applied sciences

MDPI

Article

Validation of Trade-Off in Human–Automation Interaction: An Empirical Study of Contrasting Office Automation Effects on Task Performance and Workload

Byung Cheol Lee [1], Jangwoon Park [1], Heejin Jeong [2] and Jaehyun Park [3,*]

[1] Department of Engineering, Texas A&M University - Corpus Christi, Corpus Christi, TX 78412, USA; byungcheol.lee@tamucc.edu (B.C.L.); jangwoon.park@tamucc.edu (J.P.)

[2] Department of Mechanical and Industrial Engineering, University of Illinois at Chicago, Chicago, IL 60607, USA; heejinj@uic.edu

[3] Department of Industrial & Management Engineering, Incheon National University, Incheon 22012, Korea

* Correspondence: jaehpark@inu.ac.kr; Tel.: +82-32-835-8867

Received: 21 January 2020; Accepted: 11 February 2020; Published: 14 February 2020

Abstract: Automation aims to improve the task performance and the safety of human operators. The success of automation can be facilitated with well-designed human–automation interaction (HAI), which includes the consideration of a trade-off between the benefits of reliable automation and the cost of Failed automation. This study evaluated four different types of HAIs in order to validate the automation trade-off, and HAI types were configured by the levels and the statuses of office automation. The levels of automation were determined by information amount (i.e., Low and High), and the statues were decided by automation function (i.e., Routine and Failed). Task performance including task completion time and accuracy and subjective workload of participants were measured in the evaluation of the HAIs. Relatively better task performance (short task completion time and high accuracy) were presented in the High level in Routine automation, while no significant effects of automation level were reported in Failed automation. The subjective workload by the National Aeronautics and Space Administration (NASA) Task Load Index (TLX) showed higher workload in High and Failed automation than Low and Failed automation. The type of sub-functions and the task classification can be estimated as major causes of automation trade-off, and dissimilar results between empirical and subjective measures need to be considered in the design of effective HAI.

Keywords: human-automation interaction; user experience; workload; task performance; level and status of automation; evaluation

1. Introduction

Automation enables a broad range of systems to reduce errors, improve work performance, expand human capabilities, and decrease effort and stress during operations [1]. It provides support for perceptual-cognitive and decision-making tasks, thus reducing the physiological effort and workload for the human operators as well [1,2]. However, the benefits of automation may be realized and achieved only when the automation works as designed without any errors or malfunctions [3,4]. If the automation fails, or the outcomes do not meet with operators' expectations, operators may perceive the automation as unreliable and inconsistent. Furthermore, excessive dependence on the availability and reliability of automation might interfere with skill acquisition without automation, which would make human operators misuse or disuse automation.

This unreliability and inconsistency by Failed automation depends significantly on the levels of automation (LOA) and information processing stages. Previous research has defined the LOA by the amount of automation autonomy and human physical and cognitive activity [5,6]. Sheridan and Verplank [7] suggested 10 LOA as the basis of the classic human–machine task allocation principle: at a higher level, automation can execute decision-making tasks without the aid of human operators, at a lower level, it can execute an option within the operators' controls, and at a further lower level, it may be simply performed by the human operators. Kaber and Endsley [6] proposed that automation could be classified according to the information-processing stages: perceiving the status of the system variables or scanning displays (i.e., monitoring), formulation of task-processing plans (i.e., generating), deciding on an optimal plan (i.e., selecting), and the control actions at an interface (i.e., implementing). Elaborating on this classification, Parasuraman et. al. [4] presented a four-stage model of automation to support the design of human–automation interactions (HAI) in complex systems: information acquisition, information analysis, decision and action selection, and action implementation. This classification aids the formulation of specific function allocation schemes for automation. The combination of the LOA with information processing stages led to the development of a concept known as the degree of automation (DOA) [3]. Each function in a model or an information processing stage, either by the human or machine (or some combination thereof), is responsible for the effects of automation on the task performance.

Under routine conditions or at higher LOA or DOA, human operators simply supervise the automation in an automation-driven mode, while in the case of automation failure, they are expected to override the situation control, which is referred to as a human-driven mode [8]. Generally, in an automation-driven mode, the automation enhances the overall system performance, while in a human-driven mode, it simply supports the system operation, or delivers signals or warnings to human operators to maximize the performance [9]. Broadly, automation yields a trade-off, in which better performance is exerted when all automation routinely works but increased dependence is induced, and worse performance may be produced when it fails [10]. This trade-off is analogous to the "lumber jack effect" in which "the higher trees in the forest are, the farther they fall" [3,11].

The trade-off indicates that automation is compensated between the benefits of reliable automation and the expected costs of automation failures [12]. Not only does automated task performance but also workload and loss of situation awareness follow a similar trade-off. With a higher DOA, Routine automation progressively reduce the operator's workload, the automation enables the operator to engage in other concurrent tasks, but this frequently results in the loss of situation awareness (LSA). The possible LSA renders all sorts of operators' errors and a mistrust in automation or a lack of proper understanding [13,14], that is, errors where operators failed to respond to a critical situation if the automation failed to alert them properly or where operators followed incorrect advice of automation without detecting this failure. Interestingly, modified meta-analysis confirmed that the task performance is maintained at a similar level with increasing DOA and follows a "flat" function up to a certain critical point, which mean that a human operator in Failed automation is not as vulnerable from automation failure if the automation functions at the lower level (e.g., offering options) and earlier information processing stages (e.g., information acquisition or information analysis) [3]. However, this assertion is based on the results from only 18 studies, and statistical power is not enough. Moreover, heterogeneous types of automation in the studies seem not to provide consistent performance results, and the meta-variables form performance measures and subjective measures in workload and situation awareness may not yield clear statistical conclusions.

Office automation is one of the easily accessible automated systems and widely used in various areas. Transitions to modern office automation require learning, processing of new data inputs, and substantial adaptation by the office workers. Such transformational adoption of human–system interactions in office automation are often driven by meaningful user-oriented benefits. The improved work performance by office automation is usually understood by the perspective of automation technology; however, there is little consideration of the mental and physical costs to users. Closure of this gap is necessary to better understand the implications of office automation on human–system interactions and to the day-to-day office work performance. As a result, understanding how office automation affects user (worker) experience and workload can be considered a major part of human–system integration efforts in everyday working environments [15,16]. Workload is a broad, multifaceted term that encompasses the human "cost" of performing a task [17]. Despite the concern with high workload in a modern office environment and its links to performance degradation, to date few efforts have been made to design office technologies with office workers and team workload in mind [18].

This study attempted to evaluate the relationship between user work performance and subjective workload by different perceived trust levels in office automation. There have been few empirical efforts attempting to ratify this HAI concept in general automation use. Furthermore, the trade-off can be contrastingly perceived in subjective measures (i.e., subjective workload, trust). Although a number of studies have examined the user workloads and performance in different DOAs, the results were mixed [19–21], and relatively few addressed the issues of joint human-automation performance in imperfect automation conditions. Therefore, in this study, we designed automated proofreading tasks as an exemplary model of office automations that users can easily adapt to, and evaluated the task performance measures (i.e., task completion time and accuracy) at different levels and statuses of automation. To understand how automation in human–system interactions impact demands and efforts, we also measured subjective workload by the National Aeronautics and Space Administration (NASA) Task Load Index (TLX), which includes multiple subscales to subjectively evaluate user demand [22]. Subjective workload along with perceived trust levels described the current status of operative demand and identified the potential impact of HAI.

The remainder of this article is organized as follows: in Section 2, the automation design, experimental setup, and data analysis procedures which will be used in the experiment are described. The experimental results are presented and analyzed in Section 3. In Section 4, the effects of automation on task performance and workload are discussed. Finally, a short conclusion is drawn in Section 5 and the limitations and future studies are suggested in Section 6.

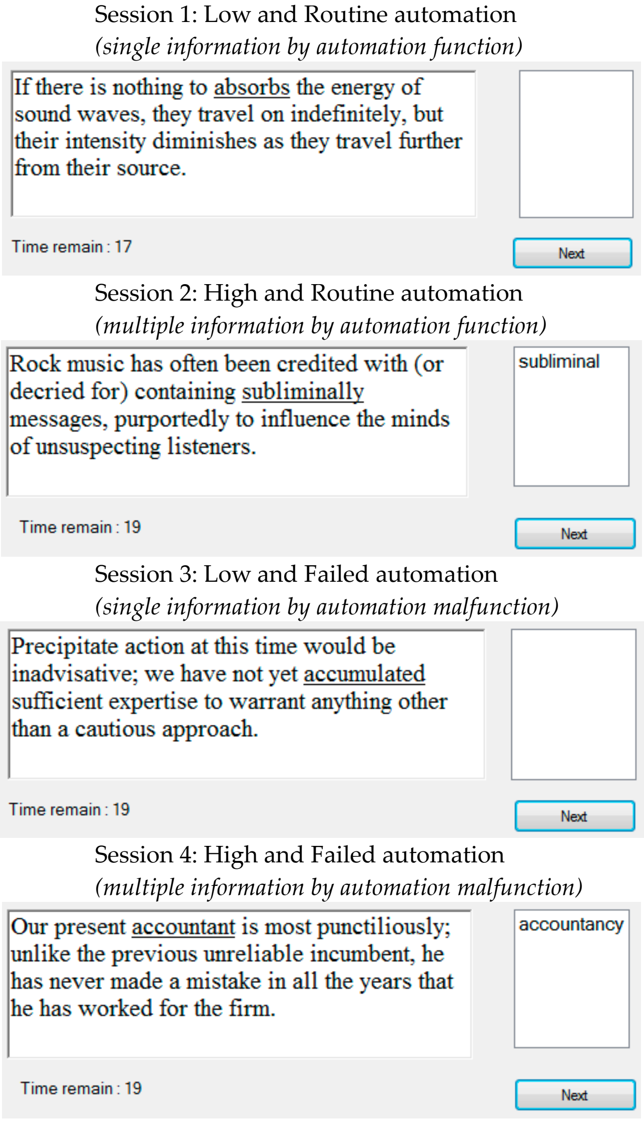
2. Materials and Methods

2.1. Participants

Staff and students in a university in the South Texas were recruited to participate. The Institutional Review Board approved this study with the following inclusion criteria for interested participants: (1) are native English speakers who have a similar level of reading comprehension capability of the English sentences and (2) are familiar with a word-processing program and auto-correction function.

2.2. Automated Tasks

Automation in this study was provided for performing proofreading, which is one of the most common types of automation experienced in word-processing. The participants were required to identify and correct typographical errors in a sentence. Generally, the automated proofreading function is already embedded in several word processors. Five proofreading tasks, similar to the AutoCorrect feature in Microsoft Word, including one non-automated and four automated proofreading tasks, were developed with a custom-built software program using Visual Studio 2016 (see Figure 1). In the non-automated task, the participants were asked to manually identify a typographical error in a sentence. The automated proofreading tasks were developed at two levels (i.e., Low and High) of automation and two statuses (i.e., Routine and Failed automation). At the Low automation, only an underlined word was provided to the participants, and, at the High automation, an underlined word was provided along with a substitute (e.g., "subliminal" and "accountancy" in Figure 1). The Routine automation status indicates that the underlined word is grammatically incorrect (e.g., "absorbs" and "subliminal"). On the other hand, Failed automation status indicates that the underlined word is grammatically correct (e.g., "accumulated" and "accountant"). Note that the substituting words in Session 4 (e.g., "accountancy") are grammatically incorrect and represent High and Failed automation.



Session 1: Low and Routine automation
(single information by automation function)

If there is nothing to absorbs the energy of sound waves, they travel on indefinitely, but their intensity diminishes as they travel further from their source.

Session 2: High and Routine automation
(multiple information by automation function)

Rock music has often been credited with (or decried for) containing subliminally messages, purportedly to influence the minds of unsuspecting listeners.

subliminal

Session 3: Low and Failed automation
(single information by automation malfunction)

Precipitate action at this time would be inadvisative; we have not yet accumulated sufficient expertise to warrant anything other than a cautious approach.

Session 4: High and Failed automation
(multiple information by automation malfunction)

Our present accountant is most punctiliously; unlike the previous unreliable incumbent, he has never made a mistake in all the years that he has worked for the firm.

accountancy

Figure 1. Four sessions of automated proofreading tasks by the combination of Low vs. High and Routine vs. Failed automation. While High automation underlines a gramatically incorrect word along with a possible substitute (multiple information), Low automation provide only a underlined word (single information). While a correct word is underlined and an incorrect substitute is presented in Failed automation, (malfunction of automation), an incorrect word is underlined and a correct substitute is provided in Routine automation.

The sentences used in the experiment were randomly selected from a sentence pool created by a two-step process. First, the sentences were extracted from well-known standardized tests (i.e., Scholastic Aptitude Test, Graduate Record Examination). Second, these sentences were categorized by two levels of readability (i.e., easy vs. difficult) using an online readability test tool [23]. As a result of the readability tests, a total of 120 sentences were collected: one half with readability scores between 9 and 14 (i.e., easy), and the remaining with readability scores higher than 14 (i.e., difficult).

2.3. Experiment Design and Procedure

A controlled laboratory experiment was conducted to investigate the effects of automation on human task performance. A 2 × 2 full-factorial design between-subject with two levels (i.e., Low and High) and two statuses (i.e., Routine and Failed) of automation were used in the experiment.

Each participant performed three sets of tasks: training, reference (i.e., non-automated proofreading), and main (i.e., automated proofreading). In the training task set, the participants were presumed to be familiar with the control of automated proofreading tasks. In the reference task set, the participants performed a non-automated proofreading task that provided a performance basis with 20 sentences that were used as a reference to assess the task completion time and the number of errors made in the task. After the reference task, the participants were allowed a two-minute break before starting the main task set. In the main task set, four task sessions (see Figure 1) were assigned to each participant by randomized orders of the sessions. Each session was composed of 20 automated proofreading tasks. After each session, the NASA TLX questionnaire with a single "trust in automation" question was provided to the participants for evaluating the subjective workload and trust level.

2.4. Dependent Variables

In this study, both task performance and subjective measures were collected. As task performance measures, task completion time measured as the delay between the displaying of the task on the screen and the clicking of the "Next" button, and task accuracy were assessed. The task completion time for each sentence was measured and the average values were computed per session. The task accuracy was calculated by the number of correct sentences per total sentences in a session. All task performance measures were compared with the measures of reference tasks and converted to the standardized performance ratios based on the non-automated condition to minimize individual difference, using the following equation:

$$Standardized\ Performance\ Ratio = \frac{Performance\ measure\ in\ automated\ tasks}{Performance\ measure\ in\ non-automated\ tasks} \quad (1)$$

Furthermore, the trust levels and subjective workload measure by NASA TLX questionnaire were evaluated as subjective operator performance parameters to understand the participants' perception of Routine/Failed automation at different automation levels. The NASA TLX questionnaire is a multidimensional subjective workload-rating method, which is composed of six subscales that reflect behavioral and subjective workload responses driven by the perception of task demands. In NASA TLX, the subjective workload is defined as the "cost incurred by human operators to achieve a specific level of performance [22]." This perceived subjective workload is evaluated as the integration of subjective responses and behaviors. These behaviors and subjective responses are guided by the perceptions of task demand, which can be quantified in terms of magnitude and importance. Table 1 showed the description of the questions for NASA TLX sub-scales and perceived trust.

Table 1. Survey questions of National Aeronautics and Space Administration (NASA) Task Load Index (TLX) subscales and perceived trust on automation.

Subscale	Description
Mental Demand	How much mental and perceptual activity was required? Was the task easy or demanding, simple or complex?
Frustration	How irritated, stressed, and annoyed versus content, relaxed, and complacent did you feel during the task?
Temporal Demand	How much time pressure did you feel due to the pace at which the tasks or task elements occurred? Was the pace slow or rapid?
Perceived Performance	How successful were you in performing the task? How satisfied were you with your performance?
Effort	How hard did you have to work (mentally and physically) to accomplish your level of performance?
Perceived Trust *	How trustworthy and helpful was the automation (indicating the word to be corrected or suggesting substitutes)?

* 3 scales: expected, less than expected, or more than expected.

2.5. Data Analysis

We measured task performance by standardized task completion time (ST) ratios and standardized task accuracy (SA) ratios. If the ST ratio is greater than 1, the automation is considered as having lengthened the task completion time; and if it is less than or equal to 1, the automation is considered as having accommodated the task completion time. Similarly, if a SA ratio is greater than 1, the automation improves the task completion accuracy; and if it is lower than or equal to 1, the automation causes more errors and deteriorates the task accuracy.

A repeated two-way analysis of variance (ANOVA) was performed to test for differences between the values of the measures, i.e., task completion time and accuracy, under different automated conditions. Prior to the analyses, all dependent variables were examined to check if the assumptions of ANOVA were met. This was achieved by checking normality using Curran's criteria of skewness (<2) and kurtosis (<7) [24], and homogeneity of variance across groups [25]. The alpha criterion of 0.05 was used to assess statistical significance. Furthermore, the Friedman test was performed as a post-hoc test using the Bonferroni method for testing both main and interaction effects. All analyses were performed using IBM SPSS Statistics version 24.

For a subjective workload measure, NASA TLX was administered post-trial, which required the participants to rate each question item. Five NASA TLX subscales including mental demand, temporal demand, effort, frustration, and perceived performance were rated on a 20-point scale (0-low, 20-high) [22]. Physical demand, one of subscales, was excluded considering task features. The "raw TLX" approach was applied without the pairwise comparison between the subscales in the NASA TLX data analysis, which is validated by many researchers [26]. The raw TLX approach is simpler to apply; the ratings are averaged or added to create an estimate of the overall workload, and an overall estimate of the subjective workload by each automation case was computed by averaging the scores describing each of the five subscales. In addition, a question to indicate a variation from perceived trust on automation, reported as "expected," "less than expected," or "more than expected," was asked to participants.

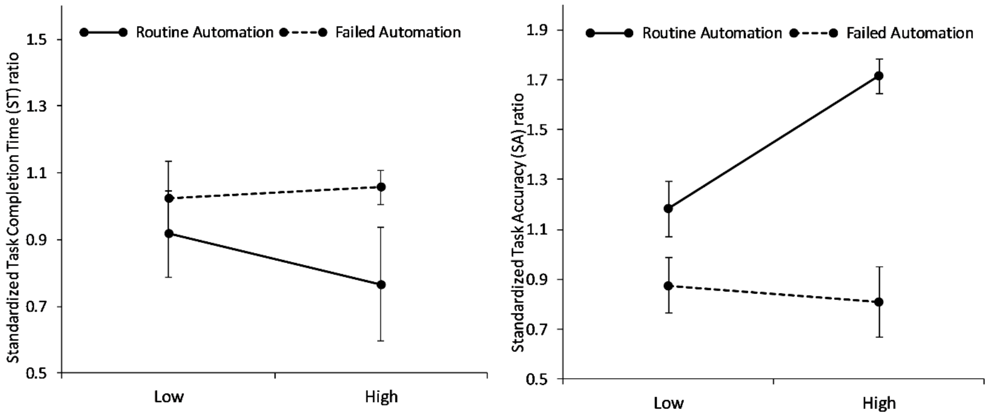
3. Results

Forty-nine college students and staffs (26 females and 23 males) in the 18–42 age range (mean (M) = 29.1 years, standard deviation (SD) = 4.3 years) took part in this study. Four participants did not complete the given task sets. Considering participants' average completion time, the tasks that required more than 60 s per sentence to complete and the sessions that took less than three minutes were excluded from further data analysis. In total, 166 ST ratios and 170 SA ratios were collected.

3.1. Task Performance Measures at Different Automation Levels and Statuses

Figure 2a describes the ST ratios at different automation levels and statuses. We observed that the ST ratios are significantly increased by the status (F (1,82) = 4.012, $p < 0.05$) and level × status (F (3,163) = 3.101, $p < 0.05$), but not by level (F (1,82) = 3.687, $p = 0.0583$). As expected, High and Routine automation showed significantly lower ST ratios than Low and Routine automation, whereas pairwise comparison of both Failed automations did not show a statistically significant difference (F (3,163) = 2.472, $p = 0.0636$).

Figure 2b describes the SA ratios at different automation levels and statuses. The SA ratios were significantly affected by level (F (1, 84) = 4.733, $p < 0.05$), status (F (1, 84) = 5.092, $p < 0.05$), and level*status (F (3, 167) = 2.756, $p < 0.05$). Similar to the ST ratio patterns, High and Routine automation shows significantly high SA ratios on automation statuses, and Low and Routine automation shows the second-highest accuracy, whereas pairwise comparison of Failed automation did not show the significant difference (F (3,167) = 3.121, $p = 0.717$). The differences in SA ratios were not significant in the Failed automation (F (1, 84) = 3.372, $p = 0.0699$).



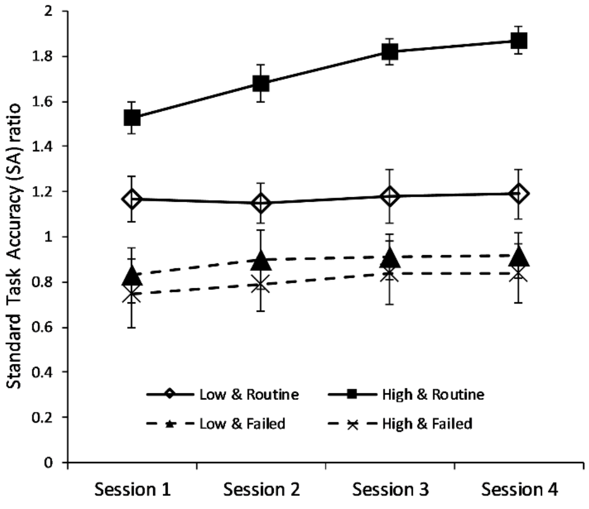
Routine	0.92 (0.11)	0.77 (0.05)
Failed	1.03 (0.13)	1.06 (0.17)

Routine	1.18 (0.11)	1.71 (0.07)
Failed	0.88 (0.11)	0.81 (0.14)

(**a**) (**b**)

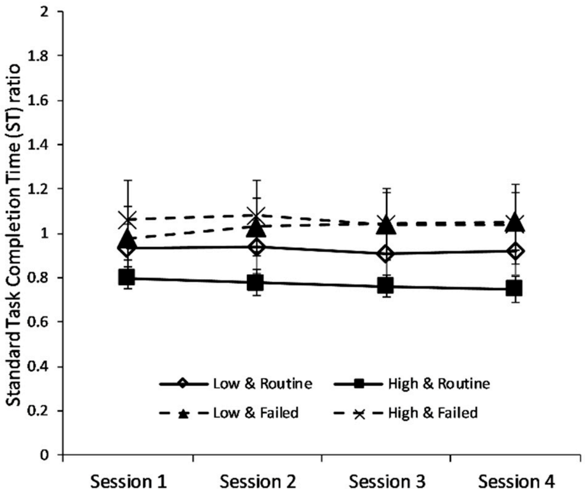
Figure 2. (**a**) Standardized task completion time (ST) ratios of four degrees of automation. (**b**) Standardized task completion accuracy (SA) ratios of four degrees of automation.

Longitudinal performance changes were examined for ST and SA ratios, as shown in Figures 3 and 4, respectively. These figures demonstrate how the users adapted to Routine and Failed automations over a prolonged period. As a result, except for SA ratios in High and Routine automation, no significant performance changes were observed. An ANOVA on SA ratios in High and Routine automation yielded a significant difference among sessions, F (3, 42) = 4.29, $p < 0.05$. A post-hoc test showed that SA ratios in the session 1 and in the session 4 differed significantly at $p < 0.05$. In addition, we could confirm that there was no longitudinal difference in statistically similar performances in Failed automation.



Low and Routine	1.17 (0.11)	1.15 (0.09)	1.18 (0.12)	1.19 (0.11)
High and Routine	1.53 (0.07)	1.68 (0.08)	1.82 (0.06)	1.87 (0.06)
Low and Failed	0.83 (0.12)	0.90 (0.13)	0.91 (0.10)	0.92 (0.10)
High and Failed	0.75 (0.15)	0.79 (0.12)	0.84 (0.14)	0.84 (0.13)

Figure 3. Variations in longitudinal SA ratios in the four sessions.



Low and Routine	0.98 (0.11)	0.94 (0.12)	0.91 (0.11)	0.92 (0.11)
High and Routine	0.79 (0.05)	0.77 (0.06)	0.76 (0.05)	0.75 (0.06)
Low and Failed	0.98 (0.14)	1.03 (0.13)	1.04 (0.14)	1.05 (0.13)
High and Failed	1.06 (0.18)	1.08 (0.16)	1.04 (0.16)	1.04 (0.18)

Figure 4. Variations in longitudinal ST ratios in the four sessions.

3.2. *Subjective Workload Measures in Different Automation Levels and Statuses*

Table 2 summarizes the descriptive statistics for overall subjective workloads by automation statuses and levels. Generally, a comparison of the effects of automation levels and statues on the workload showed that as the automation level is increased, the overall subjective workload decreased in Routine automation (F (1, 89) = 4.57, $p < 0.05$) and increased in Failed automation (F (1, 89) = 5.61, $p < 0.05$). These patterns correspond to those in the lumberjack hypothesis of automation [3]. The participants perceived that the subjective workloads are sensitive to the automation levels and Failed automation imposed more workload than Routine automation. Specifically, Table 3 indicates that "High and Failed" automation shows the highest subjective workload in the four different automation settings, in which "High and Routine" automation can be considered as the most operable condition.

Table 2. Overall subjective workload by automation level and status.

NASA TLX Scale	Automation Level and Status			
	Low and Routine	High and Routine	Low and Failed	High and Failed
Overall Subjective Workload	36.1 (6.72)	31.3 (5.58)	66.0 (10.71)	75.3 (8.01)

Table 3. NASA TLX subscales by perceived trust levels.

Perceived Trust	Total N = 45	Subjective Workload Subscale (Mean)				
		Mental Demand	Effort	Perceived Performance	Temporal Demand	Frustration
Lower than expected	7	13.8 A	9.6 A	13.4 A	13.8 A	14.0 A
As expected	11	8.6 B	8.2 B	8.4 B	9.4 B	10.2 B
Higher than expected	27	8.0 B	7.8 B	8.2 B	8.8 B	7.6 C

Difference in capital letters indicate significant difference ($p < 0.05$) in group means.

The evaluation of the subscales showed patterns different from the overall workload (see Figure 5). The effects of automation statuses were notable, while the effects of automation levels were mixed. Only perceived performance (*F* (3, 187) = 5.26, $p < 0.001$) and frustration (*F* (3, 187) = 11.879, $p < 0.001$) demonstrated significant effects of automation levels, while for the mental demand (*F* (3, 189) = 2.88, $p = 0.698$) and time demand (*F* (3, 189) = 2.78, $p = 0.708$) subscales, the automation level did not impact the measurements in Failed automation. Similarly, for the effort subscale, the participants did not recognize the effect of automation level in both Routine and Failed automations (*F* (3, 189) = 2.82, $p = 0.703$; *F* (3, 189) = 2.88, $p = 0.698$).

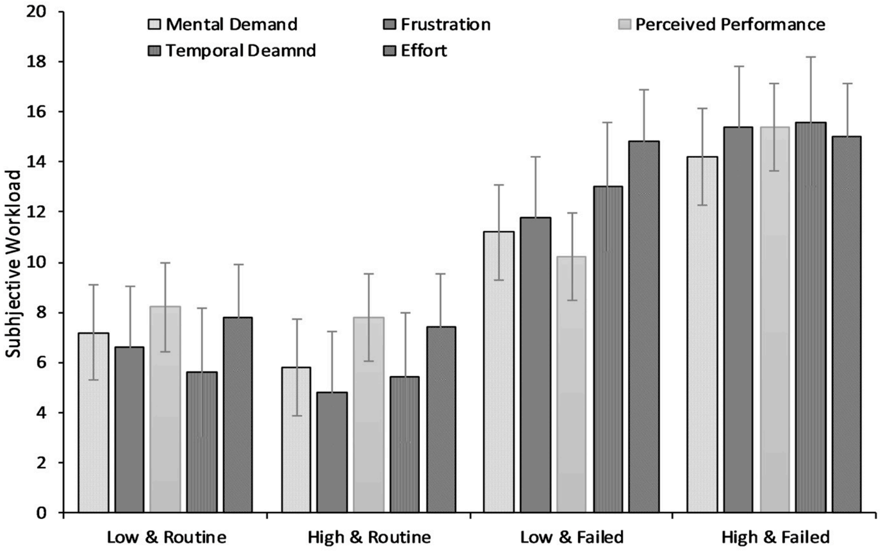


Figure 5. NASA TLX subscale scores by levels of automation.

The correlation between subscales and overall scores shows that the subscales are highly interrelated, while perceived performance is inversely correlated to other subscales and overall workload. The overall workload was highly correlated with mental demand and temporal demand (r = 0.78 and r = 0.63, respectively). In addition, temporal demand and mental demand were highly correlated (r = 0.65).

Participants reported a perceived trust in automation level as expected or higher than expected for 84% with the remaining 16% reported as lower than expected (see Table 3). Perceived trust is well correlated with participants' subjective workload scales (r = 0.79), and they also reported poorer perceived performance with unexpectedly medium (as expected) perceived trust on automation ($p < 0.01$). Frustration differed statistically across the three difficulty expectation levels, with frustration lowest when the perceived trust level was higher than expected ($p < 0.001$).

4. Discussion

4.1. Effects of Automation on Task Performance

The results showed that the automation levels marginally affected both performance measures while automation statuses moderately affected both task completion time and accuracy of proofreading tasks. Comparing task performance in Routine status, SA ratios (task accuracy) were more increased than ST ratios (task completion time) by a high level of automation (see Figures 2 and 3). Although this comparison needs to consider the ceiling and flooring effects, the sub-functions of the automation task may lead to different performance outcomes. The task used in this study comprises cognitive functions such as detecting a typographical error, correcting the error, and checking the spelling, and physical functions such as typing or writing the corrected word. At Low automation, the detecting function was automated, while at High automation, both detecting and correcting functions were automated. We observed that as the automation level increased, more cognitive functions than the physical functions are automated. Cognitive functions are vulnerable to the mistakes that are relevant to task accuracy, while physical functions easily commit errors that worsen the task completion time such as lapses, and slips [27,28]. Thus, we could infer that SA ratios are more enhanced than ST rations by the Routine automation.

Contrastingly, the almost "flat" pattern of the performances at different automation levels was observed in the Failed status. As shown in Figures 2 and 3, the average task performance in both Low and High levels of Failed automation was not significantly different. These performance patterns are closely associated with the lumberjack hypothesis suggested by Onnasch et al. [3]. According to the hypothesis, the performance follows a "flat" function up to a certain critical point regardless of the automation level in Failed automation. Therefore, the task performance in this range of Failed automation would not be affected by malfunctioned automation or unexpected automation failures, and the users could maintain their performance despite they acknowledge that the automation did not work properly.

This flat performance in Failed automation can be upheld by the skill, rule, and knowledge (SRK)-based classification. The SRK classification provides a useful framework to distinguish human behaviors or tasks based on the type of information processing demands and the different states of the constraints in working environments [27,29]. According to the SRK classification, human behaviors can be categorized into cognitive (rule/knowledge-based) and physical (skill-based) tasks [27]. Since automation is defined as the technology or the system to minimize human assistance, and it reduces information processing demands or difficult physical activities [30], rule/knowledge-based complex cognitive tasks are transformed into simple physical skill-based or rule-based tasks in Routine automation. However, Failed automation requires cognitive demanding rule/knowledge-based tasks (e.g., searching for an error). Given tasks in this study, the dexterity, knowledge, and mental resource required to complete the tasks are not complex enough to be knowledge-based tasks, and the conscious control of action with low cognitive demands (i.e., reading capability to choose the right words) and routine practice to apply the simple rule (i.e., a basic knowledge of grammar) are enough. Thus, participants can maintain their performance even in Failed automation, and this characteristic of the subtask in Failed automation results in a "flat" task performance. If the tasks depend on a higher level of automation, they will consist of more rule/knowledge-based behaviors and malfunctioned automation causes a severe decrease in the task performance.

4.2. *Effects of Automation on Overall and Subscale Workload Measurement*

While we observed a clear distinction of subjective workload between Routine and Failed automation, mixed patterns were shown by automaton levels (see Table 3). Participants perceived heavier workload in High and Failed automation than in Low and Failed, but there was no significant difference in Routine automation by different automation levels. These patterns could be interpreted as meaning that participants perceived heavy demands or more efforts are required when High level automation is failed, whereas they felt similar amounts of demands or efforts regardless of automation levels. This mixed pattern was also shown in the subscales: while all subscales in Routine automation indicated lower scores than those in Failed automation, no distinctive advantages in Low automation were reported subscale ratings. The automation levels significantly affected frustration and mental demand in both automation statues. Perceived performance and temporal demand in only Failed automation. These results suggested that the effects of automation levels in the office automation of this study were less notable than those of automation statuses.

More specifically, automation levels influenced the subscales in Failed automation. Except for effort, other subscales in High and Failed automation were higher than those in Low and Failed automation. Whereas the task performances (task completion time and task accuracy) were similar outcomes, the subscales of subjective workload were distinctively different in Low and Failed and High and Failed automations. This discrepant result may be caused by different traits of two measures.

Task performance can be measured by an empirical approach, and subjective workload refers to the portion of the individual's limited capacity and invested effort to perform a given task [31]. However, these two measures are not always well-matched. Humans tend to perceive and respond differently to the overall experience of whole-task scenarios and instant reactions [32]. Although the measuring method of the subjective workload is convenient, there are always issues as to whether any form of self-report accurately reflects respondents' "true" perceptual experiences [33,34]. To establish the validity of ratings of perceived performance, several studies suggested bringing such subjective ratings under experimental control by demonstrating their association with objective factors [33,35,36]. The gap between empirical measures and subjective self-report measures of the effects of automation is considered a potential cause of inconsistent outcomes in different automation.

The effort and perceived performance were higher than other subscales, and their scores exceeded the midpoint threshold on the NASA-TLX subscale. The midpoint threshold has been applied to access an unsustainable demand in other domains [37,38]. However, these specific sustainability thresholds have not been established within HAI. Based on the results in this study, future research should focus on minimizing effort and perceived performance. These findings suggest that providing automation for decision support may aid in smooth information processing and efficient planning procedures. If the automation delivers supportive information with positive HAI, it may reduce subjective workload for operators in unexpected automation failures.

A pattern emerged from the data between trust levels and subjective workload. Reported workload differed significantly for the participants when there was a deviation from the expected trust level. Specifically, when the trust level was lower than expected, all NASA-TLX subscales but frustration was significantly higher than cases that were rated at or higher than the expected trust level. While goal expectation has been studied in education and training and acknowledged as contributing to workload demand and workload variability, the impact of trust on task demand has not been quantified [39].

These contrasting results of subscales indicate that individual differences in operator performance are considered another cause of varied task performance in automation [40]. Although it has been adequately discussed in a wide range of psychology, sociology, human factors, and human–computer interaction literature [41,42], the effect of individual difference in human task -performance in automation has not been clearly identified. Even the same automation cannot uniformly affect task performance in different operators. In particular, due to the variance in information-processing ability and working-memory capacity, irregular patterns of task performance and variable degrees of situational awareness are established [43,44].

5. Limitations and Future Direction

The following are the potential limitations of the current study: First, the levels of automation in this study require further diversification. We designed the automation tasks with only two levels, based on the readability of the sentences. However, two levels may not be enough and too simple to describe a wide range of HAI variation. Therefore, the automation levels should specifically be defined not from the system design perspective, but from the user operation perspective. One viable user-oriented approach is Parasuraman's four stages of automation [4]. Since information processing can be acknowledged as continuous and sequential, rather than discrete processing, designing the distinctive levels of automation will be challenging. However, verifying the automation effects on task performance in more detailed and explicit levels or degrees of automation should be considered.

Second, the results of the study are limited to simple office automation and the task performance was evaluated only in terms of time and accuracy. Despite the advantage and value in the application of office automation, it does not have enough variability and flexibility in terms of the changes in the level or degree to evaluate the effects of the automation. More specified automation levels or degrees would provide detailed performance measures, so that the effects and adoption patterns of new automated functions can be identified. In addition, the task completion time and accuracy cannot be a complete set of task performance measures. Additionally, the performance indicators including relatively objective criterion measures, such as job knowledge tests and production rates, the dimensionality of job performance, and validity estimates against task performance (e.g., ability, personality, etc.) need to be considered [45].

Third, this study compared task performance with subjective workload measures in automation task performance. However, the two groups of measures were evaluated by different data collection approaches. While task performance measures gather data during the experiments, subjective workload measures collect attitudes or perceptions using a post-hoc survey questionnaire. However, the different data collection approaches may hinder a direct comparison between the two types of measures. Therefore, to minimize the possible biases from the different types of data, this study used standardized values for task performance and the Likert scale for subjective workload measures. These corrective approaches make it possible to compare the measures indirectly.

To overcome the aforementioned limitations, several further studies are suggested: first, the automation tasks should incorporate wider types of automation with specific definitions of the type, degree, and level of automation. Second, the experiment requires sufficient time to cover the patterns of adoption and learning. Such an experimental design will enable researchers to investigate task performance changes between adoption, boredom, and fatigue as the automation is extended. It will also allow researchers to investigate the changes in trust or frustration over time. Last, another task performance measuring approach can be considered, for instance, physiological measures by various sensors may supplement the objective and subjective evaluation of task performance. Since the concept of task performance seems to be difficult to define, physiological measures, such as heart rate or skin conductance, can provide additional information to aid the evaluation of the effects of automation on task performance.

6. Conclusions

In this study, we evaluated both task performance and subjective workload by automation levels and statuses to validate the automation trade-off in office automation. The automation trade-off between benefits and cost of automation is critical to understanding the overall operator performance and to develop complicated HAI design constraints. The automation tasks in this study were performed by two levels of automation (i.e., Low and High) and two automation statuses (i.e., Routine and Failed). Both task performance measures (task completion time and accuracy) showed clear benefits of automation level in Routine automation, while no significant effects of automation level were reported in Failed automation. The type of sub-functions possibly contributes to the "flat" performance in Failed automation, and the task classification may support understanding of the effects of the automation trade-off. Furthermore, task accuracy exhibited more advantages of automation than task completion time in Routine automation. The subjective workload by NASA TLX showed that higher workload was computed in High and Failed level automation that in Low and Failed automation, while mixed outcomes were shown in the subscales. The results provided important implications for future automation studies by: (1) considering the structure and the features of the sub-functions or sub-tasks in automation, (2) suggesting strategies for smooth and successful adoption and prevention of unexpected failure, and (3) providing valuable insights into the HAI system design considering operators' subjective workload and trust in automation.

Author Contributions: Conceptualization, B.C.L., H.J., J.P. (Jangwoon Park) and J.P. (Jaehyun Park); Methodology, H.J. and J.P. (Jangwoon Park); Validation, B.C.L. and J.P. (Jaehyun Park); Formal Analysis, B.L. and H.J.; Resources, J.P. (Jaehyun Park); Data curation, H.J. and J.P. (Jangwoon Park); Writing Original Draft Preparation, B.C.L.; Writing Review and Editing, B.C.L., H.J., J.P. (Jangwoon Park) and J.P. (Jaehyun Park); Visualization, B.C.L.; Supervision, J.P. (Jaehyun Park); Funding Acquisition, J.P. (Jaehyun Park). All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Incheon National University Research Grant in 2019 (Grant No.: 20190408).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hancock, P.A. Automation: How much is too much? *Ergonomics* **2014**, *57*, 449–454. [CrossRef] [PubMed]
2. Hancock, P.A. *Mind, Machine and Morality: Toward a Philosophy of Human-Technology Symbiosis*; CRC Press: Boca Raton, FL, USA, 2017.
3. Onnasch, L.; Wickens, C.D.; Li, H.; Manzey, D. Human performance consequences of stages and levels of automation: An integrated meta-analysis. *Hum. Factors* **2014**, *56*, 476–488. [CrossRef] [PubMed]
4. Parasuraman, R.; Sheridan, T.B.; Wickens, C.D. A model for types and levels of human interaction with automation. *IEEE Trans. Syst. Man Cybern. Part Syst. Hum.* **2000**, *30*, 286–297. [CrossRef] [PubMed]
5. Endsley, M.R.; Kaber, D.B. The use of level of automation as a means of alleviating out-of-the-loop performance problems: A taxonomy and empirical analysis. In Proceedings of the 13th Triennial Congress of the International Ergonomics Association, Tampere, Finland, 29 June–4 July 1997; Finnish Institute of Occuptational Health Helsinki: Helsinki, Finland, 1997; Volume 1, pp. 168–170.
6. Kaber, D.B.; Endsley, M.R. Out-of-the-loop performance problems and the use of intermediate levels of automation for improved control system functioning and safety. *Process Saf. Process* **1997**, *16*, 126–131. [CrossRef]
7. Sheridan, T.B.; Verplank, W.L. *Human and Computer Control of Undersea Teleoperators*; Massachusetts Institute of Technology Cambridge Man-Machine Systems Lab: Cambridge, MA, USA, 1978.
8. Manzey, D.; Reichenbach, J.; Onnasch, L. Human Performance Consequences of Automated Decision Aids: The Impact of Degree of Automation and System Experience. *J. Cogn. Eng. Decis. Mak.* **2012**, *6*, 57–87. [CrossRef]
9. Navarro, J.; François, M.; Mars, F. Obstacle avoidance under automated steering: Impact on driving and gaze behaviours. *Transp. Res. Part F Traffic Psychol. Behav.* **2016**, *43*, 315–324. [CrossRef]
10. Wickens, C.D.; Dixon, S.R. The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theor. Issues Ergon. Sci.* **2007**, *8*, 201–212. [CrossRef]
11. Sebok, A.; Wickens, C.D. Implementing lumberjacks and black swans into model-based tools to support human—Automation interaction. *Hum. Factors* **2017**, *59*, 189–203. [CrossRef]
12. Sheridan, T.B.; Parasuraman, R. Human-automation interaction. *Rev. Hum. Factors Ergon.* **2005**, *1*, 89–129. [CrossRef]
13. Sarter, N. Investigating mode errors on automated flight decks: Illustrating the problem-driven, cumulative, and interdisciplinary nature of human factors research. *Hum. Factors* **2008**, *50*, 506–510. [CrossRef]
14. Mosier, K.L.; Skitka, L.J.; Burdick, M.D.; Heers, S.T. Automation bias, accountability, and verification behaviors. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*; SAGE Publications: Los Angeles, CA, USA, 1996; Volume 40, pp. 204–208.
15. Salvendy, G. *Handbook of Human Factors and Ergonomics*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
16. Czaja, S.J. Human Factors in Office Automation. In *Handbook of human factors*; Salvendy, G., Ed.; John Wiley & Sons: Hoboken, NJ, USA, 1987; pp. 1587–1616.
17. Hart, S.G. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*; Sage Publications: Los Angeles, CA, USA, 2006; Volume 50, pp. 904–908.
18. Straker, L.; Mathiassen, S.E. Increased physical work loads in modern work—A necessity for better health and performance? *Ergonomics* **2009**, *52*, 1215–1225. [CrossRef] [PubMed]
19. Endsley, M.R.; Kiris, E.O. The out-of-the-loop performance problem and level of control in automation. *Hum. Factors* **1995**, *37*, 381–394. [CrossRef]
20. Kaber, D.B.; Onal, E.; Endsley, M.R. Design of automation for telerobots and the effect on performance, operator situation awareness, and subjective workload. *Hum. Factors Ergon. Manuf.* **2000**, *10*, 409–430. [CrossRef]
21. Sarter, N.B.; Schroeder, B. Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing. *Hum. Factors J. Hum. Factors Ergon. Soc.* **2001**, *43*, 573–583. [CrossRef] [PubMed]
22. Hart, S.G.; Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in Psychology*; Elsevier: Amsterdam, The Netherlands, 1988; Volume 52, pp. 139–183.
23. Readable|Free Readability Test Tool. Available online: https://www.webfx.com/tools/read-able/ (accessed on 29 December 2019).

24. Curran, P.J.; West, S.G.; Finch, J.F. The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychol. Methods* **1996**, *1*, 16. [CrossRef]
25. Warner, R.M. *Applied Statistics: From Bivariate through Multivariate Techniques*; Sage: Thousand Oaks, CA, USA, 2012.
26. Miyake, S.; Kumashiro, M. Subjective mental workload assessment technique. *Jpn. J. Ergon.* **1993**, *29*, 399–408.
27. Rasmussen, J. Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Trans. Syst. Man Cybern.* **1983**, 257–266. [CrossRef]
28. Lee, B.C.; Duffy, V.G. The effects of task interruption on human performance: A study of the systematic classification of human behavior and interruption frequency. *Hum. Factors Ergon. Manuf. Serv. Ind.* **2015**, *25*, 137–152. [CrossRef]
29. Reason, J. *Human Error*, 1st ed.; Cambridge University Press: Cambridge, UK, 1990.
30. Groover, M.P. *Fundamentals of Modern Manufacturing: Materials Processes, and Systems*; John Wiley & Sons: Hoboken, NJ, USA, 2007.
31. Schlick, C.A.; Sievert, A.; Leyk, D. Assessing Human Mobile Computing Performance by Fitts' Law. In *Mobile Computing: Concepts, Methodologies, Tools, and Applications*; IGI Global: Hershey, PA, USA, 2009; pp. 206–224.
32. Kahneman, D. *Thinking, Fast and Slow*; Macmillan: New York, NY, USA, 2011.
33. Warm, J.S.; Dember, W.N.; Hancock, P.A. Vigilance and Workload in Automated Systems. In *Automation and Human Performance*; Routledge: Hoboken, NJ, USA; Abingdon, UK, 1987; pp. 183–200.
34. Natsoulas, T. What are perceptual reports about? *Psychol. Bull.* **1967**, *67*, 249. [CrossRef]
35. Becker, A.B.; Warm, J.S.; Dember, W.N.; Hancock, P.A. Effects of jet engine noise and performance feedback on perceived workload in a monitoring task. *Int. J. Aviat. Psychol.* **1995**, *5*, 49–62. [CrossRef]
36. Hitchcock, E.M.; Dember, W.N.; Warm, J.S.; Moroney, B.W.; See, J.E. Effects of cueing and knowledge of results on workload and boredom in sustained attention. *Hum. Factors* **1999**, *41*, 365–372. [CrossRef]
37. Mazur, L.M.; Mosaly, P.R.; Hoyle, L.M.; Jones, E.L.; Chera, B.S.; Marks, L.B. Relating physician's workload with errors during radiation therapy planning. *Pract. Radiat. Oncol.* **2014**, *4*, 71–75. [CrossRef] [PubMed]
38. Mazur, L.M.; Mosaly, P.R.; Hoyle, L.M.; Jones, E.L.; Marks, L.B. Subjective and objective quantification of physician's workload and performance during radiation therapy planning tasks. *Pract. Radiat. Oncol.* **2013**, *3*, e171–e177. [CrossRef]
39. Scaduto, A.; Lindsay, D.; Chiaburu, D.S. Leader influences on training effectiveness: Motivation and outcome expectation processes. *Int. J. Train. Dev.* **2008**, *12*, 158–170. [CrossRef]
40. Rovira, E.; Pak, R.; McLaughlin, A. Effects of individual differences in working memory on performance and trust with various degrees of automation. *Theor. Issues Ergon. Sci.* **2017**, *18*, 573–591. [CrossRef]
41. Czaja, S.J.; Sharit, J. Age differences in the performance of computer-based work. *Psychol. Aging* **1993**, *8*, 59. [CrossRef] [PubMed]
42. Cullen, R.H.; Dan, C.S.; Rogers, W.A.; Fisk, A.D. The effects of experience and strategy on visual attention allocation in an automated multiple-task environment. *Int. J. Hum. Comput. Interact.* **2014**, *30*, 533–546. [CrossRef]
43. King, J.; Just, M.A. Individual differences in syntactic processing: The role of working memory. *J. Mem. Lang.* **1991**, *30*, 580–602. [CrossRef]
44. Unsworth, N.; Engle, R.W. The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychol. Rev.* **2007**, *114*, 104–132. [CrossRef]
45. Borman, W.C.; Bryant, R.H.; Dorio, J. The measurement of task performance as criteria in selection research. In *Handbook of Employee Selection*; Routledge: Abingdon, UK, 2013; pp. 439–461.