Impact of Automation: Measurement of Performance, Workload and Behaviour in a Complex Control Environment

Nora Balfe[1], Sarah Sharples[2], & John R. Wilson[2,3]

[1] Centre for Innovative Human Systems, Trinity College Dublin, Ireland

[2] Human Factors Research Group, School of Mechanical, Materials and Manufacturing Engineering, University of Nottingham, Nottingham, UK

[3] Ergonomics Team, Network Rail, London, UK

Abstract

This paper describes an experiment that was undertaken to compare three levels of automation in rail signalling; a high level in which an automated agent set routes for trains using timetable information, a medium level in which trains were routed along pre-defined paths, and a low level where the operator (signaller) was responsible for the movement of all trains. These levels are described in terms of a rail automation model based on previous automation theory (Parasuraman, Sheridan, & Wickens, 2000). Performance, subjective workload, and signaller activity were measured for each level of automation running under both normal operating conditions and abnormal, or disrupted, conditions. The results indicate that perceived workload, during both normal and disrupted phases of the experiment, decreased as the level of automation increased and performance was most consistent (i.e. showed the least variation between participants) with the highest level of automation. The results give a strong case in favour of automation, particularly in terms of demonstrating the potential for automation to reduce workload, but also suggest much benefit can achieved from a mid-level of automation potentially at a lower cost and complexity.

Impact Statement

Research in the area of automation, and in particular in the examination of human interaction with different levels of automation, has normally been undertaken in

laboratory settings using simple tasks and naïve participants where the level of automation can be easily manipulated. This research was undertaken with expert participants using complex simulation of three ecologically valid levels of automation and provides empirical field validation of some of the results found in laboratory studies.

Introduction

Automation, defined as the performance of tasks by machines (often computers) rather than human operators (Parasuraman & Riley, 1997), continues to be deployed in various industrial settings in order to increase efficiency and reduce variability. Cited benefits include the reduction of operator workload and error coupled with a reduction in labour costs (Dekker, 2004; Hollnagel, 2001). These benefits make automation very attractive to businesses wishing to increase efficiency while reducing costs. Numerous lab-based studies in the field of human factors have been undertaken to investigate the effects of automation and these have often found the benefits to be less clear-cut than might be expected (Parasuraman & Riley, 1997). For example, situation awareness may be reduced under high levels of automation (Kaber & Endsley, 2004) and workload may be increased under abnormal circumstances (Kantowitz, 1994). The level of reliability of automation is crucial, with a level below 70% believed to be worse than no automation (Wickens & Dixon, 2007). Among other weaknesses, such as the potential for programming errors (Wickens, 1992; Wiener & Curry, 1980), automation can lack the flexibility of human operators in the face of novel situations and thus difficulties can be encountered when the designers attempt to replace human problem solving abilities with automation. Hence, automation has thus far been most successful in closed loop systems, such as manufacturing systems, but humans are likely to remain vital to system performance in open loop systems, such as are commonly found in control environments, for many years (Parasuraman & Wickens, 2008).

Rail Signalling

Rail signalling is an example of an open loop system that cannot easily be fully automated. At its most basic, rail signalling involves authorising trains to move through the rail infrastructure while ensuring separation between all trains in an area. Separation in the rail context is defined in terms of sections of track (block sections)

and under normal signalling rules only one train may occupy a section at any one time. As networks become more congested signalling systems face greater challenges in terms of performance. Decisions must be made on the ordering of trains through junctions and bottlenecks, and on the most effective management of failure situations. These challenges are particularly prevalent in the British rail network due to the complexity of the infrastructure and the congestion on key routes.

Nevertheless, automation has been present in British rail signalling systems for decades. At a basic level, the interlocking systems that ensure signallers do not set conflicting routes (i.e. prevent a second train being authorised to enter a block section) for trains can be regarded as an early form of automated decision support. Mechanical forms of interlocking have been in place since the 1800s and modern computer based interlockings still perform the same function today. Early signalling systems were controlled through sets of levers directly connected to the trackside equipment. Pulling these levers changed signal aspects or the position of points, allowing signallers to change the routes of trains and give train drivers the authority to proceed. These lever frame systems were the predominant form of signalling in the UK until the 1950s when eNtry-eXit (NX) panels were introduced. NX panels reduced the physical labour involved with signalling; the signaller simply presses buttons on the panel and the physical movement of the trackside equipment is achieved automatically. In the 1980s visual display unit (VDU) based signalling was introduced in Britain facilitating the development of more advanced decision making automation in the form of Automatic Route Setting (ARS). All three forms of signalling are still in use on the British rail network but only the modern VDU form is considered here.

Rail Automation Model

Models of levels of automation have typically been used to structure investigations into the impact of different levels of automation on key cognitive ergonomics concepts such as situation awareness (SA; e.g. Durso & Sethumadhavan, 2008; Endsley & Kiris, 1995; Kaber, Onal, & Endsley, 2000; Kaber, Perry, Segall, McClernon, & Prinzel, 2006) and workload (e.g. Kaber & Endsley, 2004; Kaber et al., 2006; Kantowitz, 1994). The levels of automation identified in the models can be used to distinguish levels of independent variables in experimental designs; if sufficient levels are defined, the effect of automation can being to be described on a continuum. The levels of automation

incorporated in this study are described in terms of the model for types and levels of automation described by Parasuraman et al. (2000). The benefit of this model over those used by other researchers (e.g. Billings, 1991; Endsley & Kiris, 1995) is the ability to discriminate levels and types of automation between four functional dimensions of Information Acquisition, Information Analysis, Decision and Action Selection, and Action Implementation. Simply describing automation systems along one continuum does not give an appreciation of the different types of automation which may be present within systems and does not allow the analysis of the impact of automation at different stages of decision making.

Parasuraman et al. (2000) provide an interpretation of how automation will vary in each of these functional dimensions. For information acquisition, a low level of automation is suggested which simply helps gather the information; a mid-level is when the automation organises the information in some form, perhaps forming priorities; and a high level is where the automation filters the information so that a full set of raw data is not provided to the operator. Lower levels of information analysis automation may involve the use of algorithms to extrapolate incoming data over time or predict, and a higher level may involve integration of input variables into a single value. Automation may assist the operator with decision and action selection, for example by using conditional logic. Parasuraman et al. (2000) proposed that the decision selection automation level increases as the automation narrows the decision alternatives. Automation of the final stage, action implementation, may be the easiest type of automation to understand or observe with the level being defined by how much physical activity is replaced by automation.

The work of Parasuraman et al. (2000) was extended during this study to generate levels appropriate to the rail signalling domain in each of the four functional dimensions. A limitation of the existing scales used to describe the level of automation in each functional dimension developed by Parasuraman et al. is that they combine the functional dimensions, creating one scale for information acquisition and analysis and a second for decision-making and action implementation. This approach compromises some of the power of the four functional dimensions as the level of automation could differ independently in each. None of the other existing definitions of levels (e.g. Endsley & Kiris, 1995; Endsley & Kaber, 1999; Sheridan & Verplank, 1978) exactly matched the differences seen in the rail setting. Therefore four distinct scales were

established to describe automation at the four different decision making stages in the context of rail control. The levels defined for the rail environment are described in Table 1.

Information Acquisition	
None	Human gathers all information manually
Low	Information is gathered with assistance from ICT
Medium	Information gathering is shared between computer and human
High	Computer and technology provide most required information
Full	All information collected automatically
Information Analysis	
None	Human analyses all information
Low	Basic analysis to identify immediate control requirements
Medium	Identification of control requirements and basic prediction of future states
High	Identification of control requirements and advanced prediction of future states
Full	Full predictive analysis performed using all required data
Decision and Action Selection	
None	Human makes all decisions
Low	Computer provides decision support to help ensure decisions are safe
Medium	Computer uses basic rules to make decisions between competing demands
High	Computer makes complex decisions between competing demands under normal circumstances
Full	Computer makes complex decisions under all conditions
Action Implementation	
None	Human augments all actions
Low	Computer augments humans' physical labour
Medium	Computer implements any actions not requiring a decision
High	Computer implements most required actions
Full	Computer implements all control actions

Table 1: Levels of Automation in the Rail Automation Model

This model can be used to describe the three levels of automation used in this experiment; ARS, Auto-Routes, and Manual operation. The lowest level was Manual in which the participants were required to route all the trains under their control manually via the VDU interface. The next level was Auto-Routes in which specific routes could be set up on the workstation and any trains arriving at the start of an auto-route would be automatically routed along that path. Any trains planned over a different path required the signaller to cancel the auto-route and set an alternate route for that train. ARS was the highest level of automation. ARS has access to the planned timetable for all trains in the area and uses this information to set appropriate routes for trains arriving in its area of control. ARS also uses advanced algorithms to resolve any conflicts between trains. Figure 1 describes these three levels of automation in terms of the Rail Automation Model.
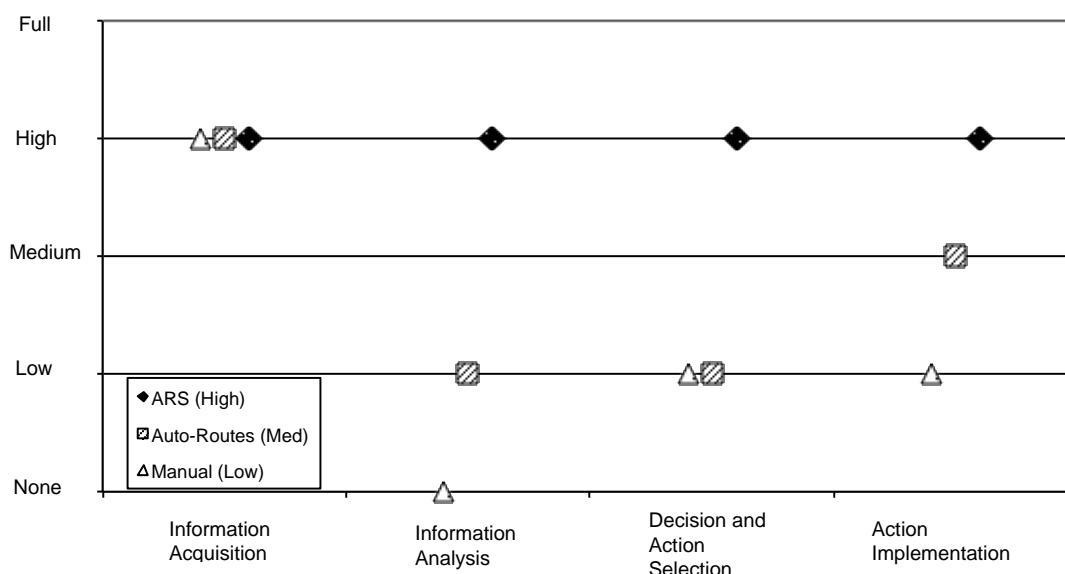
Figure 1: Levels of Rail Signalling Automation

All three LOA have high levels of information acquisition automation, as most information required is presented on the VDU. No automatic analysis of information is performed in the Manual LOA and the Auto-Routes only analyses the information to the extent of recognising a train is in the area. The ARS LOA however performs a more advanced analysis of information, comparing the relative positions and delays of trains to determine optimum routing. The interlocking provided in all signalling systems provides support for decision and action selection for all LOA, and hence Manual and Auto-routes have been assigned low levels of decision making automation. ARS provides additional automation of decision and action selection by basing the routing of trains on the outputs of its analysis, taking into account the competing demands of all the trains in the area. In terms of action implementation, the physical moving of points and signals is achieved automatically but in the Manual condition the signaller must specifically select each route for each train. Therefore this has been assigned a low level of automation. In the Auto-routes condition, any routings which do not require a specific decision are implemented automatically; hence, a medium LOA. For ARS, almost all required actions can be implemented by the automation. The main differences between the three LOA are therefore in terms of information analysis and decision and action selection, with the ARS LOA being considerably higher than both Auto-Routes and Manual, and for action implementation there are incremental raises between the three LOA. The tasks for each LOA are described in Table 2 in terms of each functional dimension.

Table 2: Tasks for each LOA

	Information Acquisition	Information Analysis	Decision and Action Selection	Action Implementation
Manual	Train Location (A) Train Identity (A)	Recognise train (M) Check route (M) Check timing (M) Check conflicts (M)	Decide route (M) Check safety (A) Decide when to set route (M)	Set Route (M)
Auto-Routes	Train Location (A) Train Identity (A)	Recognise train (A) Check route (M) Check timing (M) Check conflicts (M)	Decide route (M) Check safety (A) Decide when to set route (A)	Set Route (A)
ARS	Train Location (A) Train Identity (A)	Recognise train (A) Check route (A) Check timing (A) Check conflicts (A)	Decide route (A) Check safety (A) Decide when to set route (A)	Set Route (A)

For the remainder of this paper, the three levels of automation will be referred to as High (ARS), Medium (Auto-routes), and Low (Manual).

Effects of Automation

The study reported here follows on from previous qualitative research in the area which had found that signallers' physical workload has reduced as a result of higher levels of automation, but mental workload may have risen due to the additional need to think ahead and anticipate the actions of the automation in order to control it (Balfe et al., 2012). The same study also found that workload reduction was not achieved during disrupted train running, for example when there are train delays or infrastructure failures. A series of observations of signallers using the ARS system in the signalling environment had also been previously undertaken and found signallers engaging in different types of monitoring and signaller monitoring behaviour appeared to be driven by the quality of the automation (Balfe et al., 2008). Previous research had also suggested that the complexity of the ARS system limited the signallers' ability to understand and work cooperatively with it, with a resulting impact on performance (Balfe et al., 2012). This study considers the effect that this automation has on system performance, operator behaviour and perceived workload in VDU based signalling.

Performance in terms of mission effectiveness is a popular metric for evaluating and comparing levels of automation (Donmez, Pina, & Cummings, 2009). Without improving performance, or at least maintaining performance it is difficult to argue a benefit from automation. Data on performance are therefore an important part of evaluating an automated system. Several studies have shown that higher levels of automation result in higher levels of performance (e.g. Lorenz et al., 2002, Lee & Morey, 1992). Metzger and Parasuraman (2005) used experienced air-traffic controllers to examine the effect of conflict detection automation on operator workload and performance. This study found that automation did improve performance, but only when the automation was perfect (i.e. it correctly identified all conflicts). Research in the rail signalling environment has shown that poorly designed automation and interface design can result in reduced performance (Sandblad et al., 1997). The ability for automation to improve performance is central to its implementation and so system performance was a key variable in this research.

Well designed automation may lead to a reduction in workload; however, it is often the case that while a reduction in physical workload is achieved, there is a potential increase in mental workload for the operator (Megaw, 2005). This may be due to the need to assimilate greater quantities of information (Macdonald, 1999) or because monitoring of automation becomes burdensome (Kaber & Endsley, 2004; Warm, Dember, & Hancock, 1996). It is also the case that it is often easier to automate information acquisition and action implementation, leaving the cognitive load unchanged for operators. Automation may also lead to peaks and troughs in workload if it reduces workload during periods when workload was already low but becomes a burden during higher workload phases, a situation known as 'clumsy automation' (Wiener, 1989; Woods, 1996). Automation which assists the operator during high workload conditions is most likely to be successful (Dixon & Wickens, 2006), because this is when the largest benefits can be achieved. The ability of automation to reduce workload during normal operations is well documented in experimental studies (e.g. Kaber et al., 2000; Harris et al., 1995; Kantowitz, 1994; Röttger, Bali, & Manzey, 2009) and Kaber et al. (2006) found that the greatest workload reductions can be achieved when information acquisition and action implementation tasks are automated. This may be due to the need for operators to monitor the automation during information analysis and decision making phases and continue to make their own decisions as a basis for comparison with the automation. Kaber and Endsley (2004) also suggested that the

very act of monitoring an automated system may increase workload, perhaps due to the effort of remaining vigilant (Warm, Dember, & Hancock, 1996). Relatively few studies have directly evaluated workload and levels of automation during disruption, but Kantowitz (1994) found that automation may increase workload during incidents. The different levels of automation present in the rail signalling environment present an ideal opportunity to evaluate the effect of level of automation on workload in a complex real-world setting and to compare the results to those found in the literature.

The impact of automation on operator behaviour was also investigated in this study. Previous research (Balfe et al., 2008; Sharples et al., 2011) has noted variation in monitoring strategies during observations of signalling staff whilst using automation. Signallers varied their posture from an upright seated position, labelled 'active monitoring', to a more relaxed position labelled 'passive monitoring'. It was theorised that there was a corresponding decrease in demand from the workstation when the signaller adopted passive monitoring while interventions increased when active monitoring was more prevalent. There is relatively little existing research on operator behaviour and automation; eye-tracking data has been collected in some studies as a measure of monitoring behaviour, for example Bagheri & Jamieson (2004) studied monitoring behaviour using eye-tracking equipment and found that operators adjusted their monitoring strategy according to the automation reliability. Sarter, Mumaw & Wickens (2007) collected eye-tracking data from experienced pilots in a simulator and found that pilots do not monitor automation settings to the same degree as basic flight parameters. Metzger & Parasuraman (2006) collected eye-tracking data from experienced air traffic controllers using an automated conflict cuing system and found differences in visual attention under different workload conditions. However, these studies focussed on the specific information used by participants; Bahner, Hüper & Manzey (2008) developed an approach measuring operator sampling rates of information and used this to investigate possible automation complaceny. Röttger, Bali, & Manzey (2009) used the same approach to collect behavioural data on operator information sampling and manual interventions at different levels of automation. They found that operators reduced the frequency of information sampling and intervention when working with higher levels of automated support. This study recognised that observing operator behaviour can give valuable insight in to operator use of and interaction with automaton and the study reported here sought to further contribute to this area.

Situation awareness is frequently measured in experimental studies of levels of automation (e.g. Kaber et al., 2006; Endsley & Kaber, 1999), with higher levels of situation awareness potentially improving performance during automation failures (Kaber et al., 2000). Despite the interesting results that have emerged from lab-based studies, situation awareness was not included in this study; this was primarily due to the lack of a validated tool for measurement of situation awareness in the rail signalling environment (Golightly et al., 2012) and the need to first understand which factors should be measured with regard to situation awareness in a signalling context.

This research aimed to investigate the three dependent variables in an ecologically valid simulated setting, using a real automation system and operators who were experts in the environment and with the system. All three LOA represent methods of working which are used extensively across the UK signalling network. The experiment therefore gathered data on real world use of automation in contrast to much of the literature in the area which tends to use artificial simulations and/or non-expert participants (e.g. Lee & Moray, 1994; Sauer, Nickel & Wastell, 2013; Röttger, Bali, & Manzey, 2009; Endsley & Kaber, 1999; Bagheri & Jamieson, 2004; Beck, Dzindolet, & Pierce, 2007; Muir & Moray, 1996; Johnson et al., 2002; Meyer, Feinshreiber, & Parmet, 2003). Other studies which did use real world systems and expert operators have typically been more exploratory (e.g. Sarter & Woods, 1992; Sarter & Woods, 1994;) or examined the possible implications of a new tool (e.g. Loft, Smith, & Bhaskara, 2011; Alberdi et al, 2008) and have not directly compared different levels of automation in terms of workload, performance and operator behaviour. This experiment achieved this and also specifically examined the impact of disruption, or non-normal system operation, an aspect which is frequently omitted in the existing body of research (Sauer, Nickel, & Wastell, 2013). During disruption, ARS cannot be relied upon to make a correct decision, due to limitations in the algorithms that cannot account for all circumstances or for changes in the timetable since the algorithms were designed, and therefore this system represents that of an 'imperfect' automation system (Wickens & Dixon, 2007).

The hypotheses tested in this experiment drew on the existing findings in the literature, as well as the qualitative research previously conducted. The following hypotheses were proposed:

1. Overall workload will be reduced as level of automation increases and will increase for all LOA following the introduction of disruption
2. The support in setting routes, and consequent lower workload, will lead to higher levels of performance in the automated conditions
3. Higher levels of passive monitoring and quiet time will be observed under higher levels of automation while higher levels of active monitoring will be observed during the Manual condition to support Information Analysis

Method

Participants

Six participants took part in this study. All were male signallers from a large London based signal box. Participation was arranged in advance, although it proved extremely difficult to procure signallers for the experiment due to staff shortages in the signal box in which the experiment was conducted. For this reason, the number of participants was limited to six. The participants had a minimum of five years experience in the signal box and thus were expert signallers with familiarity of both the signalling area and the automated systems used in the experiment.

Apparatus and Materials

Simulator

Figure 2 shows the simulator used for this experiment. This simulator is typically used for training new recruits and to assess and refresh existing signallers. Although not an exact physical replica of the real signalling workstation, this simulator functions in an identical manner to a real workstation and has the same number of screens and identical input devices (i.e. trackerball and keyboard). The simulator gives a percentage score based on performance compared to the timetable.

Figure 2: Signalling Simulator

<u>Integrated Workload Scale</u>

The Integrated Workload Scale (Pickup et al, 2005) was used to measure participants' perceived workload. This is a nine-point scale developed specifically to measure perceived mental workload in the signalling environment. A high score on the IWS indicates a high workload. Pickup, Wilson, Norris, Mitchell, and Morrisroe (2005) report that this tool has proven to be a valuable measure of peaks and troughs in workload over time or within a set of scenarios. They also report that the tool is acceptable to signallers, having been developed specifically for use in the signalling environment, and maps well onto expected workload measured using other techniques. It was constructed using the Thurstone technique and so the ratings can be used as interval data. Participants were provided with a laminated copy of IWS and asked to verbally rate their workload on this scale at 2min intervals throughout the experiment.

Design

Table 2 shows the experimental design. A part-counterbalanced repeated measures design was used in which three LOA were examined; High (ARS), Medium (Auto-routes), and Low (Manual). Each condition lasted for 30min and used the same

scenario based on the same section of the timetable. After 15min disruption was introduced.

Order	Group A			Group B			Group C		
1st	High	Normal		Low	Normal		Medium	Normal	
		Disrupt			Disrupt			Disrupt	
2nd	Medium	Normal		High	Normal		Low	Normal	
		Disrupt			Disrupt			Disrupt	
3rd	Low	Normal		Medium	Normal		High	Normal	
		Disrupt			Disrupt			Disrupt	

Table 3: Study Design

In order to balance the potential learning effect the participants completed the three conditions in different orders. However, a learning effect was not anticipated as the participants were expert signallers who operate this timetable and area on a daily basis and are competent to deal with any disruption which may occur.

A form of disruption was also introduced half way through the experiment, meaning there were two levels of disruption, normal and disrupted. Choice of disruption was a key part of the experimental design as a noticeable effect on workload was required to understand whether the impact of automation was different for different levels of demand. Many disrupted conditions on the railway involve a high degree of communication and/or knowledge and application of the rules. It was necessary to control communications as far as possible to ensure that they did not affect the results as communications can contribute to increased workload and may have masked workload effects of the automation. It was also desirable to avoid application of the rules as this held ethical concerns in the event of mis-application by any participant. For these reasons, the selected form of disruption was closure of a section of track, which was a platform at a busy station. The participants were required to route trains around the closed platform and regulate this change to the service.

The disrupted condition was always second in the experiment; the participants encountered 15 min of normal running and then 15 min of disruption. It was not possible to vary the order as disruption has consequential effects and even if the

platform had reopened, the signaller would still be required to regulate the resulting delays.

The Independent Variables for the experiment were:
- Level of Automation
- Level of Disruption

The Dependent Variables for the experiment were:
- IWS Scores (perceived workload)
- Signaller Behaviour
- Performance Scores (generated by the simulator)

<u>Behaviour Coding</u>

The participants were observed during each scenario to note their activity. The coding scheme described by Sharples et al. (2011) was adapted for this experiment. The coding was based around the five basic codes of Monitoring, Interaction, Planning, Communicating, and Quiet Time. As in Sharples et al. (2011), two forms of monitoring were coded, active and passive monitoring. Active monitoring was coded when the participant was sitting up while viewing the signalling screens. Passive monitoring was coded when the participant was sitting back while viewing the signalling screens. Interaction was coded when participants used the trackerball and the purpose of each interaction was also coded; interactions coded included route setting and cancelling, auto-routes set-up, and use of reminders. Reminders are a device that can be placed on a piece of infrastructure to prevent it being used. They are frequently applied to signals on the VDU to prevent a route being set to or from that signal. Quiet time was coded whenever the participant was not involved in any aspect of the signalling task. Due to the simulated environment, planning and communication activities were limited and no analysis was performed on these. Data were coded live using a software package.

Procedure

Three researchers were used to gather the data during the experiment. The first researcher used a laptop to code signaller behaviour in real time. The second

researcher administered the verbal IWS and was responsible for timekeeping. The third researcher, a signalling subject matter expert (SME), sat in the adjoining room and gathered performance data. The SME also handled any communications with the participants required as part of the experiment. The experimental set-up is shown in Figure 3.
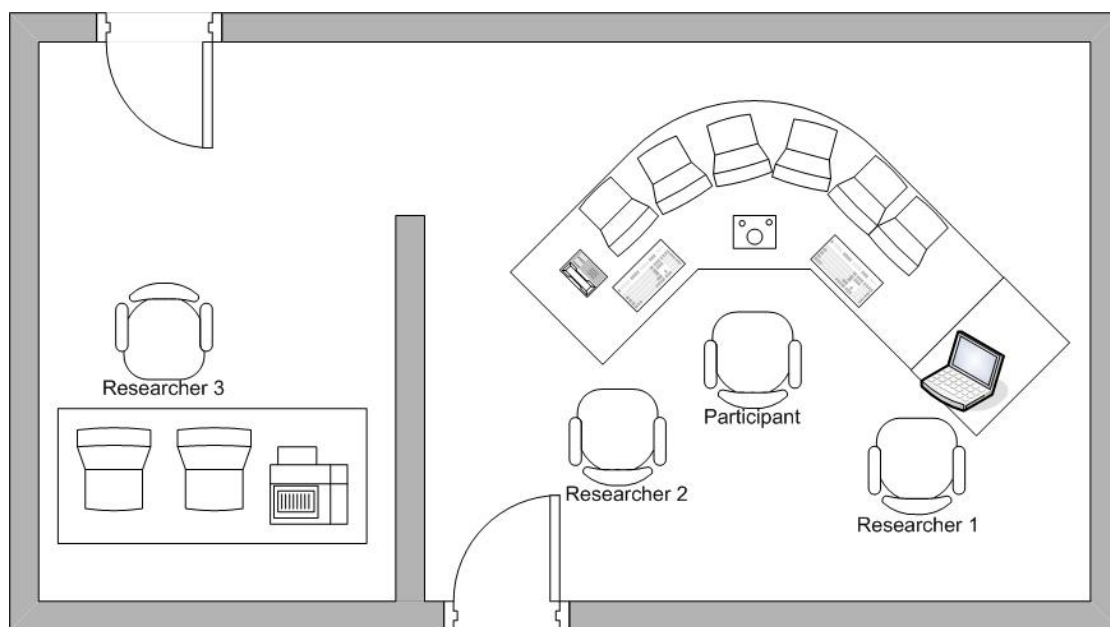

Figure 3: Experimental Set-Up

The participant was invited into the simulator room and the experiment was explained to him. He was asked to read the briefing sheet and sign the consent form. The participant then took his place at the simulator and the experiment began. At the mid-point of the experiment, the third researcher announced the closure of the platform to the signaller. The remaining half of the experiment was therefore under disrupted conditions. At the end of the experiment the simulator was paused and the performance data collected. The same procedure was followed for the second and third scenarios for which the level of automation was changed according to the group to which each participant was assigned.

Results

Workload

Participants were asked to verbally rate their workload on the IWS Scale every 2min. The results are presented here as a graph showing the average workload scores for each LOA at each 2min interval.
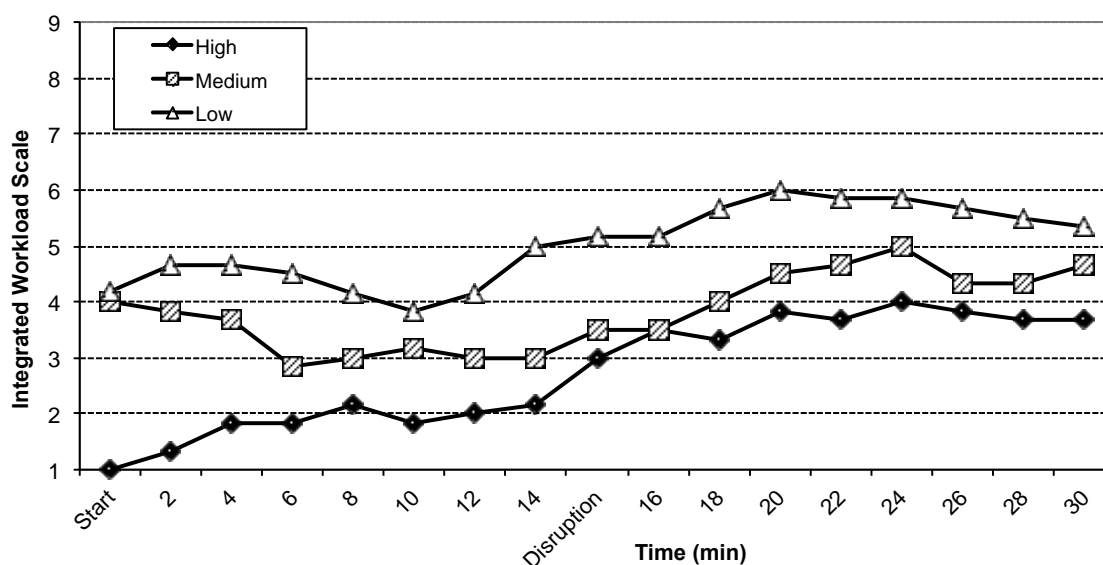

Figure 4: Mean IWS Scores for Each LOA

It is clear from Figure 4 that the High LOA was consistently rated lowest and the Low LOA was consistently rated highest. The Medium LOA initially showed increased workload scores which quickly tapered off. This corresponds with the need to set up the auto-routes at the beginning of the scenario. Once these were established the workload fell and remained reasonably consistent until the disruption was introduced. All three LOA showed an increase in perceived workload following the introduction of disruption, with the High condition showing the steepest increase. High automation workload scores increased from a mean of 1.77 before disruption to a mean of 3.69 after disruption, almost 2 full workload points. The Medium automation condition increased by just over 1 workload point, from 3.31 to 4.37, and the Low automation condition by 1.22 workload points from 4.40 to 5.52. The steep increase for the High condition corresponds with the signaller setting up protection around the affected area to prevent any trains being routed through. Workload continues to increase following the introduction of disruption as delays begin to occur, further complicating regulation.

A 2x3 repeated measures ANOVA was run on the averaged IWS data to determine whether there were any significant differences due to LOA or disruption. A significant main effect of LOA was found ($F_{(2, 10)}$ = 31.916, $p<0.001$) and a Bonferroni post-hoc showed that this was between all levels of automation (M(High) = 2.74, SD(High) = 1.01); M(Medium) = 3.82, SD(Medium) = 0.69; M(Low) = 5.02, SD(Low) = 0.69; p(High-Med) < 0.05; p(High-Low) < 0.005; p(Med-Low) < 0.01). There was also a significant effect of disruption ($F_{(1, 4)}$ = 36.462, $p < 0.005$) and a significant interaction ($F_{(2, 8)}$ = 11.636, $p < 0.005$) with Low and High automation having a significant difference in reported level of workload after the introduction of disruption.

Performance

Figure 5 describes the simulator generated performance score of each participant for each LOA. It can be seen that performance was most consistent across participants for the High LOA. This was also consistently the highest performance, followed by the Medium condition and finally the Low condition, both of which showed more variation between participants. A repeated measures one-way ANOVA was run on these data and a significant main effect of LOA was found ($F_{(2, 10)}$ = 11.516, $p < 0.005$). A Bonferroni post-hoc test showed that the difference was between the Low automation group (M(Low) = 75.17, SD(Low) = 3.97) and both higher automation groups (M(Medium) = 81.83, SD(Medium) = 5.00, $p<0.05$; M(High) = 84.83, SD(High) = 1.94, $p<0.01$) with performance lowest in the Low condition.
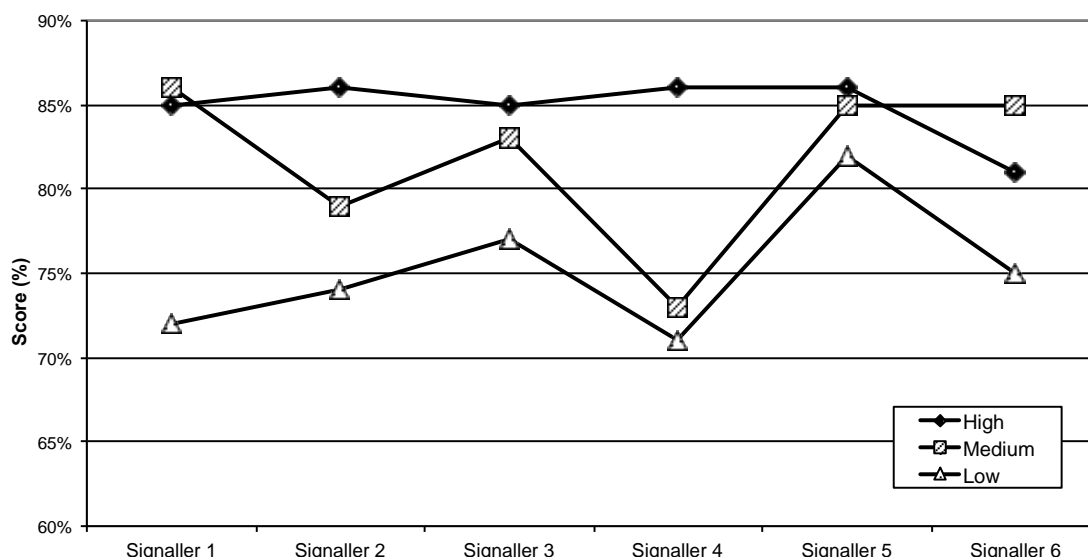
Figure 5: Performance Scores

The order of level of automation was varied for each participant to balance any potential learning effect. A learning effect was not anticipated and examining the performance data in terms of the order in which each scenario was completed (i.e. first, second, or third) shows that none occurred. Table 4 shows the average performance score for each order.

Order	Average score
1	81%
2	81%
3	79%

Table 4: Average performance score for each order of condition

Signaller Activity

The results of the behaviour observation are presented in the following sections.

Monitoring

Figure 6 shows the mean time dedicated to active monitoring for all conditions.
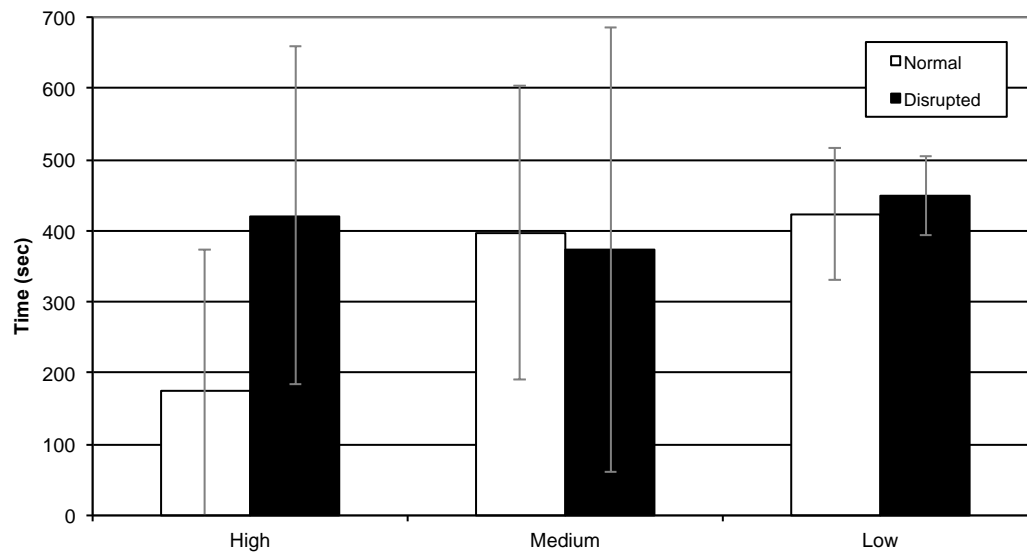
Figure 6: Mean Active Monitoring

A 2x3 repeated measures ANOVA found no significant main effect of level of automation or disruption. There was a significant interaction ($F_{(2, 10)} = 7.713$, $p < 0.5$). Observation of the data shows that active monitoring was lowest during the High LOA under normal running. The high standard deviation (SD) for High and Medium LOA indicate that active monitoring was highly variable between participants for those two conditions but was much more stable during the Low LOA.

Passive Monitoring

Figure 7 describes the mean time dedicated to passive monitoring for all conditions.
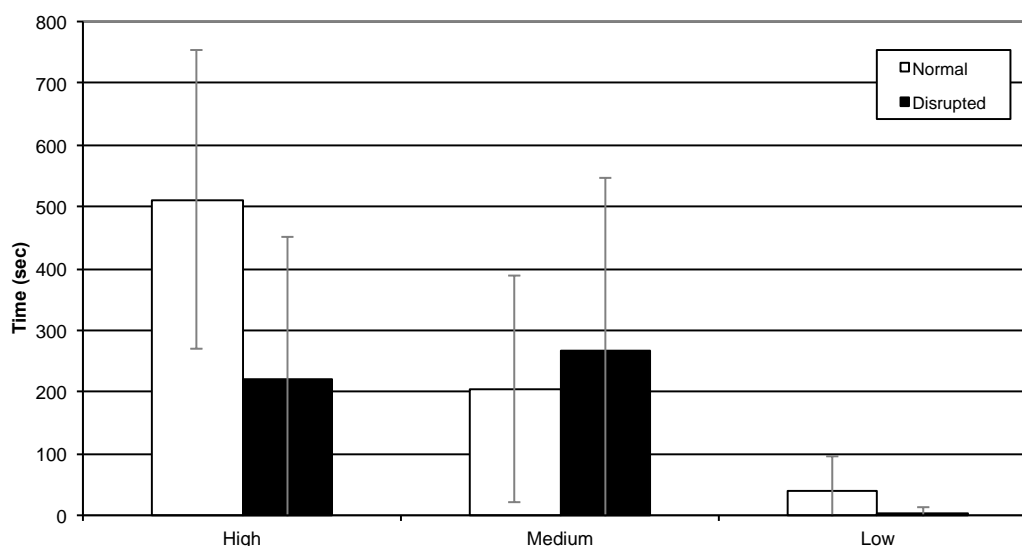
Figure 7: Mean Passive Monitoring

A 2x3 repeated measures ANOVA revealed a significant main effect of LOA for passive monitoring, (F = 11.762 (2, 10), p<0.005). Bonferroni's post-hoc comparison revealed this difference was between the High LOA and both lower automated conditions (M(High) = 365.17, SD(High) = 272.95; M(Medium) = 235.58, SD(Medium = 222.74, p< 0.05; M(Low) = 21.5, SD(Low) = 43.29, p < 0.05). A significant interaction was also found (F(2, 10) = 17.157, p < 0.001) demonstrating the particularly strong impact of disruption on the high LOA but no signification effect of disruption was found. Again, high standard deviations were found for the High and Medium LOA as compared to the Low condition.

Intervention

Figure 8 describes the use of the trackerball during the experiment. As would be expected, use was highest during the Low automation condition as participants used the device to set routes for trains in the area. The SD were reasonably small for all conditions, suggesting that intervention was driven by the scenario rather than individual preference.

A 2x3 repeated measures ANOVA found a significant effect of automation (F(2, 10) = 92.050, p <0.001) and a significant interaction (F(2, 10) = 8.358, p < 0.1). A Bonferroni post-hoc test showed that the Low LOA was significantly different to both higher levels

(M(Low) = 388.92, SD(Low) = 45.95; M(Medium) = 184.75, SD(Medium) = 59.76, $p <$ 0.001; M(High) = 107.75, SD(High) = 66.58, $p <$ 0.001).
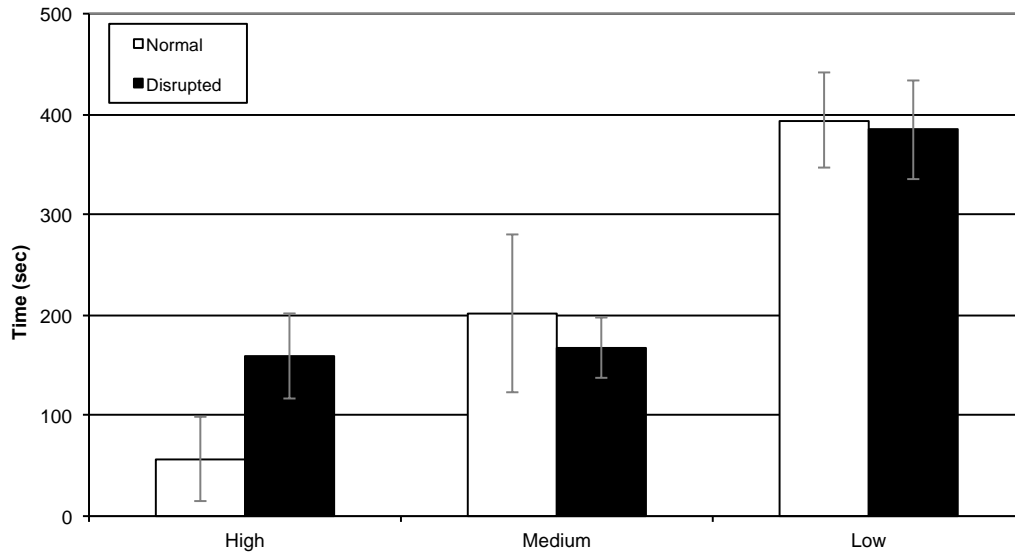

Figure 8: Use of Trackerball

The purpose of the trackerball interventions was also coded. Figure 9 describes the mean and standard deviation of the different types of intervention coded for each LOA.
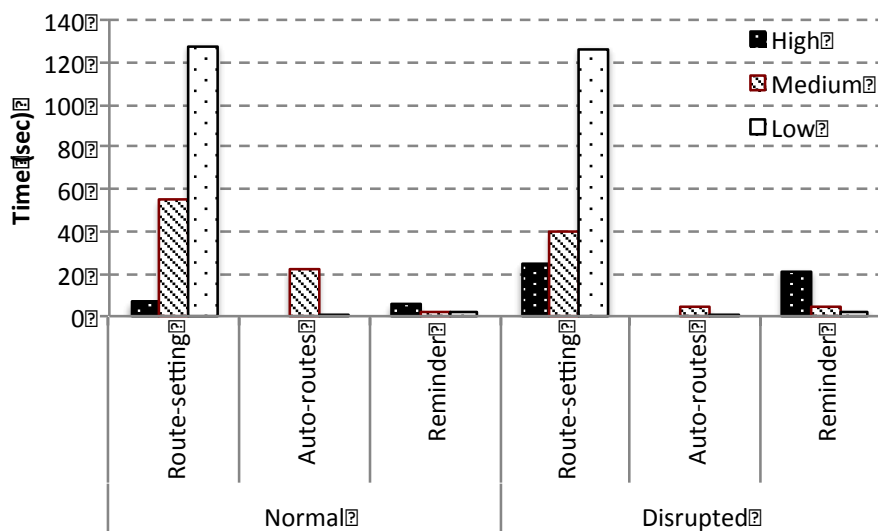

Figure 9: Mean of Intervention Types

A repeated measures two-way ANOVA was run on the data for each dependent variable. A significant effect of level of automation was found for route-setting (F(2, 10)

= 54.055, p<0.001) and a Bonferroni post-hoc showed that this was between all levels of automation (M(High)=15.667, SD(High)=2.315, M(Medium)=47.0.83, SD(Medium)=4.913, p<0.005; M(High)=15.667, SD(High)=2.315, M(Low)=127.167, SD(Low)=11.861, p<0.001; M(Medium)=47.0.83, SD(Medium)=4.913, M(Low) = 127.167, SD(Low) = 11.861, p<0.005). No significant effect of disruption was found for route-setting, nor a significant interaction. A significant effect of level of automation was also found for reminders (F(2, 10) = 76.181, p<0.001) and a Bonferroni post-hoc showed that this was between the High LOA and the two lower automation levels (M(High) = 13.750, SD(High = 1.806); M(Medium) = 2.917, SD(Medium) = 0.908, p<0.001; M(Low) = 2.000, SD(Low) = 0.753), p < 0.001). A significant effect was also found for disruption in terms of application of reminder devices (F(1, 5) = 21.818, p<0.005). An interaction effect was also found between the two independent variables for reminders (F(2, 10) = 21.602, P<0.001). Auto-routes were not further analysed; as these were only used in the Medium LOA there was obviously a difference between this condition and the others.

Quiet Time

Figure 10 illustrates the average quiet time observed during each condition. The high standard deviations indicate considerable variation between participants, but a repeated measures 2x3 ANOVA found only a statistically significant interaction (F(2, 10) = 5.009, p < 0.05). Observation of the data revealed that quiet time was highest during the High automation normal running condition, but was almost equivalent to the Medium condition during the disrupted condition. Quiet time was lowest for the Low automation condition. All three conditions showed high standard deviations for this behaviour.
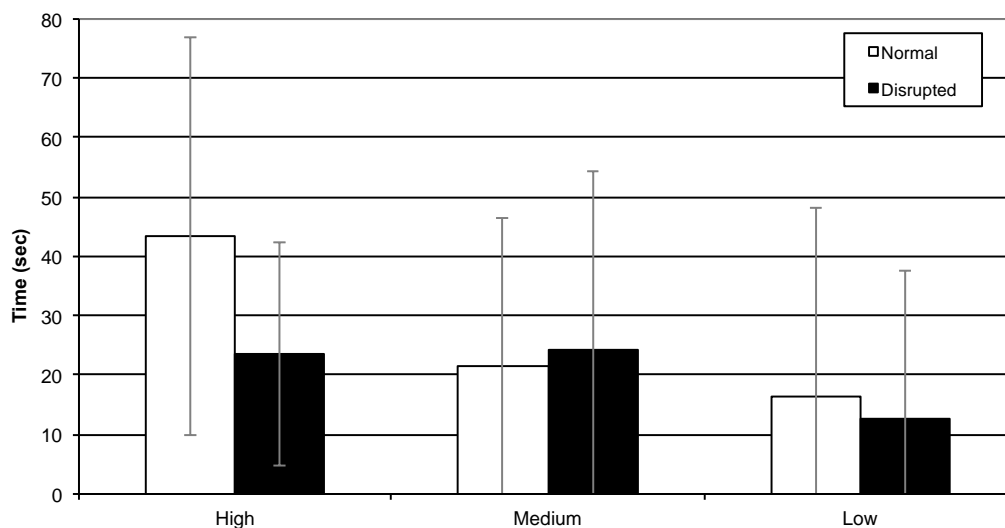
Figure 10: Mean Quiet Time

Discussion

Workload

Workload was expected to reduce as LOA increased, and this was found to be the case. A step change can be seen as the level of automation increases, and this is reflected in the significant differences found for all levels of automation. Automation may reduce workload within any of the four functional dimensions and previous research has indicated that automation is most successful at reducing workload when applied to the information acquisition and action implementation functional dimensions (Kaber et al., 2006). The three LOA are identical for the Information Acquisition functional dimension so the differences in workload must be within the remaining three. Similarly, the Medium and Low conditions were assessed as having the same level of decision-making automation, leaving only the information analysis and action implementation dimensions to explain the difference in workload for these two. The change in workload appears to closely match the step change seen in the Action Implementation functional dimension. This corresponds with existing theories in the literature (Megaw, 2005) but would imply that the main effect of workload was due to physical workload. However, the Low condition showed no increase in physical work (i.e. trackerball usage) following disruption despite the rise in workload ratings, so we can deduce that the perceived workload ratings were measuring both physical and

mental workload. Further evidence is provided by the High condition, which saw an over 200% rise in workload rating following disruption, but a 178% rise in trackerball usage. The difference in workload scores can therefore be taken to reflect both physical and mental workload decreasing as the level of automation increases, suggesting that higher levels of automation can also reduce workload effectively in the information analysis and decision making functional dimensions.

A steep increase was seen in the High automation condition when disruption was introduced and the signaller began to apply reminder devices. It was also necessary to apply these reminders in the other LOA (Medium and Low), but the same steep increase in workload is not seen on those graphs, despite the much higher levels of physical interaction. It is proposed that the increase seen on the High LOA graph represents the signaller becoming more involved in the signalling and task and being required to process more information to maintain situation awareness. When the disruption was introduced in the High automation condition the signallers suddenly had to become involved with manually routing trains around the blocked section of track. The same strategy was employed in the two lower automation conditions but as the signallers would already have been manually routing all trains in the Low condition, and some trains in the Medium condition, the increase in physical activity was not so steep. The High condition still showed the lowest workload scores overall, but this sharp increase is characteristic of the peaks and troughs associated with 'clumsy automation' (Wiener, 1989; Woods, 1996) and may represent the mental effort required to more fully engage with the system and attain an appropriate level of situation awareness.

Performance

The hypothesis in this area was that performance would increase with higher levels of automation and the results indicate that performance was significantly improved with the assistance of automation with High automation showing the most consistent performance across all signallers. However, two participants achieved higher performance under the Medium condition illustrating that high levels of automation do not always result in optimum performance for an individual. The hypothesis therefore cannot be fully accepted from this data. However, a major benefit from automation may be in its ability to achieve consistency between operators resulting in reliably high levels of system performance. This is in contrast to limited resources theory which

would suggest that performance should be improved by automation only if the task exceeds the capacity of the human operator (Wickens, Hollands, Banbery, & Parasuraman, 2013). However, the complexity of the rail signalling task means that, although within operators' capacity to perform the task optimally, there is the potential for task cues to be missed or wrongly prioritised, resulting in a reduction in performance. Automation can assist in both these areas by routing trains in areas of the workstation that may not be the current focus of the operators' attention thus leading to more consistent performance.

Unfortunately, it was only possible to measure performance for the whole experiment so the difference in performance between normal and disrupted conditions could not be investigated. It is possible that the performance decrement in the Low condition came from the time taken to set routes manually rather than reduced quality of decision making, i.e. it was related to physical workload rather than mental workload. There was not a significant difference between the High and Medium LOA, which suggests that performance may be, at least partially, influenced by the high level of physical workload involved in route setting in the manual condition rather than the mental workload involved with the task.

Behaviour Observation

Monitoring

When the two levels of monitoring (active and passive) were differentiated in previous studies (Balfe et al., 2008; Sharples et al., 2011) it was noted that active monitoring was associated with interventions; this study provides further evidence that active monitoring has a strong association with route setting, as high levels of active monitoring were sustained by all participants in the Low automation condition. Active monitoring in the High automation condition also rose to comparable levels with both other conditions after the introduction of disruption. This was as the participants became more involved with route setting around the blockage and further indicates the link between active monitoring and interaction. It can therefore be proposed that high levels of active monitoring is indicative of an increase in mental workload, and in information analysis and decision-making (as the information must be analysed and decisions made before any route can be set). Active monitoring in the High condition

rose to levels comparable with the two lower conditions after disruption was introduced, which may indicate that mental workload also rose to similar levels.

Passive monitoring was almost exclusively confined to the higher automated conditions, and a sharp reduction could be seen in the High automation condition when disruption was introduced. Interestingly, the average passive monitoring observed in the Medium condition actually rose following disruption but this was only a small rise and is likely to have been due to chance. The Medium and Low conditions were the most stable in terms of observed monitoring activity between the normal and disrupted conditions, in that there was very little passive monitoring and a high degree of active monitoring.

The results from this study suggest that signallers regulate their workload by engaging in passive monitoring. Passive monitoring may require a lower level of attention, and hence places less demand on the signallers. When the circumstances on the workstation became more demanding, signallers reduced their passive monitoring and engaged in more active monitoring. The high standard deviations also suggest that passive monitoring is a choice for signallers, with some choosing to monitor the system at a low level (passive monitoring) while others remain more actively involved or take a break completely (quiet time).

<u>Intervention</u>

The number of observed interventions was found to be significantly different between automated conditions. This is unsurprising as the requirement to set routes manually in the Low automation condition would have greatly increased the number of interventions in that condition. Even during disruption, the High automation condition was significantly lower than the Low condition as ARS continued to set routes for trains in the unaffected parts of the workstation. This demonstrates the potential value of automation in non-routine conditions. However, the intervention levels under disruption were comparable for the High and Medium conditions, indicating that the physical workload associated with route setting can be similarly reduced by a lower level of automation.

Four significant differences were found between LOA for types of intervention. Most of these differences are as expected. There was a difference between all groups for route setting, since the requirement to set routes increased as the level of automation decreased. Auto-routes were not used in the High and Low LOA so there are obviously significant differences between the Medium LOA and the others due to their use in only the Medium LOA. There was a significantly greater use of reminders in the High LOA, which is interesting as they were also required in both other conditions to protect the platform area following the introduction of disruption. The significant increase in the use of reminders in the High condition reflects their use as a control mechanism for ARS; they are primarily intended as safety appliances to prevent trains being routed over a section of track and thus they effectively constrain ARS. Previous research had already suggested that they are frequently used to constrain ARS to improve system performance when there are no safety concerns (Balfe, 2010). However, the disruption in this experiment was not anticipated to have greatly increased the use of reminders in this context and it is interesting that the effect has appeared. This demonstrates how extremely common it is for signallers to use this strategy as an easy method to direct the automation.

Quiet Time

Quiet time was highest in the High LOA during the normal condition; however, it reduced to levels equivalent to those of the Medium LOA during disruption. A similar reduction was not seen for the Medium LOA indicating that signallers felt they had the same amount of free time in normal and disrupted running and provides further evidence of the robustness of this form of automation. It is also noteworthy that quiet time was present for the Low condition demonstrating that max capacity was not reached. Despite the demands of route setting, participants did spend a small amount of time not involved with the system. This is in contrast to passive monitoring during the Low condition in which participants rarely engaged and may indicate that participants devote passive monitoring resource during automated conditions, but do not feel it is necessary when controlling manually, presumably because they are in control of any changes. This suggests that passive monitoring is a useful activity, undertaken to maintain a level of awareness of the system particularly when automation is present.

Conclusions

This experiment has shown that high levels of automation (i.e. ARS) do lead to a reduction in workload compared to lower levels of automation, and that reduction occurs in both mental and physical workload. The reduction is not as large during disrupted running suggesting the higher levels of automation could be considered 'clumsy' (Woods, 1996). In addition, performance was highest and most consistent when working with higher automation levels. The mid-level automation showed the most stable workload scores throughout, but performance was variable between participants. In terms of the Rail Automation Model, the results suggest that workload was reduced in the Action Implementation functional dimension for the highest LOA, but that workload was also reduced in the information analysis and/or decision-making functional dimensions.

Similar to workload, performance was improved with higher levels of automation, although it was not as clear-cut a result with some signallers achieving a higher score with the mid-level of automation. However, performance was most consistent across all signallers with the high level of automation suggesting that the real benefit in terms of performance is in consistently high performance.

There were obvious differences in participants' behaviours between the conditions such as increased use of the trackerball during the Low LOA and increased passive monitoring when using high levels of automation. However, the amount of time dedicated to monitoring varied between participants more for the automated conditions compared to the manual. This appears to indicate that individual signallers engage in different strategies during the automated conditions highlighting the importance of ensuring that methods to assess workload in situated tasks are able not only to measure perceived workload but also capture different strategies that may be adopted with different task configurations.

Although statistical differences were found in the results of the experiment, we acknowledge that the study did not have strong statistical power and was particularly limited by the small sample size as well as other aspects of statistical power (Baguley, 2004). Rather than focus on statistical power, the study reported here used participants with domain expertise in a naturalistic setting (Farrington-Darby & Wilson, 2006). The

results, particularly when combined with previous qualitative studies (Balfe et al., 2012), do have practical significance for the use of automation in complex control environments. The experiment was also limited by the time span simulated. Signallers would usually work for several hours and the effects of disruption may become more complex over this timeframe. Therefore, some of the more complex effects of automation during disruption may not have appeared in this experiment. A further limitation was the inability to measure the reliability of the automation in the disrupted condition. The participants took responsibility for setting routes and controlling the automation to 'force' the correct decisions but without a measure of the reliability of the automation in this condition it is difficult to generalise the results with respect to imperfect automation.

Overall the findings of this experiment support the use of automation in rail signalling and demonstrated that some of the findings of lab-based previous research hold true in real world systems. However, the advantages of high levels of automation over the mid-levels are not as great as might be assumed given the differences in complexity and cost. Future experiments could further investigate the ARS and Auto-routes LOA and involve a greater number of scenarios and participants. A more advanced simulator would be capable of giving a better analysis of performance. It is also important to develop more advanced measures of workload and performance in order to distinguish between performance decrements as a result of reaching maximum mental workload capacity and decrements due to the need to work more slowly on account of physical workload constraints. Simple measures of primary task performance are not sufficiently powerful to determine whether more physical assistance automation is required or more decision support automation is necessary to improve system performance.

The Rail Automation Model helps to abstract the results so they can be used more widely; however the range of technology employed in supporting work systems continues to expand and the current granularity of the levels of automation may not be sufficient to accurately describe subtle differences in future automation technologies. Automation systems can also vary hugely in terms of their design, not just in the type and level but also in terms of the characteristics such as degree and quality of feedback. The Rail Automation Model and the models it is based upon do not account for factors such as feedback, understandability, and interaction styles, all of which

could influence results. The results reported in this paper therefore may be generalizable to other systems but some of the results are likely to be due to the characteristics of the automation as well as the type and level. Currently there is no technique available to better represent these influences on work performance in automated systems, and this is an area which should be explored in future research.

Acknowledgements

References

Alberdi, E., Povyakalo, A.A., Stringini, L., Ayton, P., & Given-Wilson, R. (2008). CAD in mammography: lesion-level versus case-level analysis of the effects of prompts on human decisions. *International Journal of Computer Assisted Radiology and Surgery,* 3(1-2), 115-122.

Bainbridge, L. (1983). Ironies of automation. *Automatica,* 19(6), 775-779.

Bagheri, N., & Jamieson, G. A. (2004). The impact of context-related reliability on automation failure detection and scanning behaviour. In *2004 IEEE International Conference on Systems, Man and Cybernetics* (Vol. 1, pp. 212-217). IEEE.

Baguley, T. (2004). Understanding statistical power in the context of applied research. *Applied Ergonomics,* 35, 73-80.

Bahner, J.E., Hüper, A.D., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. *International Journal of Human-Computer Studies,* 66, 688-699.

Balfe, N. (2010). *Appropriate automation of rail signalling systems: A human factors study.* The University of Nottingham, Nottingham.

Balfe, N., Wilson, J.R., Sharples, S., & Clarke, T. (2012). *Development of design principles for automated systems in transport control. Ergonomics,* 55(1), 37-54.

Balfe, N., Wilson, J. R., Sharples, S., & Clarke, T. (2008). *Structured observations of* automation use. In P. D. Bust (Ed.), Contemporary Ergonomics 2008 (pp. 552-557). London: Taylor & Francis.

Beck, H. P., Dzindolet, M. T., & Pierce, L. G. (2007). Automation usage decisions: Controlling intent and appraisal errors in a target detection task. *Human Factors,* 49(3), 429-437.

Billings, C. E. (1991). *Human-centered aircraft automation: A concept and guidelines* (No. NASA Technical Memorandum 103885). Springfield, VA: National Technical Information Service.

Dekker, S. W. A. (2004). On the other side of promise: What should we automate today? In D. Harris (Ed.), *Human Factors for Civil Flight Deck Design* (pp. 183-198). Aldershot, Hampshire: Ashgate Publishing.

Dixon, S. R., & Wickens, C. D. (2006). Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. *Human Factors,* 48(3), 474-486.

Donmez, B., Pina, P. E., & Cummings, M. L. (2009). Evaluation criteria for human-automation performance metrics. In *Performance Evaluation and Benchmarking of Intelligent Systems* (pp. 21-40). New York: Springer US.

Durso, F. T., & Sethumadhavan, A. (2008). Situation awareness: Understanding dynamic environments. *Human Factors,* 50(3), 442-448.

Endsley, M. R., & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics,* 42(3), 462-492.

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors,* 37(2), 381-394.

Farrington-Darby, T., & Wilson, J.R. (2006). The nature of expertise: A review. *Applied Ergonomics,* 37, 17-32.

Golightly, D., Balfe, N., Sharples, S., & Lowe, E. (2012). Measuring situation awareness in rail signalling. In J.R. Wilson, A. Mills, T. Clarke, J. Rajan, & N. Dadashi (Eds.), *Rail Human Factors around the World* (pp. 361-369). Leiden: CRC Press.

Harris, W. C., Hancock, P. A., Arthur, E. J., & Caird, J. K. (1995). Performance, workload, and fatigue changes associated with automation. *The International Journal of Aviation Psychology,* 5(2), 169-185.

Hollnagel, E. (2001). Human-oriented automation strategies. In *Proceedings of the World Congress on Safety of Modern Technical Systems* (pp. 443-452). Cologne: TUV-Verlag, Cologne.

Johnson, K., Ren, L., Kuchar, J., & Oman, C. (2002). Interaction of automation and time pressure in a route replanning task. In *International Conference on Human-Computer Interaction in Aeronautics* (pp. 132-137).

Kaber, D. B., & Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science,* 5(2), 113-153.

Kaber, D. B., Onal, E., & Endsley, M. R. (2000). Design of automation for telerobots and the effect on performance, operator situation awareness, and subjective workload. *Human Factors and Ergonomics in Manufacturing,* 10(4), 409-430.

Kaber, D. B., Perry, C. M., Segall, N., McClernon, C. K., & Prinzel, L. J. (2006). Situation awareness implications of adaptive automation for information processing in an air traffic control control-related task. *International Journal of Industrial Ergonomics,* 36(5), 447-462.

Kantowitz, B. H. (1994). Pilot workload and flight deck automation. In M. Mouloua & R. Parasuraman (Eds.), *Human performance in automated systems: Current research and trends* (pp. 212-223). Mahwah, NJ: Lawrence Erlbaum.

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

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies,* 40(1), 153-184.

Loft, S., Smith, R.E., & Bhaskara, A. (2011). Prospective memory in an air traffic control simulation: External aids that signal when to act. *Journal of Experimental Psychology: Applied,* 17(1), 60-70.

Lorenz, B., Di Nocera, F., Röttger, S., & Parasuraman, R. (2002). Automated falult-management in a simulated spaceflight micro-world. *Aviation, Space, and Environmental Medicine*, 73(9), 886-897.

Macdonald, W. A. (1999). Train controller interface design: Factors influencing mental workload. In *Proceedings of the International Conference on Human Interfaces in Control Rooms, Cockpits and Command Centres* (pp. 31-36). London: Institute of Electrical Engineers.

Megaw, T. (2005). The definition and measurement of mental workload. In J. R. Wilson & N. Corlett (Eds.), *Evaluation of Human Work* (3rd ed., pp. 525-551). London: Taylor & Francis.

Metzger, U., & Parasuraman, R. (2005). Automation in future air traffic management: Effects of decision aid reliability on controller performance and mental workload. *Human Factors*, 47(1), 35-49.

Metzger, U., & Parasuraman, R. (2006). Effects of automated conflict cuing and traffic density on air traffic controller performance and visual attention in a datalink environment. *International Journal of Aviation Psychology*, 16(4), 343-362.

Meyer, J. L., Feinshreiber, L., & Parmet, Y. (2003). Levels of automation in a simulated failure detection task. In *2003 IEEE International Conference on Systems, Man and Cybernetics* (Vol. 3, pp. 2101-2106). IEEE.

Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics,* 39(3), 429-460.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors,* 39(2), 230-253.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans,* 30(3), 286-297.

Parasuraman, R., & Wickens, C. D. (2008). Humans: Still vital after all these years of automation. *Human Factors,* 50(3), 522-520.

Pickup, L., Wilson, J. R., Norris, B. J., Mitchell, L., & Morrisroe, G. (2005). The integrated workload scale (IWS): A new self-report tool to assess railway signaller workload. *Applied Ergonomics,* 36(6), 681-693.

Röttger, S., Bali, K., & Manzey, D. (2009). Impact of automated decision aids on performance, operator behaviour and workload in a simulated supervisory control task. *Ergonomics*, 52(5), 512-523.

Sandblad, B., Andersson, A. W., Frej, I., & Gideon, A. (1997). The role of human-computer interaction in design of new train traffic control systems. In *Proceedings of the World Congress of Rail Research* (Vol. 97, pp. 16-19).

Sarter, N.B., & Woods, D.D. (1992). Pilot interaction with cockpit automation: Operational experiences with the flight management system. *International Journal of Aviation Psychology*, 2(4), 303 – 321.

Sarter, N.B., & Woods, D.D. (1994). Pilot interaction with cockpit automation II: An experimental study of pilots' model and awareness of the flight management system. *International Journal of Aviation Psychology*, 4(1), 1-28.

Sarter, N. B., Mumaw, R. J., & Wickens, C. D. (2007). Pilots' monitoring strategies and performance on automated flight decks: An empirical study combining behavioural and eye-tracking data. *Human Factors*, 49(3), 347-357.

Sauer, J., Nickel, P., & Wastell, D. (2013). Designing automation for complex work environments under different levels of stress. *Applied Ergonomics,* 44, 119-127.

Sharples, S., Millen, L., Golightly, D., & Balfe, N. (2011). The impact of automation in rail signalling operations. *Journal of Rail and Rapid Transit*, 225(2), 179-191.

Sheridan, T. B., & Verplank, W. L. (1978). *Human and Computer Control of Undersea Teleoperators.* Cambridge, Massachusetts: Massachusetts Institute of Technology.

Warm, J. S., Dember, W. N., & Hancock, P. A. (1996). Vigilance and workload in automated systems. In R. Parasuraman & M. Mouloua (Eds.), *Automation and Human Performance: Theory and Applications* (pp. 183-200). Mahwah, NJ: Lawrence Erlbaum.

Wickens, C. D. (1992). *Engineering Psychology and Human Performance* (2nd ed.). New York: HarperCollins.

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science*, 8(3), 201-212.

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2013). Engineering Psychology and Human Performance (4th ed.). New Jersey: Pearson.

Wiener, E. L. (1989). *Human factors of advanced technology ('glass cockpit') transport aircraft* (No. Technical Report 117528). Moffett Field, CA: NASA Ames Research Center.

Wiener, E. L., & Curry, R. E. (1980). Flight-deck automation: Promises and problems. *Ergonomics,* 23(10), 995-1011.

Woods, D. D. (1996). Decomposing automation: Apparent simplicity, real complexity. In R. Parasuraman & M. Mouloua (Eds.), *Automation Technology and Human Performance* (pp. 3-17). Hillsdale, NJ: Erlbaum.