ARTICLE TEMPLATE

What Happens when Decision Support Systems Fail? — The Importance of Usability on Performance in Erroneous Systems

Philipp Brauner, Ralf Philipsen, André Calero Valdez and Martina Ziefle

Human-Computer-Interaction Center, RWTH Aachen University, Germany

ARTICLE HISTORY
Compiled February 6, 2019

ABSTRACT
With the advent of artificial intelligence (AI) methods, smart decision support systems are becoming ubiquitous. Such systems help reduce complexity for operators by automating data integration tasks and recommending actions. However, these systems are sometimes flawed. It is not sufficiently understood whether, when and why operators comply with such systems in erroneous or correct cases. We empirically investigate compliance with correct and defective decision support systems (DSS), the influence of correct and erroneous DSS's on performance and subjective factors related to compliance. In the study, a business game was used as an experimental setting in which 40 users took part. The impact of system correctness on user acceptance, trust, compliance and overall performance was investigated. The results show that the defective system reduces trust in automation (-47%), reduces usefulness (-58%), reduces acceptance (-62%) and reduces overall performance (-32%). Overall, the defective system was less user-friendly (-27%). Nevertheless, users who rated the system's usability higher, outperformed users who rated it lower. Usability is therefore an intermediary that compensates for the negative influence of erroneous decision support systems.

KEYWORDS
Decision Support System; Usability; Technology Acceptance; Trust in Automation; Automation Bias; Supply Chain Management; Business Simulation Game

1. Introduction

The 4$^{\text{th}}$ industrial revolution is reshaping manufacturing companies. Increased automation shifts the tasks and responsibilities of workers and employees from manual labor to increased controlling and managerial tasks (Bock 2015). In this article we investigate how employees interact with decision support systems, which factors govern compliance with these systems, and how they react to possibly misleading suggestions.

The emergence of cyber-physical production systems (CPPS) as the combination and tight collaboration of computational entities and physical production systems is reshaping production networks (Monostori 2014; Lee 2008). It will eventually lead to tighter integration of production processes, both within and across departments in companies, as well as across distribution networks. Further, it will integrate the whole life-cycle of a product, from design and production planning, over manufacturing, to the analysis of data acquired while using the product (Wollschlaeger et al. 2017).

CONTACT Philipp Brauner. Email: philipp.brauner@rwth-aachen.de

Although more and more planning, procurement (ensuring timely supply of resources), and manufacturing processes will eventually be automated, humans will remain necessary and essential in CPPS, at least because of two reasons: They will still be responsible for processes that cannot (yet) be automated for technical, financial, ethical, or legal reasons (Brettel et al. 2014; Davenport and Harris 2005). Further, they will have to handle exceptions of automated processes and they will be final arbitrators or referees if automated processes come into conflict (Mosier and Skitka 2018).

Such systems are built on the interaction between computers, production systems, and human operators and are consequently referred to as *socio-technical production systems* (STPS) (Frazzon et al. 2013). We assume that the performance of these systems is not solely determined by the *complexity* of the underlying system, but by the interaction between the complexity of the *cyber-physical production system*, the *interface*, and *human factors* (Brauner et al. 2016; Mosier and Skitka 2018; Brettel et al. 2014), as Figure 1 illustrates.

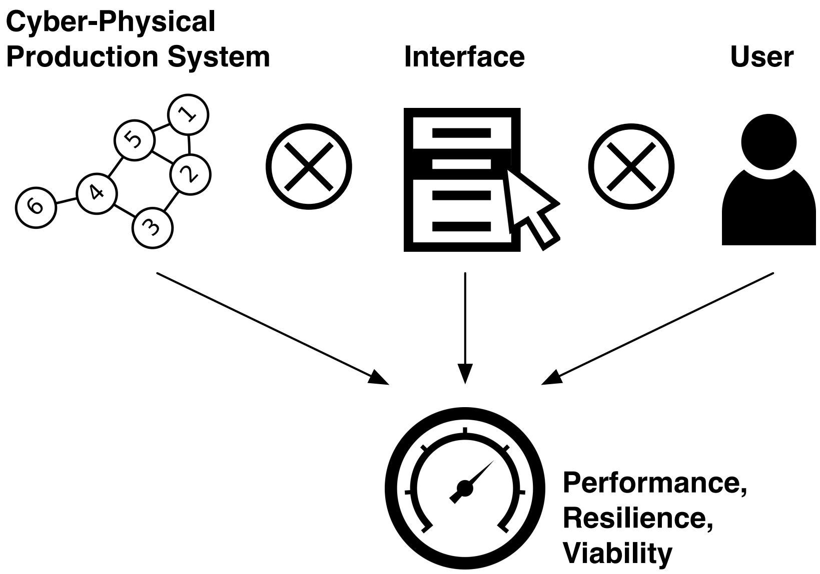

Figure 1. The three domains that influence the performance of socio-technical production systems.

One essential component to support operators in STPS are *decision support systems* (DSS), i.e., systems that automate the programmable part of an operational, tactical, or strategic decision problem and provide support for its users (Gorry and Morton 1971; Shim et al. 2002). However, research has shown that the way operators handle DSS is decisive: disuse, misuse, and obedience of such support systems can lead to lower performance, errors, and disastrous effects in terms of economic loss, efficiency, supply chain viability, or viability of companies (Lee and See 2004; Muir and Moray 1996; Ziemann et al. 2016; Mosier and Skitka 2018).

Consequently, we urgently need an increased understanding of people's interaction with decision support systems in cyber-physical production systems. An understanding of how correctness of a DSS shapes trust, compliance, and acceptance helps successful transformation towards digitalized production networks, human-oriented work conditions, as well as efficient and effective manufacturing processes (Parker and Sinclair 2001; Te'eni 1991).

But, how do humans interact with decision support in cyber-physical production systems? To understand which aspects shape both beneficial and harmful compliance with decision support systems, we experimentally investigate the influence of correct and defective decision support systems.

The remainder of this article is structured as follows: Section 2 presents related work from production network complexity, human factors in production networks, and business simulation games as the methodological foundation of this empirical work. Section 3 describes our research questions and the experimental procedure. Section 4 presents the findings of the study, and section 5 discusses the results and its implication in a broader context. Section 6 concludes this article by providing a comprehensive conclusion, limitations, and a research agenda for increasing the understanding of human interaction with decision support in cyber-physical production systems.

2. Related Work

Our study combines different concepts which are outlined the following sections. Complexity in socio-technical production systems often arises from both *supply chain disruptions* (see section 2.1) and *human factors* (see section 2.2). This complexity can be addressed by the use of decision support systems (see section 2.3). To understand the interactions between these three aspects, business simulation games can be utilized as an experimental setting (see section 2.4) to estimate the trust in, compliance with, and acceptance of such systems

2.1. *Challenges in Supply Chain Management*

First, we take a look at the contextual domain of our study: Supply chains are an integral component for manufacturing companies and cyber-physical production systems, however various threats endanger their functioning and stability.

Kleindorfer and Saad (2005) differentiated supply chain risks originating from *supply chain and demand coordination* and risks that stem from *variances and disruptions* from normal activities. Snyder et al. (2016) presented a systematic review of supply chain disruptions and found that various causes, such as unexpected demand spikes, industrial accidents, strikes, terror attacks, or natural disasters, can trigger these disruptions.

The most prominent example of a supply chain disruption is the *bullwhip effect* first described over 50 years ago (Forrester 1961). Small demand spikes at the end of a supply chain (i.e., at retailers), in combination with time delays between order and delivery, and limited communication across the supply chain can accumulate upwards the supply chain and are amplified on each level. Although this effect has been investigated well and has been part of managerial training courses for a considerable time, it is still pressing today (Lee et al. 1997; Snyder et al. 2016; Tako and Robinson 2012; Brauner et al. 2016) (see also section 2.4 and Figure 2).

In addition to the demand-related bullwhip effect, supply chain performance is prone to a variety of disruptions. These occur when the supply chain structure is sub-optimal, when the transportation of goods is disrupted (Wilson 2007), or when other uncertainties of the environment continuously accumulate (Wong et al. 2011).

The majority of methods for mitigating supply chain disruptions address technical or organizational aspects of the production network (Tang 2006), such as compatibility with fall back suppliers, strategic stocks and safety inventories to compensate fluctuations, or changing prices to shift the demand to products that are less affected by a disruption. However, these methods rarely focus on the interaction of workers with the decision support system.

2.2. *Human-Factors and Complexity in CPPS*

In this section, we show how differently humans' behaviors influence cyber-physical production systems. Research on how human factors influence performance, stability, and resilience of supply chains is scarce and not canonized, probably due to the large variety of production networks, which differ based on the size of the manufacturing companies, organizational structures, and product requirements (produced at scale vs. scope, shelf life, requirements for quality, etc.).

In the context of supply chain disruptions Blackhurst et al. (2005) identified the main causes of disruptions and three strategies to prevent them: Disruption discovery, disruption recovery, and supply chain redesign. Although disruptions are often predictable from data that is already available, the operators' ability to correctly interpret the enormous amount of information available is limited. The operators might therefore be incapable of detecting upcoming disruptions. The authors suggest an automated supply chain intelligence that predicts and visualizes potential supply chain disruptions and triggers human intervention if difficulties are foreseeable.

Kanda et al. (2009) investigated how human factors influence agile supply chains. The main factors were the ability to coordinate a supply chain, trust between buyer and supplier, the flexibility to cope with changes and resistance to change, as well as work culture and motivation of the employees.

Enterprise resource planning systems (ERP) are operators' interface to the underlying cyber-physical production system. They convey information about the underlying system, offer opportunities to control the system, and usually offer decision support through integrated decision support systems (e.g., procurement of new material if the inventory is low). However, human interaction with these systems has not systematically been investigated yet and literature on this topic is relatively sparse. For managerial decisions in the supply chain context, Mittelstädt et al. (2015) investigated the influence of task complexity, interface usability, and operators' cognitive abilities on correctness and speed in a decision task. Besides the main effect that higher task complexity leads to more errors and lower performance, a key finding is that decision complexity interacts with user interface design: Poor usability does no harm for easier tasks, but has disastrous effects for more complex decision tasks. Furthermore, the study found that perceptual speed is related to decision performance: people with higher perceptual speed were more likely to compensate negative effects of low usability and high task complexity.

Ziefle et al. (2015) conducted a study to understand the relationship between task complexity (modeled as the amount of information that has to be processed at a time), individual user factors (perceptual speed), and performance (effectiveness and speed). Higher complexity was linked to lower task speed and people with higher perceptual speed could compensate the negative influence of growing complexity better than people with lower perceptual speed.

In a follow up-study Brauner et al. (2016) investigated the effect of DSS in handling material disposition decision tasks and analyzed if these effects differed between correct and defect systems. A correct support system had a positive influence on both accuracy and decision speed compared to a baseline condition with no support system. On average, the effect of the defective system on accuracy and decision speed was negligible. Although the decision speed was only mildly affected by the defective DSS (compared to the baseline), their accuracy was devastated, effectively doubling their error rate: Despite realizing the DSS's defectiveness, the participants followed the system's suggestions. The devastating effect grew with task complexity: Participants were able to compensate the DSS's defectiveness for simpler tasks, whereas accuracy plummeted for complex tasks. Concluding, the study showed that defective systems are obeyed and obedience with these systems increases when tasks become more complex.

2.3. *Automated Systems, Decision Support Systems and Human Factors*

Next, we take a look at how intelligent systems have been used to help with the problems in our domain and what novel problems were introduced due to such systems. Parasuraman et al. (2000) introduced a model for types and levels of automation. In this model, automation can support users by augmenting human cognition in one of the four stages *information processing*, *perception*, *decision making*, and *response selection*. The suggested ten levels of automation range from no automation (i.e., complete human control, no technological support), to full automation (i.e., no human intervention possible).

The difficulties that arise from automation were first described in Bainbridge's (1983) article "ironies of automation". The identified key problem was that if processes are automated, operators' control skills deteriorated due to a lack of practice. Consequently, if automation fails, the operators will have difficulties to detect the failure *and* will have difficulties to intervene manually.

Most research focuses on the benefits of correctly working decision support systems. But what happens, if the systems do not work as intended and what shapes the use of automation?

Parasuraman and Riley (1997) defined the terms *use*, *misuse*, *disuse*, and *abuse* of automation. *Use* refers to the voluntary and appropriate usage or non-usage of automation. An appropriate use of automation improves interacting with cyber-psychical systems; it can reduce mental workload, and increase performance. *Misuse* signifies over-reliance on automation, which may lead to failing to monitor the automated system and thus unintentional obedience of automation. Consequently, this may result in higher error rates and lower performance. *Disuse* refers to the deliberate underutilization or even disuse of automation, which also may lead to lower performance, higher error rates, or higher cognitive load. On the other side, *automation abuse* refers to inappropriate implementation of automation by designers of support systems and managers that does not consider the capabilities, wants, and needs of the users of the automated system. Automation abuse may then lead to automation misuse or disuse.

Automation misuse is closely related to the concepts of automation biases and automation complacency (Goddard et al. 2012; Parasuraman and Manzey 2010). Though both concepts are connected, automation biases refer to peoples' tendency to trust and follow suggestions of an automated decision support system and to evaluate the suggested decisions as rather positively than neutrally. Whereas automation complacency refers to the perceived reliability of the system, which leads to lower attention in the monitoring of the underlying system and its automation and thus errors tend to remain undetected.

Reeves and Nass (1996) showed that people unconsciously react to computer interfaces as they would to other humans and that most effects and biases from social psychology also hold true for human-computer interaction. Therefore, depending on several factors, people attribute trust and credibility to computing systems, which then shape their interaction with these systems (Fogg and Tseng 1999). Consequently, the trust that operators attribute towards automation is inherently linked to use, misuse, and disuse of automation (Muir and Moray 1996; Lee and See 2004).

Appropriate use of automation relies on a calibration of trust in and the capabilities of automation (Lee and See 2004). If the trust exceeds the capabilities of the system, this leads to *overtrust* and a misuse of the system. If the trust falls below the capabilities, this leads to *distrust*, and consequently, disuse. Thus, trust and system capabilities should be carefully balanced. Trust, as an decisive mediator for interpersonal communication, deserves special attention in research on social-cyber-physical production systems.

Muir and Moray (1996) have shown for a machinery control task that trust that is attributed to automated processes is mainly explained by the perceived competence of the automation. Also, trust decreases if the perceived competence of the automation decreases, even if the actual influence on performance is limited. Further findings are that decreasing trust in one automated function of a component influences trust in other automated functions of the same component, but does not carry over to different components.

De Vries et al. (2003) investigated how error rates in automated and manual processes influence trust in automation and self-confidence. Higher automation error rates lead to decreased trust in the system compared to lower automation rates. Likewise, higher manual error rates were linked to lower self-confidence. Trust plays also an important role in interaction with an ERP: Mayeh et al. (2016) found that trust shaped the perceived usefulness of the system, which then was found to be the strongest predictor for technology acceptance of the ERP system.

For the case of an automated route planner, Pak et al. (2017) showed that decision making performance was influenced by correctness of the support system. Based on the assumption that compliance relates to working memory capacity, they analyzed if working memory interacts with correctness and found that people with higher working memory performed better than users with lower working memory for the correct system, but that both groups performed similar for the case of the defective system.

Concluding, operators may decide rationally, reasonably, and correctly when they have enough time and sufficient cognitive resources available to evaluate the state of the system and the quality of the suggestions from the automated support system. Under such optimal working conditions, operators are able to decide if they trust the system and follow recommendations, or rather, disregard its recommendations. However, real work settings are not optimal. Rather, operators are often inexperienced, distracted, or rushed and hence easily deflected by cognitive biases or misguiding suggestions (Gilovich et al. 2002). Under such sub-optimal working conditions, decision errors occur with negative consequences for the cyber-physical production system.

2.4. *Business Simulations and Business Simulation Games*

The following section introduces a methodological approach to studying complex systems with user interaction: business simulation games. Business simulations and business simulation games are an established method not only for conveying knowledge to learners, but also to understand how people interact with underlying business models Zyda (2005); Deshpande and Huang (2011); Brauner and Ziefle (2016). In contrast to field studies in companies, they are sufficiently complex and allow systematic manipulation of user, interface, and system factors to study their influence on relevant (game) metrics, such as production efficiency or the attained overall profit.

A prominent example is the *Beer Distribution Game* (BDG) developed by MIT's System Dynamics Group (Sterman 1989): Four players are part of a linear, multi-echelon, market driven supply chain. In the round-based game players exchange order information and materials along the linear supply chain and orders and deliveries are delayed and only possible between direct neighbors (Figure 2). Variances in the customer's orders lead to exaggerated orders for each tier of the supply chain, resulting in the *bullwhip effect*, see section 2.1.

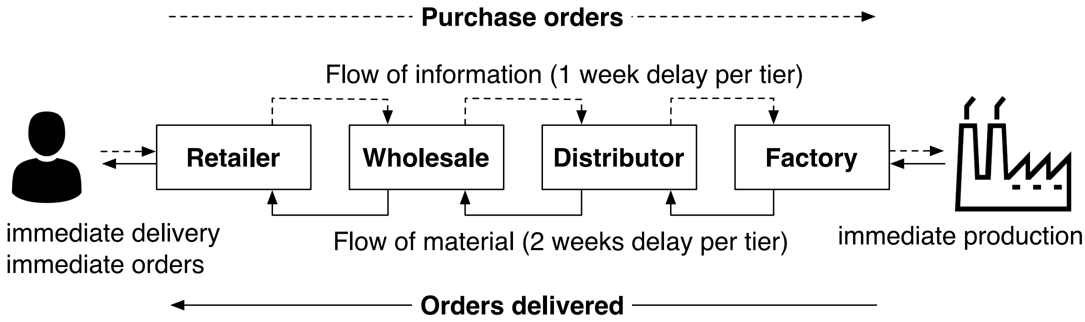

Figure 2. Illustration of Forrester's Beer Distribution Game. Four tiers of a linear production network exchange purchase information and materials with a delay and only by communicating with neighbors.

On the basis of the Beer Distribution game Sarkar and Kumar (2015) investigated the effect of upstream (i.e., from the retailer) and downstream (i.e., from the supplier) disruptions using a behavioral study in a business simulation game. They found that sharing information leads to lower variances and lower overall costs for upstream events (i.e., disruptions at the manufacturer), whereas the effect for downstream disruptions (i.e., at the retailer) was rather limited.

Ben-Zvi (2010, 2012) investigated how perceived effectiveness of decision support systems affects performance in a business simulation. The perceived effectiveness of the DSS correlated with the overall company performance. However, the study also revealed that some of the developed DSS were not effective, despite significant development efforts. Goldratt and Cox (1992) further introduced variance to a business simulation game with multiple tiers.

Based on the beer distribution game Stiller et al. (2014) developed a supply chain and quality management game with increased task complexity. Players are part of a multi-tier production network and responsible for minimizing costs for warehousing while ensuring sufficient supplies. The players need to infer the current state of production from 20 indicators, including redundant or unnecessary values and need to manage different controls for the investments. This model has been used to measure domain expertise (Philipsen et al. 2014), to empirically quantify the influence of user interface refinements (Philipsen et al. 2014), and as a method to convey supply chain and quality management expertise in higher education (Brauner et al. 2016).

2.5. *Compliance and Technology Acceptance*

Now, we look at technology acceptance and compliance, which both can be used to understand the behavior in our business simulation game. Technology acceptance research aims at predicting the adoption of products or services and to understand the underlying personal and system factors that govern this adoption process.

The most influential model is Davis' *technology acceptance model* (TAM) (Davis 1989) that shows that the *perceived usefulness*, the *perceived ease of use*, as well as the *attitude towards using* the software govern the *intention to use* and later *actual use* of software in business contexts. Consequently, if the (perceived) usefulness or ease of use is increased, the postulated relationship will positively influence the intention to use the technology and their later actual use. As the *intention to use* and the later actual *use* are strongly correlated, the later use of the system can be predicted by studying the factors that predict the intention to use.

The TAM model was refined in further iterations and adapted to different professional usage scenarios (Venkatesh and Davis 2000), personal and voluntary use of consumer technology (Venkatesh et al. 2012), as well as to serious games and business simulation games (Yusoff et al. 2010; Brauner et al. 2016) by adding additional predictors, such as price-value trade-offs, the hedonic evaluation of the technology, or the transferability of learned skills and the users' control over the learning process.

Despite the ongoing evolution of technology acceptance models and the meandering integration of new predictors, TAM and its key predictors (perceived) *usefulness* and *ease of use* are still frequently used and highly predictive.

2.6. *Research Gaps, Goal, and Research Approach*

Decision support systems help to mitigate disruptions in production systems and to manage the growing complexity of globally dispersed and increasingly interconnected supply chains. However, the *research gap* is that the influence of automation complacency and automation biases caused by correct and erroneous support systems in these contexts is insufficiently explored. Overall, both acceptance and performance must be evaluated in conjunction to fully understand the interaction of decision support and human operator in complex supply chains.

The *goal* of the study is to empirically investigate the compliance and performance with correct and defective decision support systems (DSS) and how these relate to subjective factors.

Such questions are not easily investigated in laboratory settings, yet observational studies lack the ability to purposefully vary individual factors. Thus, we have to find a trade-off between experimental necessities—such as minimal amount of factors, their valid operationalization, and the analysis of their interactions—as well as the complexity of real-life applications.

As *research approach* we designed an experimental framework around a business simulation game developed with domain experts. Using this method we can simulate complex decision making scenarios, experimentally control the DSS's correctness, and to investigate the effects we are interested in. Methodologically speaking, we can increase the variance in complexity between trials as a within-subject factor, while retaining the between-subject variance that might be high as well. In our experiment, we combined business simulation games and technology acceptance research to study if, why, and by whom correct and defective systems in manufacturing contexts are used.

3. Method

The following sections outline the methodology of our study. First, we present our research hypotheses that guided our experiment. Next, we explain the methodological framework and business simulation game used in the study. Then we report the experimental variables, the procedure, and the sample of our study.

3.1. *Hypotheses*

The following hypotheses guided our study and the subsequent analysis.

- H1: *The correctness of a decision support system influences trust in automation.* When errors are perceived and noticed, trust in the automated support should also decrease.
- H2: *The correctness of the systems relates to perceived usefulness and perceived ease of use of an automated system.* Only when a system is actually helping the user to make good decisions, it is considered to be useful. Furthermore, it is perceived as easy to use, when the system leads to a decrease in user input actions, meaning that it works correctly.
- H3: *The correctness influences the intention to use the system and the actual use of the system.* Only when the system operates correctly, will users reach the conclusion to become a future adopter. A defective system should lead users to the conclusion to not adopt the automation.
- H4: *User diversity factors (gender, self-efficacy, and trustfulness) influence the compliance with a support system.* Depending on how trusting a user is, recommendations by the systems should be followed more “blindly”. Similarly, a person that perceives themselves to have high self-efficacy, should be able to realize that errors in technology are not the users fault and thus be able to detect an erroneous system more quickly. Further we think, that women will take longer to determine whether a system is faulty. The reasoning here is, that on average women tend to show higher values of neuroticism. This in turn should make them more vulnerable to the wrong assumption, that they were responsible for mistakes that are made.

3.2. *Experimental Simulation Framework*

As a basis for this experiment we used the “Quality Intelligence” business simulation game that combines the Beer Distribution Game (Sterman 1989; Wu and Katok 2006; Sarkar and Kumar 2015) (see section 2.4) with aspects from quality management and variances in production from Goldratt’s Game (1992). As such, it addresses two relevant aspects in the inter-company flow of materials and information.

The player is part of a market-driven supply chain and has to balance the investments between procurement, inspection of incoming goods, inspection of production quality, costs for stock keeping. All this while keeping in mind the profits gained by selling the products. The players have to observe 21 (partly redundant) variables for several months (each turn in the game represents one month). The must infer the current state of the production and control the investments on three target variables (procurement, inspection of incoming goods, inspection of production quality). Figure 3 shows the interface of the game and Figure 4 illustrates the simulated relationships. An in-depth presentation and explanation of the simulation model is given in (Stiller et al. 2014).

3.2.1. Within-subject variable: Correctness of the DSS

The user’s decision making is supported by a support system presented in the game’s user interface (see Fig. 3). It focuses on material disposition and recommends the number of supplies that should be ordered. The support systems resides at the higher end on Parasuraman, Sheridan, and Wickens’ levels of automation scale (as the suggestion is entered in the system and the operator can override the suggestion) and addresses the “decision making” stage (Parasuraman et al. 2000). Both other tasks (incoming goods

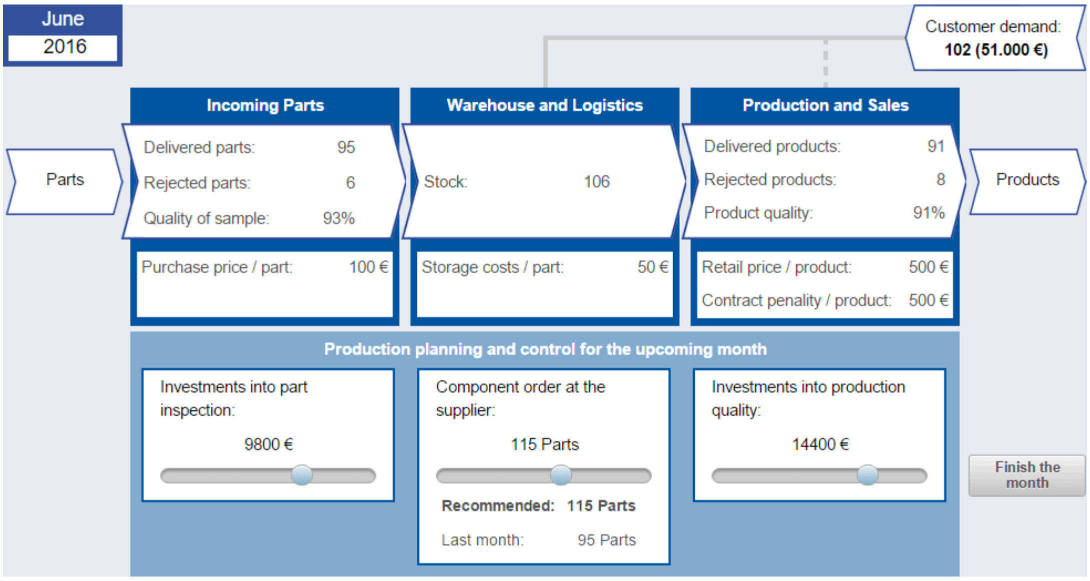

Figure 3. Interface of the game. Detailed presentation in (Stiller et al. 2014)

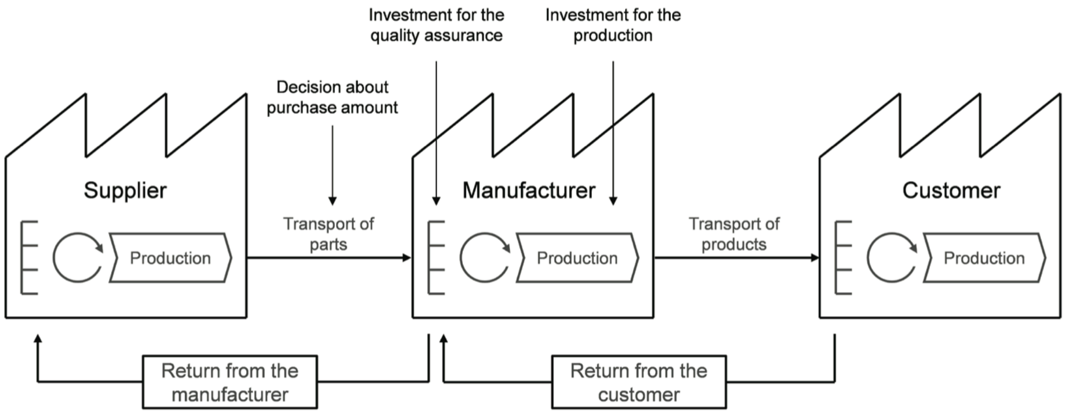

Figure 4. Schematic representation of the game's underlying simulation model. Detailed presentation in (Stiller et al. 2014)

inspection, production quality) are not assisted by the system.

The CORRECTNESS of the decision support system is varied as a within-subject variable across two rounds of the game: The DSS is either helpful in the first round and then leads the user astray in the second round, or the other way around. The order (*correct* $\rightarrow$ *defective* vs. *defective* $\rightarrow$ *correct*) is randomized and evenly distributed across the participants. As we want to study how users react to unforeseen defects, we do not inform the participants if a defect will or won't occur in a given round. Instead, the defect must be detected. Either directly from the recommendation, as the suggestion is much lower than the customer's orders, or indirectly, as the customer complaints and penalty costs will increase dramatically in the subsequent turns.

In the case of the *correct* system, the suggested number is near the theoretical optimum. Only very experienced players may find marginally better order levels. For the *defective* support system, the suggestions are correct for the first six months of the game and then get defective, yielding suggestions that are 50% below the correct recommendation. The defect occurs after six months because we wanted to investigate how people deal with a system that breaks down (as opposed to a system that is constantly broken). Consequently, the system lulls the participants into safety during the first six months.

3.2.2. Explanatory variables

To understand the influence of personality states and traits we captured the persons' age, gender, educational level, self-efficacy in interacting with technology, and trust in automation.

Age is measured as a numeric value and respondents reporting values below 18 or over 99 were excluded from the data. Gender was measured on two nominal levels—male and female.

Self-efficacy in interacting with technology (SET) is an instance of Bandura's domain-specific self-efficacy (1982) and measures a persons' perceived ability to successfully solve complex technical problems. This construct relates to a person's technology usage and is a strong predictor for performance, effectiveness, and satisfaction in interacting with interactive systems (Arning and Ziefle 2007). It is measured with eight items on a scale developed by Beier (1999) and achieves an excellent internal reliability ($\alpha = .872$ [.777, .913]).

Trust in Automation is measured with twelve items on a scale by Jian et al. (2000). It was applied three times during the experiment: Firstly, before the experiment to assess the participants' trust before using the game. Secondly, after the first round of the game. Thirdly, after the second round of the game. To measure trust in automation the first time and *before* the game, we used a scenario related to the supply chain context: A planning system that suggests the number of beverages to buy for a large event. The scale's internal reliability is high ($\alpha = .852$ [.774, .912]).

3.2.3. Dependent variables

A series of dependent variables is captured during each round of the game (metric from the game engine) and measured after each round of the game (subjective evaluations of the participants).

Following Goldratt and Cox (1992), we calculate the overall cumulated company profit as the overall performance metric. Consequently, the participants are instructed to play towards maximizing the company's profit.

In addition, we surveyed the participants' satisfaction with their performance and their perceived relative performance ("How well did you play?") compared to other players of the game. This was done without them actually knowing how others performed. The latter variable thus measures how much players think that they outperformed other players. If this measure is, on average, above 50%, the users overestimate their performance.

To understand the influence of the support system on compliance, we measured both, the participants' reported compliance ("*How often did you follow the suggestions of the support system? [in %]*"), as well as the objectively assessed compliance through the number of order changes in the business game. In other words, how often the suggested value of the support system was actually adjusted by the players.

Following Davis (1989) the participants' evaluation of the decision support system is assessed by three variables: their overall intention to use the system (see section 2.5), ease of using, and the perceived usefulness of the system (scale from 0 to 100).

Trust in Automation is measured after each round of game, again using twelve items by Jian et al. (2000) (scale from 0 to 100).

3.3. *Experimental Procedure*

The participants played two rounds of the game (see section 3.2). Each round of the game consisted of 18 turns (i.e., 18 months in the simulated company). To avoid effects of end game behaviors (Selten and Stoecker 1986), we removed the last three turns of the games for the analysis of behavior and overall profit. Otherwise some players might risk the viability of the company by clearing the warehouse and omitting orders in the last turns to maximize their profit.

Three questionnaires were administered during the study: The first captured the participants demographic data and independent variables *before* the first round of the game. The second questionnaire captured the participants evaluation of the first round. The third and final questionnaire captured the evaluation of the last round. Figure 5 illustrates the experimental procedure of this empirical user study. The administered questions can be found in Appendix 6 and the dataset of this study is publicly available (Brauner et al. 2018).

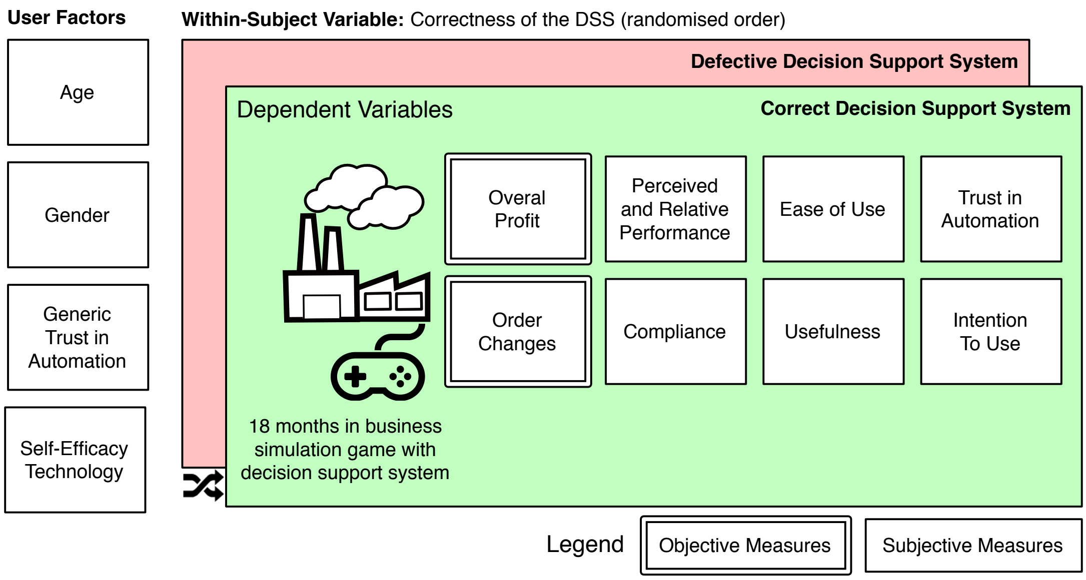

Figure 5. Illustration of the experimental procedure.

3.4. *Statistical Analyses*

The results were analyzed with parametric and non-parametric methods, using bivariate correlations (Pearson's r or Spearman's ρ), Wilcoxon tests, single, repeated multi- and univariate analyses of variance (M/ANOVA). Pillai's value V is considered for the multivariate tests. Effect sizes, as quantitative measures of the magnitude of an effect, are reported as η^2 ranging from 0 (no effect) to 1 (perfect explanation) (Kelley and Preacher 2012). If the assumption of sphericity is not met, Greenhouse-Geisser–corrected values are used, but uncorrected *df*s are reported for better legibility.

Following Cumming (2014) we report the 95% confidence interval (CI) for all statistical parameter estimates in square brackets. The error bars in diagrams represent the 95% CI. Medians are marked with *Md* and arithmetic means with M. In addition, we check for statistical significance using a level of $\alpha = .05$. Due to the comparably low sample size, we also report findings $.05 < p < .1$ as suggestive of statistical significance. As the performance from the simulation model is not normally distributed ($KS - Z_{round1} = 1.946$, $KS - Z_{round2} = 2.054$, $p < .001$) analyses of this model are performed with non-parametric Mann-Whitney U tests (MW-U).

3.5. *Sample*

The study took approximately 25 to 35 minutes to complete and participation was voluntary and not rewarded. The link to the web-based study was distributed via suitable message boards, email, and personal social networks.

Due to the voluntary participation in an online survey, we have a rather high drop-out rate: The study was started by 140 participants, but we will only consider the 40 dataset that completed all surveys and both rounds of the game (28.5%)[1]. Of course, we discuss the validity of our findings despite the drop-out rate in section 5.

From the 40 participants, 23 were male, 17 were female and the age range is between 20 and 56 years ($M = 28.5$, [25.9, 31.0]). Besides age and self-efficacy in interacting with technology ($\rho = .365$, [.061, .607], $p = .017$) all investigated user factors were uncorrelated (see Table 6). On average, the reported Self-efficacy in interacting with technology was clearly above the mean of the scale ($M = 70$, [63%, 76%]), whereas the initial Trust in Automation was near the center of the scale ($M = 54$, [49%, 60%]).

19 (47.5%) participants started with the correct DSS in the first round and finished with the defect DSS in the second round, whereas 21 participants had the opposite order (52.5%). Sequence effects can be excluded, as the order was unrelated to the user factors and order had no effect on the drop-out rate ($\chi^2 = .054$, $p = .817$).

4. Results

In a first step we analyze if performance measures—data generated from the simulation—and subjective measures—the perceived evaluations—agree with each other. Then, the influence of the correctness of the DSS on performance, compliance, and trust are evaluated. The following section 4.5 evaluates the influence of correctness on technology acceptance in general and the relationships for correct and defective systems.

4.1. *General Observations*

Firstly, we show that the measures from the underlying simulation and the reported answers from the participants are consistent. For that we look at the metrics for performance, as well as compliance with the DSS.

The average attained profit in the first round was $-3,043$ [-11,738, 5,652] ($Md = 12,275$) compared to $2,917$ [-4,367, 10,201] ($Md = 11,650$) for the second round of

[1] 140 participants followed the link to the study, 95 completed the whole first questionnaire, 54 participants finished the first round of the game, and 40 participants completed the second round and the final survey. Participants with a higher self-efficacy were slightly more likely to complete both rounds of the game ($\rho = .210$, [.009, .394], $p = .039$).

the game. Users' satisfaction with the performance for the first round was 34 [23, 45], compared to 61 [51, 71] for the second round. The perception of one own's performance compared to other players was, on average, in the 29 [22, 37] percentile in the first round and in the 53 [46, 61] percentile in the second round. Users believed they became better than other players at the game in the second round. For the the first round of the game, the actually attained PROFIT and the reported SATISFACTION with the performance ($\rho = .643$ [.415, .795], $p < .001$), as well as actual PROFIT and the perceived RELATIVE PERFORMANCE compared to other players (without knowing their true performance) are strongly related ($\rho = .470$ [.186, .681], $p < .001$). Also, RELATIVE PERFORMANCE and performance SATISFACTION are interlinked ($\rho = .745$ [.565, .857], $p < .001$). Thus, participants are satisfied, if they feel that they play the game well.

For the second round, PROFIT is strongly linked to RELATIVE PERFORMANCE ($\rho = .557$ [.298, .740], $p < .001$), but not to performance SATISFACTION ($\rho = .219$ [-.099, .496], $p = .181$). Still, RELATIVE PERFORMANCE is linked to SATISFACTION ($\rho = .530$ [.262, .722], $p < .001$). Players who make more profit, feel that they did a good job in comparison to others, however the profit did no longer make them satisfied with their performance.

Next, we found a strong *negative* relationship between the number of ORDER CHANGES from the game and the reported COMPLIANCE from the participant for the first ($\rho = -.805$ [.659, .892], $p < .001$) and second round of the game ($\rho = -.711$ [-.837, -.514], $p < .001$). This is expected, as the relationship is inverse, because participants complying *more* with the system have to make *less* changes to the suggestion by the DSS. The average number of ORDER CHANGES does not change between the first (12.0 [10.2, 13.8], $Md = 14$) and the second round of the game (12.0 [10.2, 13.7], $Md = 14$). Yet, the average REPORTED COMPLIANCE was at 34 [25, 44] in the first round compared to 44 [34, 54] for the second round.

Furthermore, measured company PROFIT ($\rho = .751$ [.574, .861], $p < .001$) and RELATIVE PERFORMANCE ($\rho = .487$ [.207, .693], $p < .001$) for the first and second round were related, indicating stable objective and perceived performances across both rounds. Players who played well during the first round, played well in the second round. However, performance SATISFACTION appears to be rather unstable ($\rho = -.038$ [-.276, .345], $p = .816$) , indicating an effect of DSS correctness.

Summarizing, the system metrics and the participants' perceived evaluations are consistent.

4.2. *System Correctness and Performance*

Next, we show that the decision support system's correctness has an actual effect on perceived and objective performance metrics.

In the first round of the game, the attained performance is significantly higher for the correct DSS ($Md = 13350$) compared to the defective DSS ($Md = 10350$) ($MW - U = 508.5$, $p = .027$). Similar effects emerge in the second round of the game (with correctness of the DSS switched): Players with correct DSS achieve higher profits ($Md = 12300$) compared to players with the defective DSS ($Md = 11650$). This effect is suggestive of statistical significance ($MW - U = 309.0$, $p = .056$). Table 1 presents the influence of correctness on performance.

This also significantly influences the players' satisfaction with their own performance for the first round of the game ($F_{1,38} = 5.009$, $p = .031$). The SATISFACTION of participants with the defective DSS was 46% lower compared to the players with the correct system. Strikingly—and different from most other findings of this article—the significant difference fades for the second round of the game ($F_{1,38} = 2.136$, $p = .152$) and the difference decreases to 19%.

A similar pattern emerges for the perceived RELATIVE PERFORMANCE. In the first round of the game, the defective system reduces the RELATIVE PERFORMANCE significantly by 38% ($F_{1,38} = 4.110$, $p = .050$), whereas the reduction is only 13% and not significant in the second round ($F_{1,38} = 1.178$, $p = .285$). Again, Table 1 shows the details of this effect.

In summary whether the support system behaves correctly influences how users perform and how they perceive their performance.

4.3. *System Correctness and Compliance*

Now we show that correctness influences the reported and actual compliance with the system: CORRECTNESS had no significant effect on the reported COMPLIANCE in the first round of the game ($F_{1,36} = .008$, $p = .928$, $\eta^2 < .001$), but in the second round ($F_{1,33} = 4.955$, $p = .033$, $\eta^2 = .131$). A similar pattern emerged for the ORDER CHANGES (i.e., measured compliance), which was not significantly influenced by CORRECTNESS in the first (MW-U $Z = -.477$, $p = .633$), but in the second round of the game (MW-U $Z = -2.245$, $p = .025$).

In the first round, the players with a defective system reported a 35.1% [21.4%, 48.7%] compliance with the system in contrast to a correct system with 34.2% [18.8%, 49.5%] compliance. In the second round, players with the defective system reported 31.4% [15.6%, 47.2%] compliance, compared to 52.6% [18.8%, 49.5%] compliance for the correct system. Likewise, the measured number of order changes in the second round of the game was lower for the correct ($Md = 8.5$, $M = 9.8$ [7.1, 12.4]) than for the defective system ($Md = 15$, $M = 14.3$ [12.2, 16.5] order changes). Table 1 and Figure 6 summarizes these findings.

Summarizing, the correctness of the support system has an effect on measured and reported compliance, although the effect only emerges in the second round.

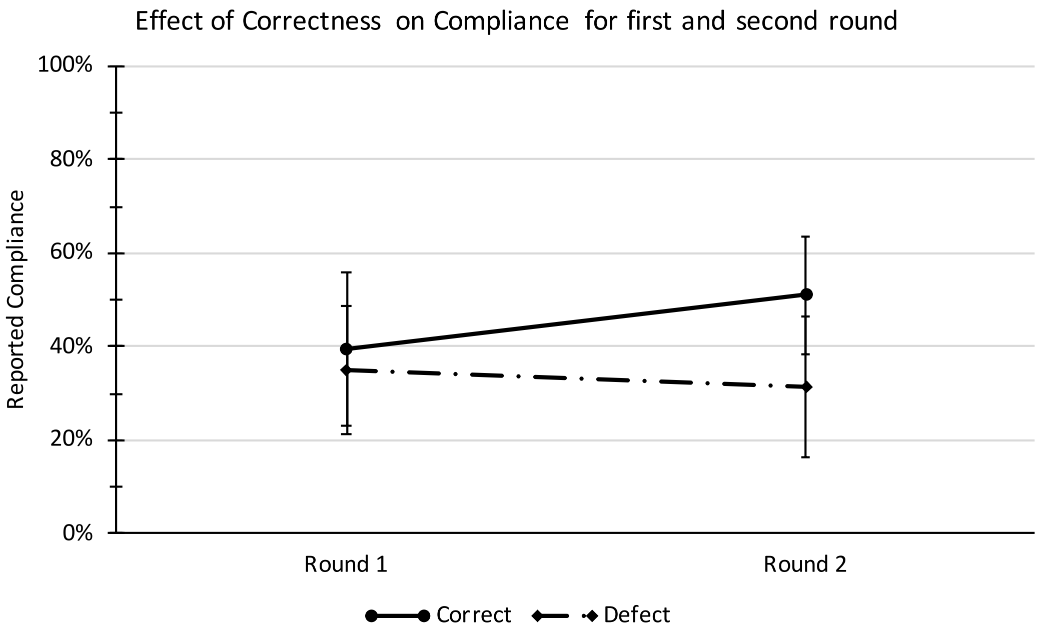

Figure 6. Effect of CORRECTNESS and ROUND on reported COMPLIANCE. If the systems behaves correctly, compliance increases. Error bars indicate the 95%-CI.

		Correct		Defect	
		Round 1	Round 2	Round 1	Round 2
Overall Profit	Median	13350	12300	10350	11650
	Mean	6589 [-498, 13677]	7605 [2650, 12560]	-10957 [-24891, 2976]	-703 [-13962, 12557]
Overall Satisfaction	Mean	47 [32, 63]	67 [55, 78]	26 [12, 39]	54 [39, 69]
Relative Performance	Mean	37 [27, 47]	57 [49, 65]	23 [13, 33]	49 [36, 62]
Number of Order Changes	Median	16	8.5	13	15
	Mean	12.2 [9.0, 15.3]	9.8 [7.1, 12.4]	11.9 [9.6, 14.1]	14.3 [12.2, 16.5]
Reported Compliance [%]	Mean	34.2 [18.8, 49.5]	52.6 [40.0, 65.2]	35.1 [21.4 48.7]	31.4 [15.6, 47.2]
Trust in Automation	Mean	48 [40, 56]	53 [44, 61]	37 [27, 47]	28 [20, 35]

Table 1. Subjective and objective PERFORMANCE and COMPLIANCE, and TRUST IN AUTOMATION based on CORRECTNESS for both game rounds. Numbers in brackets show the 95%-CI.

4.4. *System Correctness and Trust*

The correctness also influenced the participants' trust in the system. In the first round of the game, the effect is suggestive for statistical significance ($F_{1,37} = 3.401$, $p = .073$, $\eta^2 = .084$). For the second round of the game a significant effect emerges of CORRECTNESS on TRUST ($F_{1,37} = 19.763$, $p < .001$, $\eta^2 = .348$).

In the first round of the game, the reported TRUST in the correct system was slightly higher (48 [40, 56]) than the trust in the defective system (37 [27, 47]; −24%). The difference in the evaluations grew after the second round: TRUST in the correct system (53 [44, 61]) was much higher than for the defective system (28 [20, 35]; −48%). Table 1 summarizes these findings and Figure 7 illustrates the changes based on this effect.

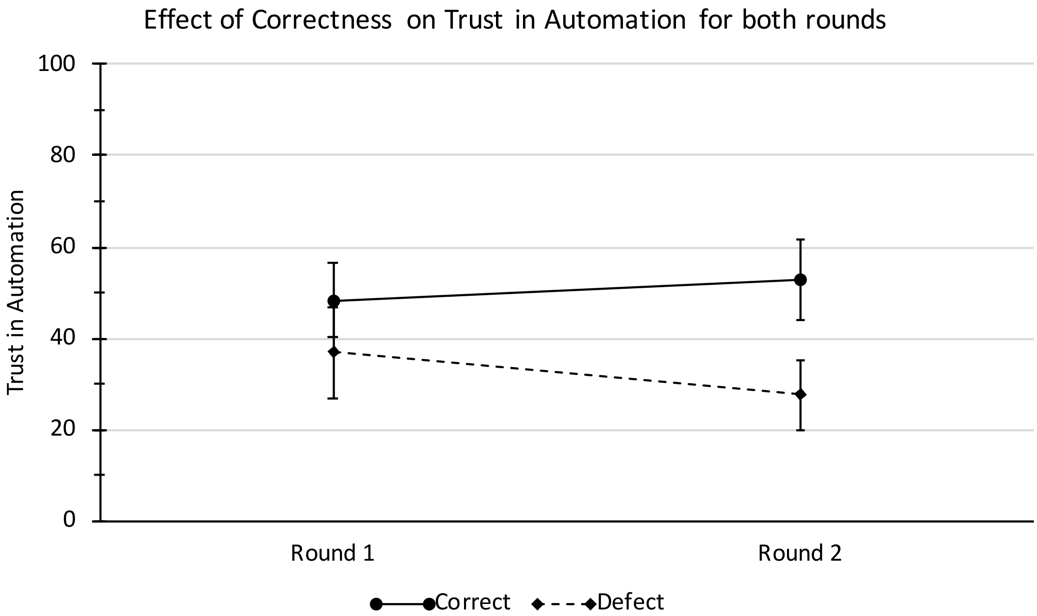

Figure 7. Effect of CORRECTNESS and ROUND on TRUST. If the system behaves correctly, trust increases in round 2. Error bars indicate the 95%-CI.

Thus, the correctness of the DSS influences the participants' trust in the system, although the effect shows up not earlier than the second round of the game.

4.5. *System Correctness and Technology Acceptance*

This section analyses the influence of correctness and trust on the acceptance of the support system on the basis of Davis' technology acceptance model (see section 2.5). For the analysis, both rounds are pooled together.

To understand if the CORRECTNESS influences the evaluation of the decision support system on the dimensions of the Technology Acceptance model, we calculated a repeated measures MANOVA with CORRECTNESS as a within-subject factor and USEFULNESS, EASE OF USE, and INTENTION TO USE as dependent variables. Overall, CORRECTNESS had a significant influence on the overall model ($V = .644$, $F_{4,33} = 14.941$, $p < .001$, $\eta^2 = .644$).

Specifically, the perceived USEFULNESS decreases significantly by 57% from 54 to 23 based on the system defect ($F_{1,36} = 41.353$, $p < .001$, $\eta^2 = .535$). Although the effect size and the relative decrease is smaller, the defect also decreases the perceived EASE OF USING the system by 26% from 60 to 44 ($F_{1,36} = 11.365$, $p = .001$, $\eta^2 = .287$).

The overall INTENTION TO USE the correct support system was 52 and decreased by 62% to 20 for the defective system ($F_{1,36} = 47.978$, $p < .001$, $\eta^2 = .571$). Table 2 summarizes and Figure 8 illustrates the effect of CORRECTNESS of the decision support system on the variables from the acceptance model.

Scale	Correct	Defect	Delta	η^2
Usefulness	54 [44, 64]	23 [15, 31]	-57%	.535
Ease of Use	60 [52, 69]	44 [35, 53]	-26%	.287
Intention To Use	52 [42, 62]	20 [13, 27]	-62%	.571

Table 2. Influence of the correctness of the decision support system on the dimensions USEFULNESS, EASE OF USE, and INTENTION TO USE. Numbers in brackets show the 95%-CI.

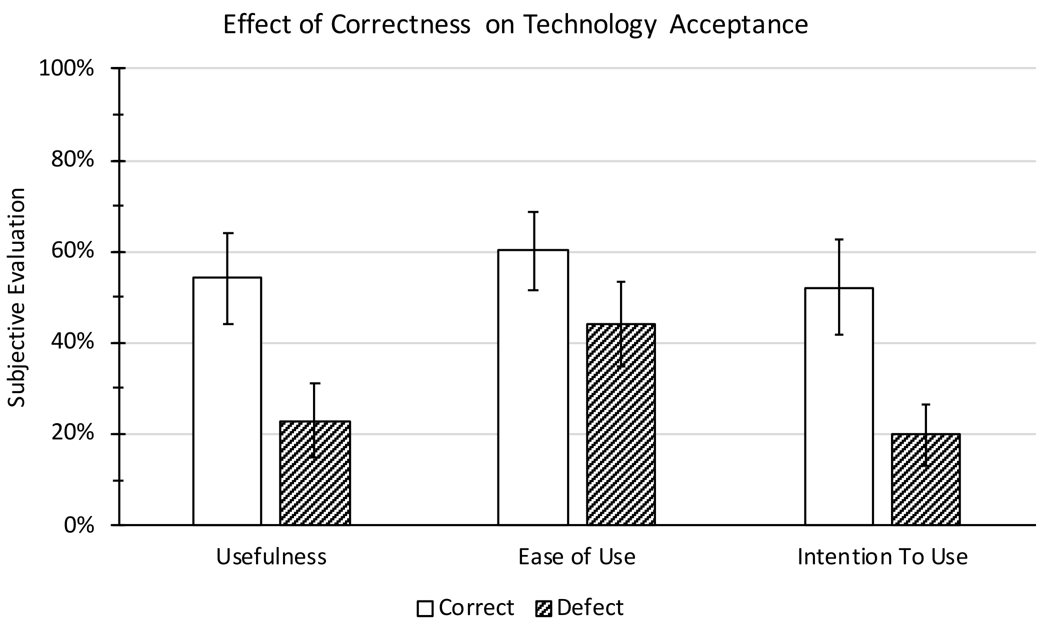

Figure 8. Evaluation of perceived USEFULNESS, EASE OF USE, and INTENTION TO USE based on CORRECTNESS of the decision support systems. Error bars indicate the 95%-CI.

It should be noted that the effect size for perceived EASE OF USE ($\eta^2 = .287$) is lower than the effect size of perceived USEFULNESS ($\eta^2 = .535$), and that of the INTENTION TO USE ($\eta^2 = .571$). Although the participants report lower EASE OF USING for the defect DSS than for the correct DSS, the effect is much less pronounced as for the other two criteria, especially than for the most decisive dimension INTENTION TO USE.

4.5.1. Correlation Analysis for the Correct *Decision Support System*

For the correct decision support system, both, the perceived USEFULNESS and the EASE OF USING relate to the INTENTION TO USE: Still, perceived USEFULNESS is linked more strongly ($\rho = .832$ [.703, .908], $p < .001$) than EASE OF USING the system ($\rho = .439$ [.148, .660], $p = .005$).

Furthermore, the INTENTION TO USE was strongly linked to the reported COMPLIANCE with the system ($\rho = .616$ [.377, .778], $p < .001$).

The additional component TRUST in the system shapes the INTENTION TO USE the system ($\rho = .719$ [.526, .841], $p < .001$) , as well as the reported COMPLIANCE with the system ($\rho = .453$ [.164, .669], $p = .004$). Users who trust the system and comply with it, intend to use it later.

In addition, the EASE OF USE, USEFULNESS, and TRUST in the system are interconnected ($\rho \geq .397$ [.098, .630], $p < .001$). Although no significant effect of EASE OF USE on reported COMPLIANCE was discovered ($\rho = .236$ [-.081, .510], $p = .147$), the perceived USEFULNESS had a strong influence on the reported compliance ($\rho = .634$ [.402, .789], $p < .001$). Users who perceived the system as useful, complied with it more often.

Contrary to expectations, neither INTENTION TO USE ($\rho = .022$ [-.291, .331], $p = .896$), nor COMPLIANCE with the system ($\rho = -.174$ [-.145, .460], $p = .295$) were linked to the overall company PROFIT. Profit was independent of whether users followed the system or later report that they would rely on it again. However, there was a positive effect of INTENTION TO USE ($\rho = .361$ [.056, .604], $p = .022$) and a effect suggestive of statistical significance of COMPLIANCE ($\rho = .284$ [-.030, .547], $p = .080$) on RELATIVE PERFORMANCE. Hence, participants complying with the support system at least achieved a higher *perceived* profit in the game.

The left side of Figure 9 illustrates the relationships between the system's evaluations and compliance with the correctly functioning support system.

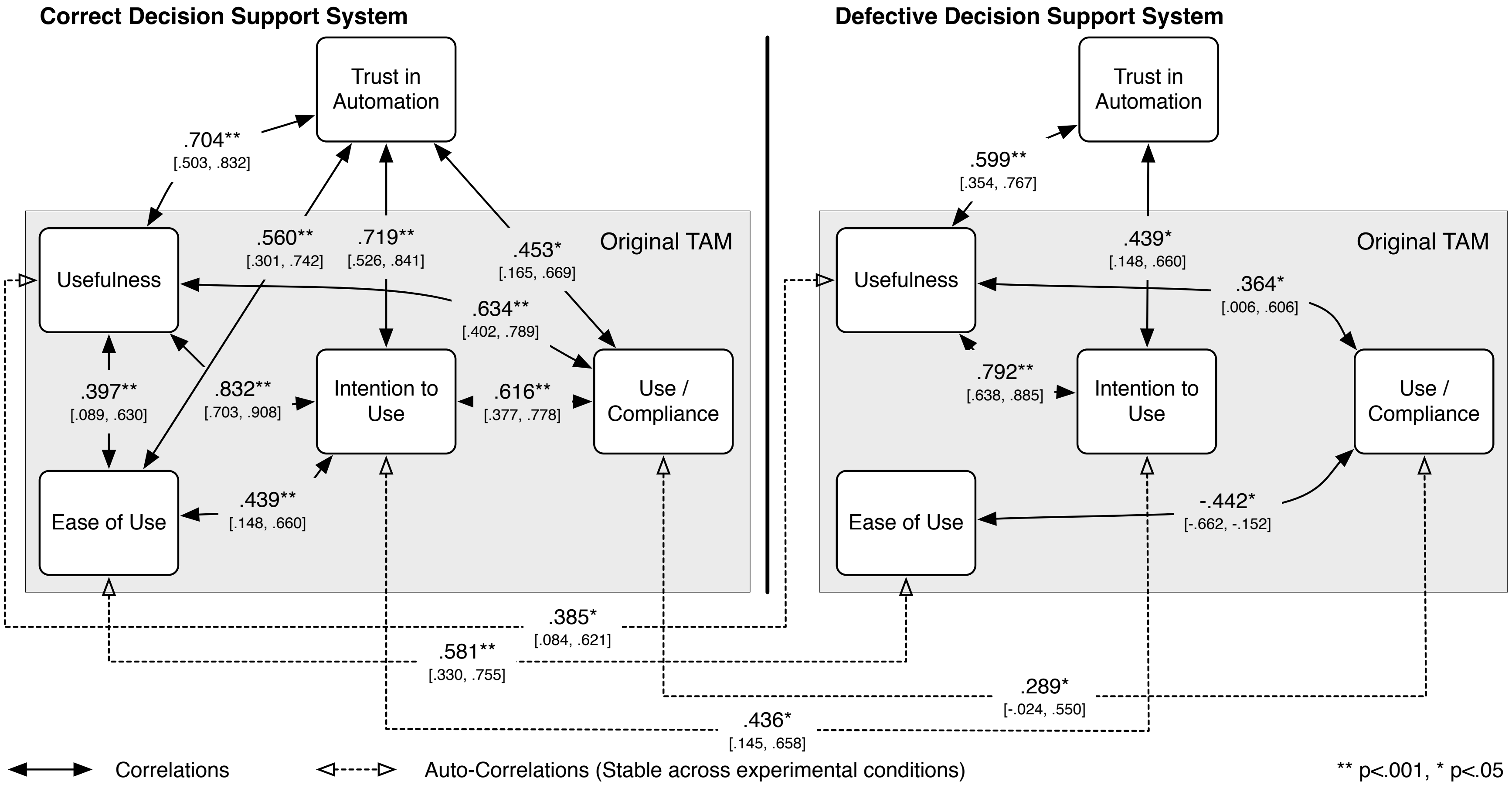

Figure 9. Correlation networks of TRUST, USEFULNESS, EASE OF USE, INTENTION TO USE, and COMPLIANCE for *correct* (left) and *defective* (right) decision support systems. Values in brackets show the 95% confidence interval.

4.5.2. *Correlation Analysis for the* Defective *Decision Support System*

For the defective decision support system, the perceived usefulness was strongly related to the intention to use the system ($\rho = .792$ [.638, .885], $p < .001$). The more useful the system, the more likely users would use it. Intention to use was also influenced by the participant's trust in the system, although the strength of this effect is weaker ($\rho = .439$ [.148, .660], $p = .005$). The more I trust the system, the more I am likely to rely on it. Surprisingly, the perceived ease of use was not related to the intention to use the system ($\rho = -.022$ [-.331, .291], $p = .892$).

As above, there is a link between usefulness and trust ($\rho = .599$ [.354, .770], $p < .001$), but no correlation between trust and ease of use ($p = .581$), and ease of use and usefulness ($p = .376$).

No significant link was found between intention to use the system and the reported compliance ($\rho = .026$ [-.287, .334], $p = .134$). Still, ease of use and usefulness were found to influence the compliance: There is a medium positive influence of perceived usefulness on compliance ($\rho = .364$ [.060, .660], $p = .034$), whereas the medium influence of ease of using the system is negative ($\rho = .442$ [.152, .662], $p = .010$).

The intention to use the support system and the actual *company profit* are unrelated ($\rho = -.139$ [-.180, .431], $p = .393$). Higher compliance with the defective system is strongly related with lower overall company profits ($\rho = -.668$ [.451, .810], $p < .001$). And, although intention to use was not related to relative performance ($\rho = .079$ [-.238, .381], $p. = 630$), there was a link between the reported compliance and ($\rho = -.339$ [.031, .588], $p = .049$) and relative performance. Thus, users that followed the broken recommendations, felt that they had performed rather well.

5. Discussion

Prior work has shown that decision support system increase efficiency and effectiveness in decision tasks in various contexts (Sharda et al. 1988; Garg et al. 2005; Pick 2008; Röttger et al. 2009; Onnasch et al. 2014). Generally, DSS are essential to cope with the increasing task complexities in cyber-physical production systems and supply chains (see sections 2.1, 2.1, and 2.2). Our work has empirically investigated how the compliance with a decision support systems relates to trust and technology acceptance for both correct and defective decision support systems. In the following we discuss the main findings with regard to the relevant body of knowledge. The first finding of the study is the proximity of *perceived* and *measured* metrics of the study: In general, the participants in our study could correctly assess their compliance with the decision support system and their overall performance in the game.

The results show that the defect of the decision support system had a strong effect on almost all investigated dependent variables. A correct system yielded higher trust, higher usefulness, and higher ease of use, which in turn increased the intention to use the support system and the overall compliance with the system. Most importantly, the correct support system leads to a higher performance satisfaction, and a higher overall company profit.

Thus, the self-evident—if not trivial—conclusion is that decision support systems should work as reliably as possible and should not provide misleading or wrong suggestions to the operators of cyber-physical production systems. Yet, this best-case is not achievable, due to errors in programming, sensors, and specifications, or due to unforeseen events in a production network. Therefore, it is essential to understand who complies with correct *and* defective systems. It is crucial to understand why and under which conditions compliance occurs, in order to empower operators to detect and intentionally disregard faulty suggestions.

The analysis of the relationships that govern acceptance and use of the correct decision support system is in line with the underlying theory and our expectations: The intention to use the system is governed by the participants' trust in the automated system, their perceived usefulness of the system, as well as its ease of use. As one would expect, trust and usefulness are the strongest predictors. In addition, ease of use also unfolds a strong positive influence on the trust in the automated system, as well as in the intention to use the system. The reported compliance with the system, which matched the measured compliance from the simulation model, also correlates with trust in the system, the reported usefulness, and the intention to use the system.

At this point, one obvious finding becomes apparent: The more a user disobeys a defective support system, the more profit his company makes (see section 4.5.2). However, users comply with the systems to a large extent in both the defect and the correct case. This shows that some users tend to follow the misguiding suggestions of the support system, despite evident feedback through customer complaints in the game (see section 3.2.1; Te'eni (1991)).

But what drives this obedience of the defective system? People have different reasons for complying with bad recommendations. Compliance increases when the system is *perceived* as useful and *decreases* when the system is perceived as easy to use. Apparently, some people attribute the system a higher usefulness, despite its evident malfunction. It is not yet sufficiently understood what drives this misconception, but it allures users into compliance and blind obedience of a defective support system.

It is both striking and remarkable that *ease of use* plays a different role in the correct setting than in the defect setting. People who find the system easy to use comply *less* with the defective system and then attain higher profits. This is in line with previous work that showed that ease of use and interface usability have considerable positive influence on performance, although often not directly measurable and only unfolding in more complex settings (cf. section 2.2 and (Parker and Sinclair 2001; Mittelstädt et al. 2015; Brauner et al. 2016)). Our study suggests that higher *perceived* ease of use enables users to compensate the defects as users intentionally disregard the misleading suggestions by complying less, and thereby generating higher profits.

Interestingly, the effects of the correctness of the DSS did not show up in the first round of the game, but most prominent in the second round. There are two explanations for this effect: On the one hand, one could speculate that participants stay with their strategy from the first round despite the defectiveness of the system. Then, performance and acceptance measures would decrease not earlier than in the second round. On the other hand, one could suggest that it takes some time (the first round of the game) until participants become aware of the defectiveness of the system, and effects can be seen not earlier than in the second round. The second suggestion is more probable, as the performance in the first round was quite low and this contradicts the idea that participants keep their high performance strategy from the first round before their performance decreases in the second round.

Naturally, correlation does not imply causality. Within the scope of this experiment, the usability of the interface was not systematically varied and the positive effect emerged for *perceived* ease of use. Future work will have to investigate if the participants' evaluation of the perceived ease of use is higher because of higher profits or if players that find the interface easier to use have sufficient cognitive resources to spare to detect and compensate the defect (cf. Chandler and Sweller (1991)). Hence, an increased focus on the usability of interfaces of automated cyber-physical production systems may yield higher acceptance in case of correct automation, lower compliance in case of incorrect automation, and may consequently reduce Bainbridge's *Ironies of Automation* (1983) (see section 2.3).

Lastly, none of the investigated user factors influenced the intention to use the support system or the compliance with the system. This contradicts previous findings on the impact of user factors on performance in ERP systems, in which especially processing speed impacted performance (Ziefle et al. 2015; Mittelstädt et al. 2015). Beyond age and gender, which are quite generic, we captured the persons' self-efficacy in interacting with technology as well as trust in automation. On the base of the results we can only speculate why user factors did not impact the results. Several reasons could account for the finding: Due to the comparably small sample size and the relative complexity of the task, effects could have been veiled. Or, the sample could have been too homogeneous with regard to the user factors under study and therefore, did not influence the dependent measures significantly. Future work will have to reassess how personality and traits influence human performance in cyber-physical production systems. Especially the role of trust in automated systems should be systematically evaluated. In this work, trust is only shaped by the correct or defect behavior of the support system, but we were not able to predict trust in advance and, surprisingly, all trust measures captured at the beginning and after the game rounds were unrelated. This would indicate that trust is determined only by the functioning of the system and not by individual differences or trust dispositions.

In summary, this article shows that the defect of a DSS in a cyber-physical production system has a strong negative effect on user perceptions, such as trust, usefulness, ease of use, and intention to use, but also in the attained performance of the system. This is striking, as the system supported only one (and a comparatively simple) of multiple decisions in a rather complex experimental setting. A further key contribution of this study is the analysis of the factors that govern compliance with the system: The results show that perceived ease of using the system is negatively related to compliance in the defect case: People who find the interface easy to use seem to have enough spare cognitive capacity to detect and compensate the defect system and therefore show higher performance in managing the cyber-physical production system and attain higher profits.

Why is this relevant? While the positive effects of good interface design to compensate system errors are well studied in some domains, such as medical informatics or aviation (Goddard et al. 2012), the emerging fields of the Industrial Internet and cyber-physical productions systems often neglect the human factors perspective, especially in less tangible contexts as cross-company collaboration and supply chain management. In that sense, the main function of this article might be a call to action to study the influence of system, interface, and user factors on performance, to transfer and validate the findings from other research domains, and to canonize the results for the design of cyber-physical production systems.

6. Limitations and Outlook

Of course, this study is not without limitations. The first limitation relates to the relatively high drop out rate that lies above the (reported) rates of many other studies.

In contrast to traditional laboratory settings, web based studies require high motivation of the participants (Crump et al. 2013), especially when the experiment is long and complex and participation is voluntary and not gratified.

While we acknowledge that the sample size of 40 participants is quantitatively limited, the quality of the findings seem unaffected because of two reasons: The first reason is a methodological one and refers to the fact that the study focuses on the within-subject factor CORRECTNESS. This factor is neither related to the dropout rate nor to the other investigated user factors. The second reason is related to participants' motivation and the compliance with the experimental task: Those participants that kept up and finished the experiment probably were more involved and took the tasks seriously, thereby resembling the attitude of real workers that have to handle their daily business.

Still, the analysis of user diversity effects and other interesting relationships, such as the influence of the disposition to trust, could not be sufficiently investigated because of the small sample size. Therefore, a follow-up study with a larger sample is the next step to replicate the findings and to provide valuable insights in the effects of user diversity for researchers and practitioners.

Furthermore, we have focused on the influence of system correctness, but without manipulating the complexity of the underlying simulation and without investigation different application fields and contexts for decision support systems (e.g., medical technology). Acceptance and compliance with decision support systems might not only be shaped by system correctness, but also by how necessary the decision support is from a user's perspective. Consequently, further research should address the transferability of the findings to lesser or more complex environments and different fields of applications, such as decision support in health care or financial controlling.

We have shown that understanding how humans behave with automated systems in cyber-physical-production-systems is essential. It ensures viability, competitiveness, and economic growth of manufacturing companies and societies building on these industries.

Acknowledgments

Authors owe gratitude to participants for their commitment and dedication to contribute to this research. Also, thanks are devoted to Sebastian Stiller, Quoc Hao Ngo, Marco Fuhrmann, and Robert Schmitt for in-depth discussions and valuable feedback on this work. Further thanks go to Julia Offermann-van Heek, Anne Kathrin Schaar, Patrick Halbach, Fabian Comanns, and Sabrina Schulte for their research. Finally, thanks to the anonymous reviewers for their critical and very helpful comments on this article.

The German Research Foundation (DFG) funded this work within the Clusters of Excellence *Integrative Production Technology for High-Wage Countries* (EXC 128), the subproject *Cognition Enhanced Self-Optimizing Production Networks* (Schlick et al. 2017), and *Internet of Production* (EXC 2023).

The dataset is publicly available (Brauner et al. 2018).

References

T. Bock, The future of construction automation: Technological disruption and the upcoming ubiquity of robotics, Automation in Construction 59 (2015) 113–121.

L. Monostori, Cyber-physical production systems: Roots, expectations and R&D challenges, Procedia CIRP 17 (2014) 9–13.

E. A. Lee, Cyber Physical Systems: Design Challenges, in: 2008 11th IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing (ISORC), pp. 363–369.

M. Wollschlaeger, T. Sauter, J. Jasperneite, The future of industrial communication: Automation networks in the era of the internet of things and industry 4.0, IEEE Industrial Electronics Magazine 11 (2017) 17–27.

M. Brettel, N. Friederichsen, M. Keller, M. Rosenberg, How virtualization, decentralization and network building change the manufacturing landscape: An industry 4.0 perspective, International Journal of Mechanical, Industrial Science and Engineering 8 (2014) 37–44.

T. H. Davenport, J. G. Harris, Automated decision making comes of age, MIT Sloan Management Review 46 (2005) 83.

K. L. Mosier, L. J. Skitka, 10 human decision makers and automated decision aids: Made for each other?, Automation and human performance: Theory and applications (2018) 120.

E. M. Frazzon, J. Hartmann, T. Makuschewitz, B. Scholz-Reiter, Towards socio-cyber-physical systems in production networks, Procedia CIRP 7 (2013) 49–54.

P. Brauner, A. C. Valdez, R. Philipsen, M. Ziefle, On Studying Human Factors in Complex Cyber-Physical Systems, in: Workshop Human Factors in Information Visualization and Decision Support Systems Held as Part of the Mensch und Computer 2016, Gesellschaft für Informatik, 2016.

G. A. Gorry, M. S. S. Morton, A Framework For Management Information Systems, Sloan Management Review 13 (1971) 50–70.

J. P. Shim, M. Warkentin, J. F. Courtney, D. J. Power, R. Sharda, C. Carlsson, Past, Present, and Future of Decision Support Technology, 2002.

J. D. Lee, K. a. See, Trust in automation: Designing for appropriate reliance., Human Factors 46 (2004) 50–80.

B. M. Muir, N. Moray, Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation, Ergonomics 39 (1996) 429–60.

M. Ziemann, Y. Eren, A. El-Osta, Gene name errors are widespread in the scientific literature, Genome Biology 17 (2016) 177.

C. Parker, M. Sinclair, User-centred design does make a difference. the case of decision support systems in crop production, Behaviour & Information Technology 20 (2001) 449–460.

D. Te'eni, Feedback as a source of control in decision support systems: an experiment with the feedback specificity, Behaviour & Information Technology 10 (1991) 373–382.

P. R. Kleindorfer, G. H. Saad, Managing Disruption Risks in Global Supply Chains, Production and Operations Management 14 (2005) 53–68.

L. V. Snyder, Z. Atan, P. Peng, Y. Rong, A. J. Schmitt, B. Sinsoysal, OR/MS Models for Supply Chain Disruptions: A Review, IIE Transactions 48 (2016) 89–109.

J. W. Forrester, Industrial Dynamics, MIT Press, Cambridge, 1961.

H. L. Lee, V. Padmanabhan, S. Whang, Information Distortion in a Supply Chain: The Bullwhip Effect, Management Science 43 (1997) 546–558.

A. A. Tako, S. Robinson, The application of discrete event simulation and system dynamics in the logistics and supply chain context, Decision Support Systems 52 (2012) 802–815.

P. Brauner, R. Philipsen, A. Fels, M. Fuhrmann, H. Ngo, S. Stiller, R. Schmitt, M. Ziefle, A Game-Based Approach to Meet the Challenges of Decision Processes in Ramp-Up Management, Quality Management Journal 23 (2016) 55–69.

M. C. Wilson, The impact of transportation disruptions on supply chain performance, Transportation Research Part E: Logistics and Transportation Review 43 (2007) 295–320.

C. Y. Wong, S. Boon-Itt, C. W. Y. Wong, The Contingency Effects of Environmental Uncertainty on the Relationship Between Supply Chain Integration and Operational Performance, Journal of Operations Management 29 (2011) 604–615.

C. Tang, Robust strategies for mitigating supply chain disruptions, International Journal of Logistics 9 (2006) 33–45.

J. Blackhurst, C. W. Craighead, D. Elkins, R. B. Handfield, An empirically derived agenda of critical research issues for managing supply-chain disruptions, International Journal of Production Research 43 (2005) 4067–4081.

A. Kanda, A. Barve, R. Shankar, The role of human factors in agile supply chains, European Journal of Industrial Engineering 3 (2009) 2–20.

V. Mittelstädt, P. Brauner, M. Blum, M. Ziefle, On the Visual Design of ERP Systems – The Role of Information Complexity, Presentation and Human Factors, in: 6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015, pp. 270–277.

M. Ziefle, P. Brauner, F. Speicher, Effects of Data Presentation and Perceptual Speed on Speed and Accuracy in Table Reading for inventory Control, Occupational Ergonomics 12 (2015) 119–129.

P. Brauner, A. Calero Valdez, R. Philipsen, M. Ziefle, Defective Still Deflective – How Correctness of Decision Support Systems Influences User's Performance in Production Environments, in: F. F.-H. Nah, C.-H. Tan (Eds.), HCI in Business, Government, and Organizations (HCIGO), Held as Part of HCI International 2016, Springer International Publishing Switzerland 2016, 2016, pp. 16–27.

R. Parasuraman, T. B. Sheridan, C. D. Wickens, A Model for Types and Levels of Human Interaction with Automation, Systems, Man and Cybernetics 30 (2000) 286–297.

L. Bainbridge, Ironies of Automation, Automatica 19 (1983) 775–779.

R. Parasuraman, V. Riley, Humans and Automation: Use, Misuse, Disuse, Abuse, Human Factors: The Journal of the Human Factors and Ergonomics Society 39 (1997) 230–253.

K. Goddard, A. Roudsari, J. C. Wyatt, Automation bias: A systematic review of frequency, effect mediators, and mitigators, Journal of the American Medical Informatics Association 19 (2012) 121–127.

R. Parasuraman, D. H. Manzey, Complacency and bias in human use of automation: An attentional integration, Human Factors 52 (2010) 381–410.

B. Reeves, C. Nass, The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places, Cambridge University Press, New York, NY, USA, 1996.

B. J. Fogg, H. Tseng, The Elements of Computer Credibility, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '99, ACM, New York, NY, USA, 1999, pp. 80–87.

P. de Vries, C. Midden, D. Bouwhuis, The Effects of Errors on System Trust, Self-Confidence, and the Allocation of Control in Route Planning, International Journal of Human Computer Studies 58 (2003) 719–735.

M. Mayeh, T. Ramayah, A. Mishra, The Role of Absorptive Capacity, Communication and Trust in ERP Adoption, Journal of Systems and Software 119 (2016) 58–69.

R. Pak, A. C. McLaughlin, W. Leidheiser, E. Rovira, The Effect of Individual Differences in Working Memory in Older Adults on Performance with Different Degrees of Automated Technology, Ergonomics 60 (2017) 518–532.

T. Gilovich, D. Griffi, D. Kahneman, Heuristics and Biases: The Psychology of Intuitive Judgment, Cambridge University Press, 8th edition, 2002.

M. Zyda, From Visual Simulation to Virtual Reality to Games, USC Information Sciences Institute 38 (2005) 25–32.

A. A. Deshpande, S. H. Huang, Simulation games in engineering education: A state-of-the-art review, Computer Applications in Engineering Education 19 (2011) 399–410.

P. Brauner, M. Ziefle, How to Train Employees, identify Task-relevant Human Factors, and Improve Software Systems with Business Simulation Games, in: D. Dimitrov, T. Oosthuizen (Eds.), Proceedings of the 6th International Conference on Competitive Manufacturing 2016 (COMA '16), CIRP, Stellenbosch, South Africa, 2016, pp. 541–546.

J. D. Sterman, Modeling Managerial Behavior: Misperceptions of Feedback in a Dynamic Decision Making Experiment, Management Science 35 (1989) 321–339.

S. Sarkar, S. Kumar, A Behavioral Experiment on Inventory Management With Supply Chain

Disruption, International Journal of Production Economics 169 (2015) 169–178.
T. Ben-Zvi, The efficacy of business simulation games in creating Decision Support Systems: An experimental investigation, Decision Support Systems 49 (2010) 61–69.
T. Ben-Zvi, Measuring the Perceived Effectiveness of Decision Support Systems and their Impact on Performance, Decision Support Systems 54 (2012) 248–256.
E. M. Goldratt, J. Cox, The Goal: A Process of Ongoing Improvement, North River Press, Great Barrington, Massachusetts, 2nd edition, 1992.
S. Stiller, B. Falk, R. Philipsen, P. Brauner, R. Schmitt, M. Ziefle, A Game-based Approach to Understand Human Factors in Supply Chains and Quality Management, Procedia CIRP 20 (2014) 67–73.
R. Philipsen, P. Brauner, S. Stiller, M. Ziefle, R. Schmitt, Understanding and Supporting Decision Makers in Quality Management of Production Networks, in: Advances in the Ergonomics in Manufacturing. Managing the Enterprise of the Future 2014 : Proceedings of the 5th International Conference on Applied Human Factors and Ergonomics, AHFE 2014, CRC Press, Boca Raton, United States, 2014, pp. 94–105.
P. Brauner, R. Philipsen, M. Ziefle, Projecting Efficacy and Use of Business Simulation Games in the Production Domain Using Technology Acceptance Models, in: C. Schlick, S. Trzcieliński (Eds.), Advances in Ergonomics of Manufacturing: Managing the Enterprise of the Future, Springer International Publishing, 2016, pp. 607–620.
F. D. Davis, Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology, MIS Quarterly 13 (1989) 319–340.
V. Venkatesh, F. D. Davis, A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies, Management Science 46 (2000) 186–204.
V. Venkatesh, J. Y. L. Thong, X. Xu, Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology, MIS Quarterly 36 (2012) 157–178.
A. Yusoff, R. Crowder, L. Gilbert, Validation of Serious Games Attributes Using the Technology Acceptance Model, in: Second International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES), IEEE Computer Society Press, Braga, 2010, pp. 45–51.
D. Y. Wu, E. Katok, Learning, communication, and the bullwhip effect, Journal of Operations Management 24 (2006) 839–850.
A. Bandura, Self-Efficacy Mechanism in Human Agency, American Psychologist 37 (1982) 122–147.
K. Arning, M. Ziefle, Understanding Age Differences in PDA Acceptance and Performance, Computers in Human Behavior 23 (2007) 2904–2927.
G. Beier, Kontrollüberzeugungen im Umgang mit Technik [Locus of control when interacting with technology], Report Psychologie 24 (1999) 684–693.
J.-Y. Jian, A. M. Bisantz, C. G. Drury, Foundations for an Empirically Determined Scale of Trust in Automated System, International Journal of Cognitive Ergonomics 4 (2000) 53–71.
R. Selten, R. Stoecker, End behavior in sequences of finite Prisoner's Dilemma supergames - A learning theory approach, Journal of Economic Behavior and Organization 7 (1986) 47–70.
P. Brauner, R. Philipsen, A. Calero Valdez, Z. Martina, What happens when decision support systems fail? — the importance of usability on performance in erroneous systems [data file available at https://osf.io/q3x2b/], 2018.
K. Kelley, K. J. Preacher, On effect size, Psychological Methods 17 (2012) 137–152.
G. Cumming, The New Statistics: Why and How, Psychological Science 25 (2014) 7–29.
R. Sharda, S. H. Barr, J. C. McDonnell, Decision Support System Effectiveness: A Review and an Empirical Test, Management Science (1988).
A. X. Garg, N. J. Adhikari, H. McDonald, M. P. Rosas-Arellano, P. J. Devereaux, J. Beyene, J. Sam, R. B. Haynes, Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: A systematic review, JAMA 293 (2005) 1223–1238.
R. A. Pick, Benefits of Decision Support Systems, Springer Berlin Heidelberg, Berlin, Heidel-

berg, pp. 719–730.

S. Röttger, K. Bali, D. Manzey, Impact of automated decision aids on performance, operator behaviour and workload in a simulated supervisory control task, Ergonomics (2009).

L. Onnasch, C. D. Wickens, H. Li, D. Manzey, Human performance consequences of stages and levels of automation: An integrated meta-analysis, Human Factors (2014).

P. Chandler, J. Sweller, Cognitive Load Theory and the Format of Instruction, Cognition and Instruction 8 (1991) 293–332.

M. J. Crump, J. V. McDonnell, T. M. Gureckis, Evaluating amazon's mechanical turk as a tool for experimental behavioral research, PloS one 8 (2013) e57410.

C. Schlick, V. Stich, R. Schmitt, G. Schuh, M. Ziefle, C. Brecher, M. Blum, A. Mertens, M. Faber, S. Kuz, H. Petruck, M. Fuhrmann, M. Luckert, F. Brambring, C. Reuter, N. Hering, M. Groten, S. Korall, D. Pause, P. Brauner, W. Herfs, M. Odenbusch, S. Wein, S. Stiller, M. Berthold, Cognition-Enhanced, Self-optimizing Production Networks, in: C. Brecher, D. Özdemir (Eds.), Integrative Production Technology - Theory and Applications, Springer International Publishing, Berlin, Heidelberg, 2017, pp. 645–743.

I. Ajzen, The Theory of Planned Behavior, Organizational behavior and human decision processes 50 (1991) 179–211.

Appendix

Characteristics of the Sample

	Descriptives	1	2	3	4
1. Age	M=28.4±8.4 (20–56 years)	—			.365*
2. Gender	23 male, 17 female		—		
3. Trust in Automation	M=54±16% (3–78%)			—	
4. Technical Self-efficacy	M=70±19% (15–95%)				—

Table 3. Characteristics of the sample with arithmetic mean (M), standard deviation (±), minimum and maximum values, and Spearman's ρ correlations.

Questionnaires and Items used

Scale [Source]	Reliability	Item
Usefulness (PU) [Davis (1989)]	$\alpha \geq .940$	Using the system improves my performance in the game. Using the system in the game increases my productivity. Using the system enhances my effectiveness in the game. I find the system to be useful in the game.
Ease of Use (PEU) [Davis (1989)]	$\alpha \geq .799$	My interaction with the system is clear and understandable. Interacting with the system does not require a lot of my mental effort. I find the system to be easy to use. I find it easy to get the system to do what I want it to do.
Intention to Use (ItU) [Ajzen (1991); Davis (1989)]	$\alpha \geq .864$	Assuming I had access to the system, I intend to use it I plan to use the system the next time I play the game. Given that I had access to the system, I predict that I would use it.
Self-efficacy technology (SET) [Beier (1999)]	$\alpha = .867$	I am able to solve most of the technical problems I am faced with on my own. I really enjoy cracking a technical problem. As I have coped well with technical problems in the past, I feel optimistic about future technical problems. As I feel quite helpless towards technical devices, I keep my hands off them. It is difficult to find a technical problem that I am not able to solve. I only solve technical problems because I have to. I really like to solve new technical problems. In my circle of friends I am the one with the highest technical abilities.

Trust in Automation (TiA) [Jian et al. (2000)]	$\alpha \geq .795$	The system is deceptive. The system behaves in an underhanded manner. I am suspicious of the system's intent, action, or outputs. I am wary of the system. The system's actions will have a harmful or injurious outcome. I am confident in the system. The system provides security. The system has integrity. The system is dependable. The system is reliable. I can trust the system. I am familiar with the system.
Compliance / Use [Davis (1989)]	—	Approximately, how often did you follow the suggestion of the decision support system during the game (in %)
Subjective Performance	—	Are you satisfied with your performance in the game?
Relative Performance	—	How does your performance compare to the performance of other players?

Table 4.: Applied scales, item texts, and internal reliability Cronbach's α. Minimal reliability is reported for scales measured repeatedly. USEFULNESS, EASE OF USE, and INTENTION TO USE.