AT THE LIMIT? USING OPERATIONAL DATA TO ESTIMATE TRAIN DRIVER HUMAN RELIABILITY

Chris Harrison[1], Julian Stow[2], Xiaocheng Ge[1], Jonathan Gregory[1], Huw Gibson[1] Alice Monk[1]

[1]RSSB, The Helicon, The Helicon, One South Place, London EC2M 2RB
[2]Institute of Railway Research, University of Huddersfield

Abstract

Human reliability analysis plays an important role in the safety assessment and management of rail operations. This paper discusses how the increasing availability of operational data can be used to develop an understanding of train driver reliability. The paper derives human reliability data for two driving tasks, stopping at red signals and controlling speed on approach to buffer stops. In the first of these cases, a tool has been developed that can estimate the number of times a signal is approached at red by trains on the Great Britain (GB) rail network. The tool has been developed using *big data* techniques and ideas, recording and analysing millions of pieces of data from live operational feeds to update and summarise statistics from thousands of signal locations in GB on a daily basis. The resulting driver reliability data are compared to similar analyses of other train driving tasks. This shows human reliability approaching the currently accepted limits of human performance. It also shows higher error rates amongst freight train drivers than passenger train drivers for these tasks. The paper highlights the importance of understanding the task specific performance limits if further improvements in human reliability are sought. It also provides a practical example of how big data could play an increasingly important role in system error management, whether from the perspective of understanding normal performance and the limits of performance for specific tasks or as the basis for dynamic safety indicators which, if not leading, could at least become closer to real time.

The Landscape of Human Reliability

Human reliability is defined as the probability of humans correctly conducting specific tasks with satisfactory performance, e.g., within a time limit. Human error is contrary to human reliability and human reliability assessment provides a methodology for predicting the likelihood of human errors based on characteristics of the task being performed.

There is quite rightly a debate around the value and use of the term 'human error' within the research community (Read et al, 2021) and that performance variability is a more appropriate and less judgemental term. For the purposes of this paper we will use the term human error, as it is integrated within the human reliability assessment methods which form the context for this paper. Whilst using the term, the paper recognises that human variability/human error is only one part of a wider systems view of adverse system outcomes and that human reliability needs to take account of this wider systems view. Also, this paper focuses on the collection of success as well as 'error' data to generate human reliability data, which is a key feature of understanding positive and negative variability identified as a way forward by Read et al (2021).

A EUROCONTROL white paper (EUROCONTROL, 2009), highlights that railways are a more 'tractable' system, when compared to air traffic management, military operations and healthcare for example. The white paper identifies that existing safety methods, such as human reliability assessment, are more relevant to tractable systems such as rail. The continued use and development of human reliability assessment and supporting human reliability data may be more relevant in the railway context than for other more tightly coupled and intractable systems.

Human reliability data can be generated from incident data, observation and automated reporting systems. In addition, Human Reliability Assessment (HRA) methods provide approaches and data to support the assessment of human reliability as part of safety assessment or design studies. These approaches have been reviewed in a Health and Safety Executive (HSE) report (Bell and Holroyd, 2009). Examples of these approaches include the Human Error Assessment and Reduction Technique (HEART) (Williams, 1986), Cognitive Reliability and Error Analysis Method (Cream) and the Technique for Human Error Rate Prediction (THERP) (Swain and Guttman, 1983).

The HEART method for HRA is based around generating human reliability estimates, recognising that the likelihood of error occurring may be influenced by one or more error producing conditions such as operator inexperience, equipment design or low workforce morale. The task being analysed is mapped to generic task types (which have a Human Error Probability (HEP) value

associated with them), before a proportion of the effect of the Error Producing Conditions (EPCs) are added to produce a final HEP value (Williams, 1986, Hasibuan et al., 2020; Walker and Strathie, 2013).

The Railway Action Reliability Assessment (RARA) tool is based on HEART which is tailored towards tasks commonly undertaken by railway staff (Gibson et al, 2012, 2013). This railway focus makes RARA the most relevant tool to support the discussions in this paper, as the values used for quantification are consistent with the HEART approach, but the generic task type and EPC descriptions have been described in further detail for the railway context. The RARA tool was developed as part of a research programme managed by RSSB and funded by the UK Department for Transport. Some examples from RARA are presented in Table 1 which illustrate how performance reliability can vary by task type. These data highlight that for simple tasks where warning systems support the train driver, the best expected human error probability is 2.0E-05 or an error rate of 1 in 50,000 opportunities. This reliability value is used in a number of industries to express the limits of human reliability, in terms of the best human performance that can be expected. As can be seen in Table 1, there are quite wide bounds associated with this 'limit' ranging from (approximately) 1 in 1100 to 1 in 167,000. Human performance is significantly worse when there is an element of confusability in simple tasks (e.g. saying a signal number as 'one six five' when 'one five six' was intended) with an estimate of 3.0E-03 (1 in 33 opportunities). Finally, where human performance requires decision making under uncertainty, performance values such as 1.6E-01 (1 in 6 opportunities) or higher may be experienced. These levels of performance may be experienced during railway fault finding in degraded modes or emergencies.

Table 1 Example of task human reliability estimates from the Railway Action Reliability Assessment tool

Human Error Probability	Bounds	Error Rate (1 in …)	Task Description	Example Task
2.0E-05	6.0E-06 to 9.0E-04	50,000	Respond correctly to system command even when there is an automated system providing accurate interpretation of system state.	Locate and act on railway signals outside the cab on a route frequently driven by the driver, which is optimally sighted, is not located with other signals and is preceded by an audible warning to alert the driver to the signal.
3.0E-03	2.0E-03 to 4.0E-03	33	Skill-based tasks (manual, visual or communication) when there is some opportunity for confusion.	Communicating a signal number.
1.6E-01	1.2E-01 to 2.8E-01	6	Complex task requiring a high level of understanding and skill,	Diagnosing a train fault which presents complex indications to the driver.

The predictive power of these techniques is most obvious when no human error data is available in relation to a proposed design or redesign. However, even when human error data are available, for example from incident or near miss reporting systems, theoretical error rate predictions from error quantification methods can be useful. Comparing error rates experienced within the system to the rate predicted by human error quantification methods helps to set the experienced error rate in context, and to understand the extent to which human reliability improvements can be expected by addressing error producing conditions. For example, if the real-world data identify more reliable performance than the most reliable performance identified in a technique, then it may be necessary to consider fundamental redesign for the task and recognise that improved training or improvement to environmental factors may have a minimal impact. However, the fundamental redesign needs to take account of adverse effects that a fundamental change may have on operator or system performance, in particular if the operator task is automated (Bainbridge, 1983).

There are limitations in the use of HRA methods, for example, in the age of the data the models are based on. On this topic Hickling and Bowie (2013) identify that underpinning data for HRA approaches do not reflect the wider adoption of

human-computer interfaces for process control. There is also the need to develop domain-specific approaches, for example in the air traffic control domain reliability data have been collected and used to support human reliability tool development (Kirwan et al, 2008). Also, there have been significant advancements, not only in technology but also in the understanding of the complexity of human behaviour (French et al., 2011). This does not mean that there is no value in using these approaches, and there have been validation studies which have reviewed human reliability assessment approaches (e.g. Kirwan et al., 1996, 1997, Forester et al, 2014). However, it highlights that there needs to be continued efforts to collect real world human reliability data to reflect changes to tasks and build on the known limitations in the data underpinning HRA techniques. Elements of the HEART approach data have been updated by Williams and Bell (2015).

Increasing availability of large operational data sets provides an opportunity to revisit some human error reliability data in RARA (Gibson et al, 2012). Walker and Strathie (2013) take this a step further by discussing the idea of a 'living database' which means that error probability calculations are based on performance databases. This would signal a move away from the use of generic task types based on static databases to live databases specific to individual tasks and contexts. These databases have the potential to become a continuous source of performance measurement rather than used at specific time points (Walker and Strathie, 2013) and are being developed in other safety critical industries. For example, Walker and Strathie (2014) found great potential in using flight data monitoring which pairs the data from black boxes at the end of every journey and couples it with human factors methodologies to detect risks in advance of accidents happening. Walker and Strathie (2016) have also completed a proof-of-concept study which identified human factors leading indicators using on track data recorder (OTDR) data. The availability of this type of data could highlight tasks where design changes are required to enable better human performance or emerging risk areas and to even track the impact of change. This could be an aspiration for the rail industry and feasibility in applying this type of data to RARA analyses would be worth exploring.

The following sections look at some specific examples of train driver task reliability and their context, the associated HEPs that can be estimated along with an overview of how the data to achieve this was collected – which is not trivial.

Quantifying the human reliability of train drivers stopping at red signals

An event where a train passes a signal showing a stop aspect without authorisation is known as a *signal passed at danger* (SPAD). SPADs can range

from minor incidents where a signal is passed by only a few metres, to serious incidents where longer overruns give rise to the chance of collision with other trains. One of the most serious examples of the type of incident that can result from a SPAD is the accident at Ladbroke Grove, UK, in 1999, which resulted in 31 fatalities and 417 injuries. The accident report (Cullen, 2000) identified key failings in the design of the signalling system, signal sighting and driver's training as causes of the accident.

As the driver controls the train brakes, in most cases the immediate precursor for a SPAD or TPWS intervention approaching buffer stops is the actions of a driver. The frequency with which such errors occur is the HEP. It is now widely recognised, for example by Gibson et al (2016) that the driver's 'error' in passing a signal is usually the endpoint of a range of systemic errors which increase the probability that a SPAD may occur. These encompass a wide range of factors including shift patterns which increase fatigue, poor signal sighting and poor communications. The driver is therefore one part of a complex sociotechnical system, and it is the failure of one or more aspects of that system, as a whole, which lead to a SPAD. The authors recognise that the concept of 'human error' or 'driver error' is simplistic in this regard and that industry has generally moved away from this deterministic approach. Indeed, it is argued later in the paper that train drivers are probably reaching the limits of human performance with regard to stopping at red signals or, more accurately, the system (manually driven trains with fixed block lineside signalling with in-cab warning devices) is reaching the limits of its performance. The ability of the human train driver to correctly respond to signals, judge braking distances and adapt to the prevailing conditions remains at the heart of this system. For convenience the term HEP is therefore used throughout this paper although it is recognised that this alone does not adequately represent the complexities of the system in which the driver is working.

Stanton and Walker (2011) identify that SPADs have been occurring for over a hundred years and described many of the SPADs happening today as 'stubbornly resistant to a wide range of well-intended safety measures'. There are a range of research studying the nature of the underlying systems causes and human performance factors associated with SPADs, including (Gibson, 2016 Balfe et al, 2017, Buck, 1963). A recent review of SPAD investigation reports in the UK (Gibson, 2016) concluded that SPADs feature multiple causes, and there was a bias in these reports towards investigating driver performance rather than underlying factors. These underlying factors may be direct/indirect or latent/active, but they combine and interact to enable an incident to occur (Lawton and Ward, 2005; Santas-Reyes and Beard, 2006). Most importantly, Gibson (2016) has recommended several approaches to link the human factor data to SPAD risk assessment.

This paper focusses on trains approaching red signals although the importance of reactions to the preceding signals, signal warning systems and underlying systemic factors should be considered in a holistic analysis of incidents or precursors to incidents. It follows that knowing the number of trains that approach signals displaying a red aspect (the 'red approach rate') is fundamental to the understanding of SPAD risk at individual signals. It also aids in the normalisation of SPAD data, both locally and nationally, for trending and benchmarking. Until recently, there has been limited data on the number of red aspects approaches; instead train km have been used as a proxy for the red aspect approach rate. For example, the European Rail Agency collect and publish SPADs per million train km for each EU country every two years, the most recent data available being European Rail Agency (2022). For the UK, the figure for 2019 is given as 0.57271 SPADs per million train km.

It should be noted that using train km has some significant drawbacks. It does not account for the likely differences in red approach rates that different types of train operation (e.g., passenger or non-passenger trains) and different types of operators (e.g., rural, suburban commuter, or long distance) experience. A better normaliser is the number of red aspect approaches; however, these have not been routinely collected and a way of undertaking this is needed. To be able to have an accurate view on the red aspect approach rate, the Red Aspect Approaches to Signals (RAATS) tool has recently been developed using a big data approach.

Overview of the RAATS tool

Network Rail (the GB mainline railway infrastructure manager) provides publicly available live data feeds that give various information on the movement of trains. At the most fundamental level, the source of the information used by the RAATS toolkit is train describer (TD) data. A TD is an electronic device connected to each signalling panel which provides a description of each train (its 'headcode') and which section of track (or 'berth') it currently occupies. The TD is responsible for correctly displaying the train movements from berth-to-berth to the signaller and for ensuring that the train's identity is correctly passed to the next signaller's panel when it leaves the current signalling area.

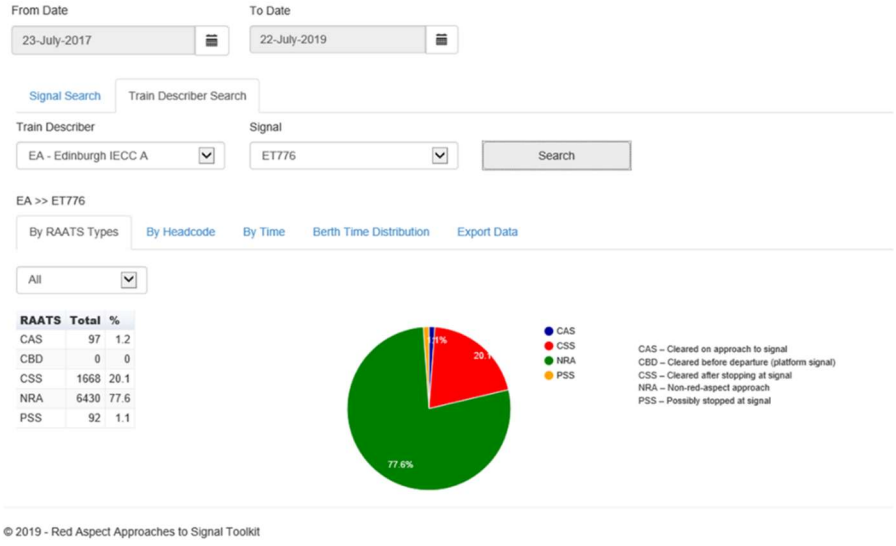
RAATS uses two separate TD data feeds, termed C-Class and S-Class messages. C-Class messages record train movements between individual track berths, whilst S-Class messages record the times at which signal aspects change. The S-Class data only shows whether a signal is *on*, showing a red aspect or *off*, showing a proceed aspect (single or double yellow, or green). Both C-Class and

S-Class messages are transmitted through the live feed with a combined total of approximately 5.2 million messages being sent per day.

The RAATS toolkit takes these data and applies an algorithm (Harrison, 2015; Zhao et al, 2016, 2018) that models and classifies each approach by a train to a signal. The initial classification is whether the train approached a red signal or not. This can be simply determined by the status of the signal the train is approaching when it enters a berth. Instances where a train is approaching a red signal are further broken down according to whether the train is likely to have come to a complete halt, or whether the signal is likely to have cleared to a proceed (non-red) aspect before the train came to a stand. The tool also uses supplementary information to enable these red approaches to be broken down by other factors such as the train type, the time of day, the day of the week. Results for a selected signal or signals can also be exported from the tool for further analysis. Figure 1 shows a screenshot of the RAATS toolkit.

The RAATS database currently contains around 610,000,000 train approaches to approximately 12,000 signals from June 2017 onwards. This represents approximately one third of the signals on the GB network and is limited by the availability of S-class signal aspect data. Nevertheless, RAATS provides a large data source which provides sound estimates of the overall red aspect approach rate for GB railways.

Figure 1: Screenshot of the RAATS toolkit



The outputs from the RAATS toolkit provide a wealth of statistics that enable SPAD risk and the causes of SPADs to be put into the context of the number of red aspect approaches.

Using RAATS to understand human reliability for SPADs: *an analysis of RAATS and SPAD data from 2019*

There were 336 SPADs recorded on the GB rail network during 2019. Of these 324 can be attributed to 'driver error', the other 12 mostly being due to communication related errors by other railway staff. It should be noted that the error rates for approaching red signals are based only on events that resulted in a SPAD. In this study they do not include overspeeding on the approach to signals where a possible SPAD was prevented by the intervention of the automatic TPWS system which are worthy of further investigation. There will also be a number of events where trains nearly overran a non-TPWS fitted signal. Unless reported by the driver, these will generally remain unknown.

The outputs from RAATS can be used to make a good estimate of the total number of red aspect approaches over the same period. The operational data feeds provide train movement information on approximately 85% of the country. Based on this data for 2019 there were 590,804,638 approaches to signals. Renormalising this to 100% (assuming the missing 15% is broadly similar in characteristics to the rest of the country) to account for the whole country gives an estimate of approximately 695 million approaches by trains to signals in 2019.

The RAATS outputs themselves only cover around a third of the signals on the network. In 2019 there were 223,676,641 approaches to signals in the RAATS dataset. This gives a factor of 3.11 less compared with the estimated overall total number of approaches stated previously and is broadly in line with the estimated RAATS network coverage of a third. Of these approaches around 90.5% were to non-red aspects (greens, yellows, double yellows). Of the 9.5% of approaches to a red aspect (the signal being approached is displaying a red), around a fifth (2.0% of all approaches) were to a red aspect where the train is predicted to eventually stop and wait for the signal to clear according to the RAATS algorithm. For these approaches a SPAD would have occurred if the train had not stopped, so an initial estimate of the number of approaches where there was an opportunity for a SPAD for 2019 may be given as 2.0% of 695 million (13.6 million).

This assumes that the average red approach rates across the country can be used when accounting for the missing two thirds of the network. An alternative approach is to look at smaller regions of the country (known as train describer

(TD) areas, of which 114 are available in RAATS) and work out individual red aspect approach rates for these and then apply these to the total approaches in them. This way of estimating is more robust as the variables (drivers, trains, signals) within a TD are more representative for scaling a local estimate to account for unknowns compared with taking a national average and scaling for a national unknown (as the method before was doing). Using this localised approach gives a total number of red aspect approaches of 4,502,809. Using the factor of 3.11 derived earlier, this can then be scaled to give an estimate of the national number of red approaches of 14.0 million. This figure can be considered more accurate as an estimate as it takes into account the local variation in red aspect approach rates then factors up for this accordingly as opposed to applying an average factor nationally as presented initially. The 114 train describer areas also cover diverse areas of the country and are representative of all train operations and types of route (eg commuter, long distance, rural). Therefore, scaling from them can be considered to be reasonable and representative of a national red approach rate.

In the context of SPADs, the HEP can be defined as:

$$\frac{\textit{Number of signals passed at danger}}{\textit{Number of train approaches to signal red aspects where the train was required to stop}}$$

Using the 14.0 million red aspect approaches as the total opportunity for a SPAD with the actual number of SPADs allows the human reliability of stopping at a signal at danger to be calculated. Based on data for 2019, this is 324/14,000,000 to give a HEP of 2.32E-05 or a SPAD rate of around 1 in 43,000 red approaches. This is the overall rate considering all driver error SPADs and all red aspect approaches to red signals for 2019.

The train approach data can be broken down in different ways to explore several underlying factors that may influence the SPAD rate. An example of such a breakdown is given in Table 2.

Table 2: Breakdown of SPADs and approaches to signals in 2019

	Passenger Train	Empty Coaching Stock (ECS)	Freight
SPADs	66.7%	21.0%	12.3%
All approaches to signals	88.3%	5.0%	6.7%
Non-red aspect approaches to signals	88.9%	4.4%	6.7%
Red aspect approaches to signals	82.3%	10.4%	7.3%
Red aspect approaches to signals where the train stopped	72.7%	11.4%	15.9%

Table 2 shows a breakdown of the proportions of all SPADs (324 from the incident data) and the proportions of RAATS approaches (from the 223,676.641 approaches) to signals during 2019 by three different types of train service: passenger, empty coaching stock (ECS) and freight[1]. The first two lines of the table show that although passenger trains account for the majority of approaches to signals they account for a lower percentage of SPADs. Conversely freight trains and ECS account for a higher proportion of SPADs than the proportion of approaches might suggest.

The last line of Table 2 shows the proportion of red aspect approaches to signals where the train stopped. This profile is markedly different to that shown by the second line showing the profile based on all approaches to signals. Comparison shows that the relationship between all approaches and red aspect approaches is not linear. This illustrates the point made earlier about the inadequacy of using train miles to normalise SPAD risk (as train miles can be considered in general to be proportional to the number of signal approaches).

Table 3 shows a breakdown of the different types of SPADs by train service based on the proportions shown in Table 2 and the total SPADs for 2019 of 324. The corresponding red aspect approaches where the train stopped for the different train services can be made by applying the proportions derived from the available RAATS data shown in the last line of Table 2 and applying to the estimated number of 14 million red aspect approaches where the train stopped in 2019. These estimates can then be used to make an estimate of the HEP and SPAD rates for the different type of train services, which are shown as the last two lines in Table 3 respectively.

1 The vast majority of trains within this category are freight trains, but there are some non-passenger trains included that strictly speaking would not be classified as freight.

Table 3: Breakdown of SPADs and red approaches to signals where the train stopped in 2019

	Passenger Train	Empty Coaching Stock (ECS)	Freight
SPADs	216	40	68
Red aspect approaches to signals where the train stopped	10,169,548	2,222,856	1,599,857
Human error probability	2.12E-05	1.80E-05	4.25E-05
SPAD rate (1 in …)	47,000	55,600	23,500

Table 3 shows that in general passenger and empty coaching stock services have a broadly similar SPAD error rate of around 1 in 47,000 and 1 in 55,600 respectively, whereas freight trains have a SPAD error rate 2 to 2.4 times these being around 1 in 23,500.

There are a number of reasons that may explain this observation including:

- Passenger train services tend to be in the peak and off-peak hours of the day. By their nature, ECS services tend to be at the beginning or end of the day when passenger services are coming into or out of service. Freight trains can operate at any time of the day, with many running during the night when drivers may be more prone to fatigue related errors.
- Passenger trains tend to have a regular and predictable service pattern. This means that day-to-day the aspect a particular signal is likely to be displaying on approach by a passenger service is somewhat predictable. The same is not true for freight services, where the routing and timing of services can vary considerably day-to-day leading to less predictable signal approach patterns.
- Freight trains typically have a different speed profile approaching red signals, as drivers may approach slowly to try and avoid time lost through accelerating heavy trains from a stop. This provides greater opportunity for loss of concentration when approaching signals at low speed.
- Freight train brakes take longer to apply and have lower performance than modern multiple unit passenger trains, providing more opportunities for drivers to misjudge stopping distances and less opportunity to correct misjudgments when they occur.

Quantifying the human error reliability of train drivers approaching buffer stops

A similar human reliability analysis was carried out for approaches to buffer stops on the GB rail network. It is a requirement for the approach to every buffer stop in a station to be fitted with a Train Protection & Warning System (TPWS) Overspeed System Sensor (OSS) grid to provide protection against a

train coming in too fast and potentially striking the buffer stop. The OSS grids are set to intervene at 10mph and if a train passes at a speed over this limit then the emergency brakes will be applied.

There is a corresponding requirement for train drivers to approach a buffer stop at less than 10mph and this requires the driver to control their train speed below this limit. If they don't, a TPWS activation or intervention will be triggered and recorded as an approach to a buffer stop at greater than 10mph. In the time period of the years 2015 and 2016 there were a total of 133 such incidents.

Based on an analysis of the timetable and the trains planned to run it is estimated that there were in total approximately 4.4 million approaches by trains to buffer stops in stations per year in 2016. Assuming a similar figure for 2015 allows the HEP for a train driver not adequately controlling their train on approach to a buffer stop to be estimated. Based on data for 2015-2016, this is 133/2/4,400,000 to give a HEP of 1.53E-05 or an overspeed on approach to buffer stop rate of around 1 in 66,000 buffer stop approaches.

This is the overall rate considering all driver errors on approach to TPWS fitted buffer stops for 2015-2016. The buffer stop approach data can also be broken down as described previously for SPADs. Table 4 shows this breakdown, which is based on partitioning the previously mentioned 2016 timetable analysis by train type (as this information is provided in the schedules) and scaling by a factor of 2, assuming that 2015 was broadly similar.

Table 4: Breakdown of overspeeds on approach to buffer stops and approaches to buffer stops in 2015-2016

	Passenger Train	Empty Coaching Stock (ECS)	Freight
Overspeed on approach to buffer stop events	122	10	1
Total approaches to buffer stops	8,000,000	650,000	55,000
Human error probability	1.52E-05	1.54E-05	1.84E-05
Overspeed on approach to buffer stop rate (1 in …)	66,000	65,000	54,000

As before with SPADs, it is possible to make an estimate of the human error probability and overspeed on approach to buffer stop rates for the different types of train services which are shown in the final two lines of Table 3. These show that in general passenger and empty coaching stock services have a broadly similar error rate of around 1 in 66,000 and 1 in 65,000 respectively, whereas freight trains have an error rate slightly higher than both of these being around 1 in 54,000. It should be noted that the freight estimates are very

uncertain as there are comparatively fewer approaches to TPWS equipped buffer stops by these type of trains per year and only a single recorded error on approach. It is notable that the error rates and HEP are broadly similar and comparable to the SPAD error rates presented earlier in the paper.

Discussion

Comparison with previous human reliability studies

The HEP estimates presented in this paper can also be compared with previous studies of train driver reliability. A study (Gibson et al. 2017) derived HEP estimates for failure to call at a station, station overruns and wrong side door releases - Table 5 summarises the main findings.

Table 5: Estimates of HEP from (Gibson et al. 2017)

	Failure to call at a station	Station Overrun	Wrong side door release at station
Number of events	95	73	42
Exposure	12,117,190	12,117,190	12,599,599
Human error probability	7.84E-06	6.02E-06	3.33E-06
Error rate (1 in …)	128,000	166,000	300,000

It is notable that the human error rates (HER) for approaching red signals appears to be around 3 – 7 times those found by Gibson et al for station overruns, failure to call and wrong side door release events. This is discussed further below.

HEP in RARA and Limits of Human Performance

Extensive work has been carried out to quantify HEPs across a range of industries and task types. As noted by Gibson (2016), RARA identifies that the limits of human performance for well designed, simple tasks where there is an automated supervisory system in place is around 2E-05, or once in every 50,000 opportunities.

Using the RARA approach, the calculation approach would normally involve input from task experts and review of the task in context and in detail. For the purposes of producing an approximate and generic estimate for comparison within this paper, the following has been calculated by the authors using the RARA manual methodology.

Following the guidance in the manual, for the task of stopping at a red signal (the task associated with a SPAD) the generic task type R1 “Respond correctly

to system command even when there is an automated system providing accurate interpretation of system state" from RARA could be selected, based on the RARA guidance which identifies "Locate and act on railway signals outside the cab" as an example for this this generic task type. The probability is not modified with error producing conditions for this type of generic assessment, again based on the guidance from the approach. The calculated value from RARA would therefore be 2E-05 for the task of stopping at a red signal. This aligns very closely with the value generated from data in this paper.

For the task of overspeeding on approach to buffer stop, the generic task type R2 "Completely familiar, well designed, highly practiced task which is routine" could be selected as the closest match, based on the driver examples provided in the approach (eg "Controlling train speed under normal operating conditions on green signals on a familiar route."). Again, error producing conditions would not be applied in this type of generic assessment and the value calculated from RARA would be 4E-04. This is a higher error probability that that found in the data of 1.52E-05. It is beyond the scope of the paper to explore this difference in detail, it could for example, relate to the under-reporting of buffer stop incidents.

The data collected for this study for train drivers stopping at a red signals and for approaching a buffer stop without over speeding are strikingly close to the lowest human error probability value in both RARA and HEART of 2E-05. This suggests that as cohort drivers, who are by definition skilled, experienced and trained in both technical and non-technical skills, are operating near to the accepted limits of human reliability. This is significant as it may explain the difficulty faced by the industry in achieving further reductions in the number of SPADs and therefore the resulting SPAD risk. It also demonstrates the importance of understanding the underlying and systemic factors that have contributed to SPADs as SPADs are rarely just a case of individual driver error. It should be noted however that Gibson et al (Gibson et al, 2017) found evidence of drivers performing tasks at higher level of reliability, reaching or even exceeding the bounds suggested by RARA, a finding supported by the buffer stop approach HEPs reported in this paper. Further work would be required to understand exactly why these particular tasks have lower HER than key tasks such as stopping at signals. A key consideration would be to understand any potential differences due to underreporting in the error rates due to how the events are reported (automatically detected by the signalling or train systems versus manual reporting by a member of the workforce).

These important considerations warrant further investigation, both in terms of the apparent differences between train types, but also whether further significant improvements may be achieved through training or only by the introduction of

additional technology to assist drivers. An important related consideration will be the level of investment that could be justified based on the current risk profile.

Other related research and further work

The preceding analysis has considered some of the factors that influence the SPAD error rate from the perspective of the type of service. It is also possible to look at the error rates in terms of the properties of the signal. A previous study (Nikandros and Tombs, 2007) observed that in general signals with a high proportion of red aspect approaches actually have a relatively low chance of a SPAD occurring for each red aspect approach because drivers are accustomed to approaching the signal at red. Conversely, signals with a low proportion of red aspect approaches have a relatively high chance of a SPAD occurring when the signal is approached at red because drivers have less experience of stopping at these signals. The analysis presented in Table 1 presumes that signals with more red aspect approaches are proportionately more likely to have a SPAD event. In reality, it is likely to be a combination of this and the proportion of approaches to a signal that are red aspect approaches together that influence the chance of a SPAD at the signal. To look at this further requires a more in-depth signal-by-signal analysis and this is work that RSSB and the University of Huddersfield are actively pursuing.

Analysis of RAATS data has highlighted that the type of train is a factor that needs to be considered when calculating error probability. This has raised the question of what is exactly meant by SPAD probability, as the answer depends on the context in which it is answered. For example, SPAD probabilities associated with different types of train approaches to a single signal (as presented in this paper) is one view. Alternatively, SPAD probability being calculated from an individual train operator's perspective along a route or entire operation, taking into consideration all of the train approaches to all the signals on a journey across all of the operator's services is another view.

Work is also ongoing for GB rail to collect better data on the underlying causes of SPAD incidents using the causal classification framework based on the 10 incident factors (Gibson et al, 2019). Linking these data with RAATs data will create a richer picture, which links the driver reliability data with underlying causes related to safety management systems.

More generally, it is likely that further data sources will become available allowing greater insight into the HEPs associated with various train driving tasks. For example, El-Rashidy et al (2018) highlight how recent development of automated train driver competence indicators based on on-train data recorder analysis will provide very large data sets to quantify how often drivers deviate

from a range of rules and procedures. Further developments of the RAATS tool are planned to examine red aspect approaches by train service instead of by individual signal.

Finally, as noted earlier, the error rates for approaching red signals are based only on events that resulted in a SPAD. They do not include overspeeding on the approach to signals where a possible SPAD was prevented by actions of the driver or TPWS. Being able to collect and assess this data would enable refinement of some of the HEP presented in this paper.

The use of 'Big Data' in system error management

This study provides a useful example of how 'big data' can inform and change approaches to understanding and managing errors in complex sociotechnical systems. At the most straightforward level, the careful use of signalling data has been shown to provide a better understanding of red signal approach rates and how this can build confidence in error rates based on its use. It is likely that this approach is applicable to a wide range of human-centred tasks both within and beyond the railway industry and that using such an approach may allow the understanding of the reliability thresholds to be refined further. It is already the case for example that a much wider data set is available within the railway industry than is utilised in this study given the availability of continuously logged data from on-train data recorders and signalling centres. However, the use of such data also presents considerable challenges, particularly when several large data sets must be accurately synchronised or where data quality issues arise.

However, the increasing availability of such data opens up considerable opportunities for understanding errors and managing risk. Where, as in the example described in this paper, the limits of reliability are approached, big data may permit a more in-depth analysis of the 'normal' range of performance with the aim of early identification of when either the human or technical parts of the system are deviating from that norm. In the case of signal approaches that could mean, for example, an increase in the red approach rate due to timetable or train regulation changes, later braking by trains approaching restrictive aspects or drivers responding more slowly to AWS warnings due to changes in depot shift patterns. This in turn lays the foundations for new types of dynamic, data driven risk management systems which bring the possibility of moving closer to, if not into, real-time applications. However, such possibilities will inevitably have to compete with technological solutions, such as the European Train Control System (ETCS) Level 3, which reduce the opportunities for the human in the system to make errors which lead to trains exceeding their movement authority.

Conclusions

The paper has illustrated how operational data, and the increasing availability of *big data*, can be used to understand human reliability of two closely linked train driving tasks, stopping at red signals and avoiding overspeeding on approach to buffer stops. An important enabler of accurate estimation of error rates approaching signals is the RAATS tool which contains many millions of train approaches which can be used to establish accurate estimates of how often trains approach, and stop at, red signals.

The findings show that an estimate of the overall human error probability for stopping at a red signal is 2.32E-05 (around 1 in 43,000 approaches) and for overspeeding approaching buffers of 1.53E-05 (1 in 66,000 approaches). Breaking these down by train type highlighted some notable variations. In each case freight trains show higher error rates than passenger trains, up to 2.4 times greater in the case of red signal approaches.

There are a number of operational reasons that may explain this observation, including:

- Differences in how these train services are routed and prioritized.
- The times, predictability and regularity of the routes and signals approached by the different train services.
- The braking characteristics of freight trains.

It is clear from the analysis presented that train drivers are operating at or close to the accepted limits of human reliability. However, these limits may be quite task specific and there is evidence from other studies that some driving tasks which are on the face of it quite similar, achieve even higher levels of reliability. Understanding the differences between tasks and the limits of human reliability when carrying them out is an important factor when determining whether improvements require better training, re-design of tasks or provision of automated driver assistance technology. It is likely that the increasing availability of data from train, infrastructure and signaling systems will support developments in this area in the near future.

More widely, the use of red approach data provides a practical example of how big data could play an increasingly important role in system error management, whether from the perspective of understanding normal performance and the limits of performance for specific tasks or as the basis for dynamic safety indicators which, if not leading, could at least become closer to real time. Further work is required to understand how this better knowledge of spatial and temporal variations in risk profile can be exploited to improve railway safety.

References

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

Balfe, N., Geoghegan, S., Smith, B., 2017. SPAD Dashboard: A tool for tracking and analysing factors influencing SPADs, in: Charles, R. and Wilkinson, J. (Eds) Contemporary Ergonomics and Human Factors 2017.

Bell, J., Holroyd, J. 2009. Review of human reliability assessment methods. RR679. HSE Books, Norwich.

Buck, L., 1963. Errors in the perception of Railway Signals. Ergonomics. 6, 181-192.

Cullen, W., 2000. Ladbroke Grove Rail Enquiry, Part 1 – Report. Her Majesty's Stationary Office, Norwich

El-Rashidy et al., 2018, Automated train driver competency performance indicators using real train driving data, chapter in Haugen, S., Barros, A., van Gulijk, C., Kongsvik, T., & Vinnem, J.E. (Eds.), Safety and Reliability – Safe Societies in a Changing World (1st ed.). CRC Press. https://doi.org/10.1201/9781351174664

EUROCONTROL, 2009. A White Paper on Resilience Engineering for ATM. EUROCONTROL, Bretigny-Sur-Orge.

European Rail Agency, 2022. Common Safety Indicators data (2006-20), https://www.era.europa.eu/file/8104/download_en?token=PvM7IGMq accessed 03/03/2022

Forester et al, 2014. The International HRA Empirical Study. USNRC: NUREG-2127. US Nuclear Regulatory Commission, Washington.

French, S., Bedford, T., Pollard, S., Soane, E., 2011. Human Reliability Analysis: A critique and review for managers. Safety Science. 49, 753-763.

Gibson, W.H., 2012. Railway Action Reliability Assessment user manual, A technique for the quantification of human error in the rail industry. RSSB, London.

Gibson, W.H., Mills, A., Smith, S., Kirwan, B. K., 2013. Railway Action Reliability Assessment A Railway - Specific Approach to Human Error Quantification in: Dadashi, N. Scott, A., Wilson, J.R. Mills, A. (Eds.). Rail Human Factors Supporting Reliability, Safety and Cost Reduction. Taylor & Francis Group, London, 671-676.

Gibson, W.H., 2016. Industry Human Factors SPAD Review - Project Summary Report. RSSB, London.

Gibson, W.H., Willett, J., Lewis, G. and Harrison, C., 2017. Exploring the limits of train driver reliability in: Sixth International Human Factors Rail Conference, 6-9 November 2017, London.

Gibson, W.H., Mills, A., Monk, A., Waters, S., Smith, J. 2019. Understanding signals passed at danger through a human factors lens in: 12th World Congress on Railway Research (WCRR2019) Tokyo, Japan.

Hasibuan, C., Daeng, P.Y., Hasibuan, R., 2020. Human Reliability Assessment Analysis with Human Error Assessment and Reduction Technique (HEART) Method on Steriliser Station at XYZ company in: The 2020 International Conference on Information Technology and Engineering Management, Batam, Indonesia.

Kirwan, B., 1996. A validation study of three human reliability quantification techniques: THERP, HEART and JHEDI-Part1-Technique descriptions and validation issues. Applied Ergonomics, 27(6), 359-373.

Kirwan, B., Kennedy, R., Taylor-Adams, S., Lambert, B., 1997. A validation study of three human reliability quantification techniques: THERP, HEART, and JHEDI – Part II – Results of validation exercise. Applied Ergonomics, 28(1), 17-25.

Kirwan, B., Gibson, W.H., Hickling, B., 2008. Human error data collection as a precursor to the development of a human reliability assessment capability in air traffic management. Reliability Engineering and System Safety, 93, 217-233.

Lawton, R., Ward, N.J., 2005. A systems analysis of the Paddington railway accident. In Accident Analysis and Prevention, 37, 235-244.

Nikandros, G., Tombs, D., 2007. Measuring Railway Signals Passed at Danger. 12th Australian Conf. on Safety Critical Systems and Software, Adelaide. Conf. In T. Cant (ed.) Research and Practice in Information Technology (CRPIT), vol. 86.

Read, G.J.M, Shorrock, S., Walker, G.H., Salmon P.M., 2021. State of science: evolving perspectives on 'human error'. Ergonomics 64, 9, 1091-1114.

Stanton, N., Walker, G. 2011. Exploring the psychological factors involved in the Ladbroke Grove rail accident. Accident Analysis and Prevention, 43, 1117-1127.

Swain, A.D. and Guttman, H.E., 1983. A handbook of human reliability analysis with emphasis on nuclear power plant applications. NUREG/CR1278. USNRC, Washington D.C..

Walker, G. and Strathie. A., 2013. Human Factors leading indicators.Flight data monitoring, On Train Data Recording and Human Factors. EPSRC EP/1036222/1. Heriot Watt University, Edinburgh.

Walker, G., Strathie, A., 2014. Combining human factors methods with transport data recordings in: Proceedings of the 5th International Conference on Applied Human Factors and Ergonomics AHFE2014. Krakow, Poland.

Walker, G., Strathie, A., 2016. Big data and ergonomics methods: A new paradigm for tackling strategic transport safety risks. Applied Ergonomics, 53, 298-311.

Williams, J.C., 1986. HEART – a proposed method for assessing and reducing human error in: Proceedings of the 9th advances in reliability technology symposium. University of Bradford, UK.

Williams. J.C. and Bell, J.L., 2015. Consolidation of the Error Producing Conditions used in the human error assessment and reduction technique. The Journal of the Safety and Reliability Society. 35(3), 26-76.

Zhao, Y., Stow, J., Harrison, C., 2016. Estimating the frequency of trains approaching red signals: a case study for improving the understanding of SPAD risk. IET Intelligent Transport Systems, 10,(9), 579-586.

Zhao, Y., Stow, J., Harrison, C., 2018. A method for classifying red signal approaches using train operational data. Safety Science 110, Part B, 67-74.