity Cultures are found in areas such as Western Europe and North America. In Dignity cultures the worth of an individual is not determined by the opinions and values of others and cannot be altered by other people, which is only evaluated by the individual who has his/her own standards (Leung & Cohen, 2011). Dignity cultures are associated with independence and focusing on personal and individual goals (Schwartz, 1992). People in Dignity cultures tend to make the "swift trust" assumption: others deserve to be trusted until they prove otherwise (Dirks, Lewicki, & Zaheer, 2009).

Face Cultures are widespread in East Asian because in these cultures self-worth is primarily interdependent with a person's role in a stable social hierarchy and on fulfillment of role obligations (Heine, 2001). In Face cultures, self-worth is stable and extrinsically derived based on social interactions with others, and what is important is the view that others have of you (Leung & Cohen, 2011). In cultures such as these, people can lose face if another person or group of people believes they have acted out, and other people can lose face because of your own views of their behavior (Leung & Cohen, 2011). Face cultures are high on collectivism with in-groups (Triandis, 1996), thus they would have high in-group trust, which could be related to the presence of stable institutions that foster cooperation.

Honor Cultures depend on the social interaction and therefore involve externally driven self-worth. Norms and values of Honor culture fall between Dignity and Face cultures, because Honor culture takes norms and values from both Dignity and Face culture to generate its own unique cultural prototype. Cultures like these can be found in the Middle East Latin America, and Mediterranean countries along with Southern United States. Honor cultures manifest with a reputation for toughness in protecting the self and family and involve not letting others take advantage of you (Cohen & Nisbett, 1997). The social context of Honor cultures is unstable social hierarchies. Members of Honor culture tend to have low interpersonal and institutional trust.

The defining characteristics of cultural syndromes have elements that are also present in Hofstede's dimensions. For instance, the elements that affect self self-worth are also found in definitional elements of individualism. Cultural syndromes therefore could bring relevant elements in addition to the Hofstede dimensions that may provide the basis for greater discriminatory power.

Cultural differences

As the hypotheses based on Hofstede's dimensions (1991) and a more recent theory of cultural syndromes (Leung &

Cohen, 2011; Triandis, 1996) suggest, it is reasonable to expect culture to affect trust and use of automation in a variety of ways. These cultural characteristics that have been identified as influencing inter-personal trust will guide our research in how cultural factors may influence trust and use of automation and help formulate the further studies.

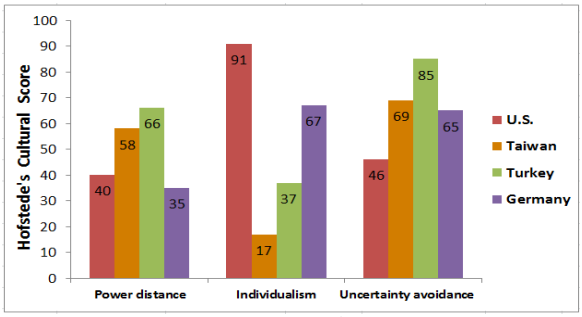


Figure1. Country comparisons in Hofstede cultural dimensions

The present research was conducted in US, Germany, Taiwan, and Turkey to provide representatives of each of the identified cultures. The selection of these countries was done based on two sets of criteria. The first criterion was the differences they exhibit in their values on the most well studied Hofstede cultural dimensions, as shown in figure 1. Second, each of these countries belongs to a different cultural syndrome; US and Germany (Dignity), Taiwan (Face) and Turkey (Honor). The above comparisons have provided decent supports for conducting the studies among these countries.

METHOD

Conducting cross-cultural studies is a complex problem. The first task is to find reliable and valid measures for a base population. Unfortunately, standardized tasks sensitive to trust effect(s), subjective measures of trust such as NASA-TLX (Hart & Staveland, 1988) for workload have yet to be developed. Extending measures to another culture must determine whether or not concepts are similar (typically defined by dimensionality and item loadings) and whether measurements are comparable (scale invariance). These two problems interact with dimensionality being dependent on item appropriateness. Therefore it will be reasonable to follow standard methods for adapting cross-cultural tests leading to one (or more) instruments of known psychometric properties. The following sections describe the scale development and cross cultures validation in greater detail.

Initial Data Collection

To refine the population of items, the initial phases of instrument development were conducted via Amazon Mechanical Turk (Mturk). After eliminating the unengaged responses (e.g., failed to answer the verifiable questions correctly), a total of 172 responses were collected, in which 77% of the participants are American.

Exploratory factor analysis (EFA) was used to determine the dimensionality of the data. To capture the effects of trust on various contexts, the purpose of the automation was categorized into two types, general and specific use of automation. The general attitudes toward automation largely involve predisposition to trust, whereas the attitudes invoked after human participants had been cued to think about particular instances of automation belong to the specific cluster. Four dimensions were included in the general group, performance, process, purpose, and task contexts, in which 32 items were selected, 8 items each construct. In the specific cluster, the involved items also related to the three systematic variables (performance, process and purpose), in which 18 items were retrieved, 6 items for each construct. To examine these retrieved items, reliability was measured with Cronbach's alpha (Venkatesh, Morris, & Davis, 2003) and the test of average variance extracted (AVE) was used to evaluate the item convergent validity (Bagozzi & Yi, 1988). The results revealed the items succeeded in both reliability and validity tests, table 1 Mturk column. Further details of the initial phases of scale development and refinement can be found (Chien, Lewis, Semnani-Azad, & Sycara, 2014a, 2014b).

Cross Country Scale Validation

To examine the instrument, another round of test was conducted in the U.S., Taiwan and Turkey. To avoid the language issues, before conducting the study, the instrument was translated into Chinese by two instructors in the Department of East Asian Languages and Literatures at the University of Pittsburgh; however, the English version was used for the Turkish population. Taiwanese participants were recruited from Chengchi University and Turkish participants were recruited from Ozyegin University. Students participated in the study to earn extra course credit. In order to increase the sample diversity, US army war college students were also recruited. A Smartphone was introduced as a context for considering the general use of automation; whereas GPS were referred to for the specific automation scales.

After eliminating the unengaged responses, a total of 311 student participant responses were collected and used to refine the proposed instrument. Reliability and validity tests were conducted to ensure the item consistency. As shown in table 1 (column United States, Taiwan and Turkey), the Cronbach's reliability alpha has the values higher than the threshold value of 0.7 in all dimensions. However, some of the constructs failed to pass the validity tests, such as purpose influence and task contexts.

Initial Application- Automated Driving System

This study was conducted in a Highly Automated Driving System (HADS), in which the HADS provided lateral and longitudinal control, including lane changes and overtaking. At the beginning of each experiment, participants were briefed on the driving simulator and were told that the system would not require monitoring during highly automated driving. Any necessary driver intervention would be announced with sufficient time to react by a take-over request (TOR) that combined sinusoidal sound and visual icon. TOR was used to simulate imperfect automation (i.e., automation failures) and measure participants' reliance on automation. Data were collected in single one-hour experiments and the virtual driving scenario was adopted for all sessions. A total of 66 participants BMW Group employees voluntarily participated in the study. The instrument was first translated into German by the BMW's translation department. After the experimental session, participants were asked to complete the posteriori Automation Trust Scale.

TABLE 1. Cross country scale rating comparisons (suggestive threshold values: Cronbachs α > 0.7 and AVE > 0.5)

General Auto (num of participants)	Mturk (172)		United States (100)		Taiwan (120)		Turkey (91)		Germany (66)		OVERALL (Mturk+US+ Taiwan+Turkey)	
	Cronbachs α	AVE	α	AVE	α	AVE	α	AVE	α	AVE	α	AVE
Performance expectancy	0.966	0.667	0.888	0.619	0.862	0.527	0.878	0.552	N/A		0.897	0.600
Process transparency	0.962	0.668	0.869	0.546	0.856	0.503	0.855	0.513			0.889	0.566
Purpose influence	0.946	0.585	0.844	0.500	0.777	0.409	0.850	0.491			0.849	0.488
Task contexts	0.948	0.567	0.704	0.432	0.743	0.415	0.800	0.440			0.759	0.425
Specific Auto	Mturk		United States		Taiwan		Turkey		Germany		OVERALL (Mturk+US+Taiwan +Turkey+ ***Germany***)	
	Cronbachs α	AVE	α	AVE	α	AVE	α	AVE	α	AVE	α	AVE
Performance expectancy	0.987	0.738	0.847	0.587	0.859	0.594	0.903	0.675	0.854	0.585	0.889	0.648
Process transparency	0.962	0.718	0.813	0.531	0.824	0.539	0.886	0.639	0.883	0.637	0.870	0.608
Purpose influence	0.979	0.664	0.809	0.516	0.840	0.560	0.887	0.642	0.871	0.614	0.864	0.596

RESULTS

MANOVA was used to examine the differences across scales for the different cultural datasets. Because the German participants took both general and specific scales of the trust survey after already experiencing the automated driving studies, in which several automation failures were injected judgments on the general items were contaminated by the recently experienced failures, The U.S. participants by contrast were asked to rate based on their general experience with smart phone a (general auto) or with a specifically cued navigation system (specific auto) leading to a clear distinction between predisposition and trust based on experience. Because of these distinctions only responses to use of specific automation (table 1) are included for the German sample. The analyses found significant cultural differences in all of the comparisons ($p<.001$), including the dimensions of performance, process, purpose and task contexts in general clusters as well as in the specific groups.

The data were reanalyzed to examine the purpose of the automation use (i.e., general vs. specific use of automation). As shown in figure 2a, the overall comparison included general and specific groups. The general construct included performance, process, purpose, and task contexts, whereas the specific cluster only included three systematical factors (performance, process and purpose). The analyses revealed significant cultural differences ($p<.001$) in all three comparisons- overall, general and specific use of automation. Pairwise T-tests revealed significant differences in most of the analyses, except the below comparisons:

Overall:

Mturk ≈ U.S. ($p=.31$); Taiwan ≈ Turkey ($p=.07$)

General:

Mturk ≈ U.S. ($p=.84$)

Specific:

Mturk ≈ Germany ($p=.26$); U.S. ≈ Germany ($p=.40$)

Taiwan ≈ Turkey ($p=.42$)

Additionally, items were also analyzed by its systematical constructs. For instance, the performance variable (figure 2b) included the items from the general and specific uses of performance, regardless of task contexts construct. MANOVA again showed significant cultural differences ($p<.001$) in all comparisons, performance, process, and purpose. Pairwise T-tests showed significant differences in most of the comparisons, excluding the below comparisons:

Performance:

Mturk ≈ U.S. ($p=.16$)

Process:

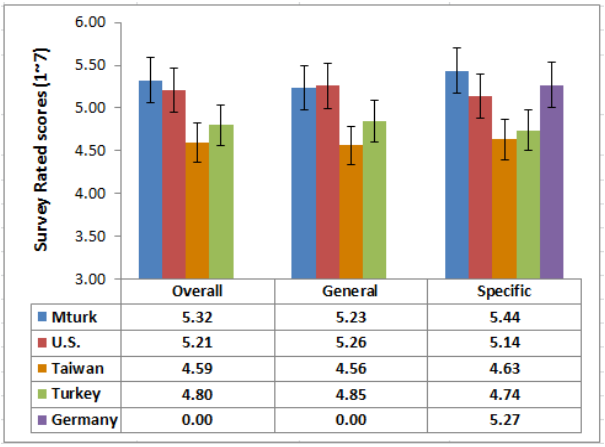
Taiwan ≈ Turkey ($p=.64$)

Purpose:

Mturk ≈ U.S. ($p=.68$)

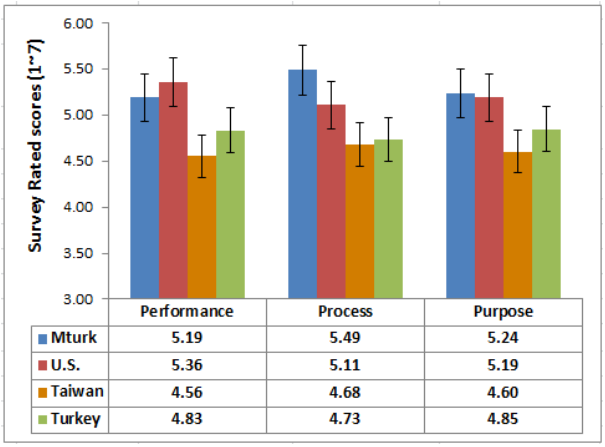
DISCUSSION

The connotations of trust in various system configurations and cultural dynamics may greatly affect an individual's (initial) trust attitude and resulting behaviors, even the strategies in relying on autonomous assistance. Researchers have been devoting considerable work on how trust relationships influence the social dynamics of automation uses as well as how trust in the systems impacts autonomous robustness (Chien, Lewis, Mehrotra, & Sycara, 2012, 2013). The present study provided a scale-invariant instrument to study both theoretically and most crucially empirically trust in automation, in order to develop a fundamental understanding of general principles and factors pertaining to trust in automation, and how trust mediates reliance on automation across cultures. Through various rounds of scale validations, the developed instrument showed great consistency (table 1) in performance and process in general use of automation as well as in the retrieved dimensions in specific use of automation across all the cultural validations, although the purpose and task contexts in the general cluster failed to fully satisfy the validity tests (table 1). These promising results confirmed that the developed instrument is capable to differentiate the noticeable cultural distinctions. According to the cultural syndrome, both Germany and the U.S. belong to the dignity cultures. The results confirmed this cultural phenomenon (figure 2a: specific cluster).



	Overall	General	Specific
Mturk	5.32	5.23	5.44
U.S.	5.21	5.26	5.14
Taiwan	4.59	4.56	4.63
Turkey	4.80	4.85	4.74
Germany	0.00	0.00	5.27

Figure2a. Grouped by *the purpose of automation uses*
General: 3 systematical dimensions + 1 task context
Specific: 3 systematical dimensions



	Performance	Process	Purpose
Mturk	5.19	5.49	5.24
U.S.	5.36	5.11	5.19
Taiwan	4.56	4.68	4.60
Turkey	4.83	4.73	4.85

Figure2b. Grouped by the *systematical constructs*
e.g., Performance: General + Specific performance dimensions

Figure2. Cross Country scale rating comparisons

As shown in the cultural syndrome, the member of Dignity culture has higher level of initial trust, which was again confirmed by our instrument (figure 2a), in which the U.S. group rated significantly higher in overall and general trust. Due to the large portion (77%) of the Mturk participants are American, the overall and general trust remained no statistical differences (figure 2a). The analysis showed marginal differences in overall trust (p=.07) and no statistical differences in specific trust (p=.42) between Taiwan and Turkey population, although the cultural characteristics were categorized into two groups, Face culture and Honor culture respectively. A recent study (Aslani, Ramirez-Marin, Semnani-Azad, Brett, & Tinsley, 2013) concluded that Face and Honor cultures are closely related, which can explain the non-significant statistical results.

A limitation is that most of the responses (except the German samples) were not examined by capturing the instantaneous trust through the empirical studies. However, the present study still provided a reliable instrument that can be used across cultures to measure trust and its antecedents. Next, we will conduct theoretically guided experimental studies in the U.S., Taiwan and Turkey to empirically determine how cultural factors affect various aspects of trust and reliance on automation.

ACKNOWLEDGMENT

This research has been sponsored by AFOSR FA9550-13-1-0129.

REFERENCE

Aslani, S., Ramirez-Marin, J., Semnani-Azad, Z., Brett, J. M., & Tinsley, C. (2013). 10. Dignity, Face, and Honor cultures: implications for negotiation and conflict management. *Handbook of research on negotiation*, 249.

Bagozzi, R. P., & Yi, Y. (1988). On the evaluation of structural equation models. Journal of the academy of marketing science, 16(1), 74-94.

Chien, S. Y., Lewis, M., Mehrotra, S., & Sycara, K. (2012). Effects of unreliable automation in scheduling operator attention for multi-robot control. *IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC 2012)*.

Chien, S. Y., Lewis, M., Mehrotra, S., & Sycara, K. (2013). Imperfect Automation in Scheduling Operator Attention on Control of Multi-Robots. *Human Factors and Ergonomics Society Annual Meeting*.

Chien, S. Y., Lewis, M., Semnani-Azad, Z., & Sycara, K. (2014). An Empirical Model of Cultural Factors on Trust in Automation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 58, No. 1, pp. 859-863). SAGE Publications.

Chien, S. Y., Semnani-Azad, Z., Lewis, M., & Sycara, K. (2014). Towards the development of an Inter-Cultural Scale to Measure Trust in Automation. *In Cross-Cultural Design* (pp. 35-46). Springer.

Cohen, D., & Nisbett, R. E. (1997). Field experiments examining the culture of honor: The role of institutions in perpetuating norms about violence.

Dirks, K., Lewicki, R., & Zaheer, A. (2009). Reparing Relationships Within and Between Organizations: Building A Conceptual Foundation. *Academy of Management Review*, *34*(1), 68–84.

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. *The Journal of the Human Factors and Ergonomics Society*.

Fulmer, C. A. (2010). *Dynamic Trust Processes after Violation: Trust Dissolution and Restoration* (Doctoral dissertation).

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload*.

Heine, S. (2001). Self as cultural product: An examination of East Asian and North American selves. *Journal of personality*, *69*(6), 881–906.

Hofstede, G. (1991) *Cultures and Organizations: Software of the Mind*, Maidenhead, UK: McGraw-Hill.

House, R., Javidan, M., Hanges, P., & Dorfman, P. (2002). Understanding cultures and implicit leadership theories across the globe: an introduction to project GLOBE. *Journal of world business*.

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243–1270.

Lee, J., & See, K. (2004). Trust in automation: designing for appropriate reliance. *Human factors*, *46*(1), 50–80.

Leung, A. K. Y., & Cohen, D. (2011). Within- and between-culture variation: Individual differences and the cultural logics of honor, face, and dignity cultures. *Journal of Personality and Social Psychology*.

Li, Dingjun, Rau, P. L., & Li, Y. (2010). A Cross-cultural Study: Effect of Robot Appearance and Task. *International Journal of Social Robotics*.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: an attentional integration. *The Journal of the Human Factors and Ergonomics Society*, *52*(3).

Triandis, H. (1996) The psychological measurement of cultural syndromes, *American Psychologist*, 51(4), 407-415

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view.

Wasti, S., Tan, H., Brower, H., & Onder, C. (2007). Cross-cultural measurement of supervisor trustworthiness: An assessment of measurement invariance across three cultures. *The Leadership Quarterly*.

Yamagishi, T., and Yamagishi M. (1994) Trust and commitment in the United States and Japan. Motivation and Emotion, 18, 129-166.