Safety Science 192 (2025) 106958

Contents lists available at ScienceDirect

Safety Science

journal homepage: www.elsevier.com/locate/safety

Re-evaluating fatigue measurement: A comparative study of subjective and objective fatigue tests

J.L. König [*,1], J. Bowen [2], A. Hinze [3]

School of Computing and Mathematical Sciences, University of Waikato, New Zealand

ARTICLE INFO

Keywords:
Fatigue
Subjective fatigue tests
Objective fatigue tests
Participatory study

ABSTRACT

Fatigue poses a significant risk in hazardous industries, with forestry being a particularly under-examined domain. Despite the availability of subjective and objective fatigue tests, inconsistencies in their application, selection rationale, and performance remain largely unaddressed in existing literature. This paper investigates the utility and challenges of subjective and objective fatigue assessments through a review of existing literature and two case studies: one intensive longitudinal study with a single participant and another broader study involving 31 participants. Our results reveal strong internal consistency across subjective tests but variable outcomes for objective tests, raising questions about test sensitivity and context-specific reliability. We argue for clearer guidance on fatigue test selection and propose criteria to inform future research in complex, real-world settings like forestry.

1. Introduction

Fatigue is a critical safety issue in hazardous work environments, contributing to reduced performance and increased incident rates. In industries such as forestry, where both mental and physical exertion are high, fatigue has been implicated in accidents caused by lapses in judgment, poor concentration, and physical mishaps like slips or falls (Anonymize, 2021). Despite the acknowledged danger, fatigue in forestry is often inferred from incident report patterns or survey data (Bentley et al., 2005; Lilley et al., 2002), with little research deploying validated fatigue tests directly within this context.

Two primary approaches are used to assess fatigue: subjective tests, which capture perceived tiredness through self-reporting tools; and objective tests, which evaluate performance metrics like reaction time. While both types of tests are common in other hazardous sectors such as aviation, construction, and military contexts (Storm, 1983; Darbandy et al., 2020; Good et al., 2020), their usage in forestry remains limited and fragmented. Complicating matters further, there is little consensus on which tests to use, how many to deploy, or how to interpret inconsistent results between them.

This paper aims to examine the challenges of fatigue testing by reviewing existing literature across hazardous industries and conducting two targeted case studies. The first is an intensive five-day study involving a single participant, and the second includes a larger sample assessed across multiple sessions. Together, these studies explore the relationships and discrepancies between different fatigue measurement tools and raise important questions about test selection, technological reliability, and cross-test comparability. Our goal is to provide practical recommendations and methodological clarity for researchers and practitioners working in real-world, high-risk environments.

The paper is structured as follows. First, we introduce related work on fatigue, measuring fatigue, and fatigue testing techniques, both within the forestry industry, and within other hazardous industries. Next, we investigate the challenges of working with fatigue tests, and outline our investigation. After this, we introduce and evaluate our two case studies. Finally, we discuss our results, limitations, and future work.

2. Related work

Fatigue tests have been used for a long time in numerous industries. However, there are still major challenges associated with their use. In this section we first discuss fatigue and how fatigue is measured. We then discuss how fatigue is typically measured in the forestry industry and other hazardous industries.

* Corresponding author.
E-mail addresses: jemma.konig@waikato.ac.nz (J.L. König), judy.bowen@waikato.ac.nz (J. Bowen), annika.hinze@waikato.ac.nz (A. Hinze).
[1] Lecturer (Software Engineering).
[2] Associate Dean Academic & Associate Professor (Software Engineering).
[3] Head of School & Professor (Computer Science).

https://doi.org/10.1016/j.ssci.2025.106958
Received 22 May 2025; Received in revised form 2 July 2025; Accepted 28 July 2025
Available online 1 September 2025

2.1. Fatigue

Fatigue is a physiological state, caused by physical or mental exhaustion, that results in reduced performance capability. Mental fatigue affects mental performance and cognitive processes, and is caused by prolonged cognitive exertion (Marcora et al., 2009). Physical fatigue affects physical performance, and is caused by intense physical exertion (Hawley, 1997). Both are cumulative, and increase the more a task is performed. However, both also lessen after a period of rest.

2.2. Measuring fatigue

There are two main approaches to measuring fatigue: subjective fatigue tests, and objective fatigue tests. Subjective tests measure *perceived* fatigue, which usually involves some form of self-reporting. These types of tests involve participants indicating their perceived level of tiredness, weariness, exhaustion, etc. Holtzer et al. (2017).

In contrast, objective fatigue tests measure quantifiable performance, which usually involves performance-based tests such as reaction speed (Holtzer et al., 2017). The Simple Reaction Time test is one example. In this test, participants are evaluated based on their reaction speed. There are several different versions of the test, from the original stick-drop-test, which involved catching a falling stick and measuring the distance between the start of the stick and where it was caught (Johnson et al., 1985), to computerized tests that show a stimulus on the screen and record the time it takes for the participant to react by pushing a button.

2.3. Fatigue in forestry

Both mental and physical fatigue play a role in the occurrence of incidents in the forestry industry. Mental fatigue is often reported in incident causes such as "a lack of concentration" or "errors in judgment", while physical fatigue can be seen reported in incidents such as "slip, trip, and fall" (Bentley et al., 2005; Bell, 2002; Bentley et al., 2002). Driscoll et al. report "errors of judgment" as one of the leading causes of fatalities in forestry workers (Driscoll et al., 1995), while Bentley et al. (2005, 2002) identify peaks in incident occurrences around mid-morning and mid-afternoon, which they attribute to fatigue.

Very little research has been done with fatigue tests in the forestry industry. Rather than conducting fatigue tests, either on-site or in a laboratory environment, fatigue research in the forestry industry tends to focus on analyzing incident data reports and survey responses (Bentley et al., 2005, 2002; Lilley et al., 2002). The New Zealand forestry industry has a national reporting scheme where forestry companies log any incidents that occur on site (Ministry of Business Innovation Employment, 2012). This type of data has been used by several researchers to identify trends in patterns and/or causes of incidents in forestry. For example, Bentley et al. used incident reports in their 2002 study to investigate skid work injuries (Bentley et al., 2002). Although their primary goal was not to identify fatigue in the forestry industry, they did identify patterns where injuries tended to peak at around 8 am–10 am and 2 pm–3 pm, which they attribute to fatigue.

In a similar study in 2005, Bentley et al. used incident reports to investigate felling injuries (Bentley et al., 2005). In this study, they also found a peak in the times that incidents occurred, this time between 9 am–11 am. They provide the following rationale for why these peaks may be occurring. In New Zealand, forestry work tends to occur in remote locations, which requires workers to drive two hours or more to site each morning. This results in an extended work day. Workers may leave home as early as 5 am and not take a break until lunch. They then tend to work through until the end of the day. The fuel provided by breakfast and lunch can only sustain them for up to 4 h, suggesting that they may be more likely to experience fatigue in the periods before their lunch break and the end of the day (Bentley et al., 2005, 2002). [4]

We found similar patterns in a study we conducted in 2021 (Anonymize, 2021). We analyzed eight years worth of New Zealand incident reports, which we used to identify incident causes that involved worker-failure and worker-fatigue. We focused on the cause of, and time that incidents occurred and found that 70% of incidents could be attributed to worker-failures. This included causes such as *lack of concentration*, *poor technique*, and *poor evaluation of hazards*. Of these incidents, we found that 78% showed indications of worker-fatigue. Worker-fatigue was identified using the time that incidents occurred. We found that worker-failure based incidents showed a strong trend of occurrences peaking at 10 am and 2 pm – indicating, like Bentley et al. (2005, 2002), that fatigue plays a role in incidents in the forestry industry.

As mentioned, most fatigue research in the forestry industry tends to center on either analyzing incident data reports or analyzing survey responses. As an example of the latter, Lilley et al. conducted a survey to investigate the relationship between fatigue and workplace accidents in the New Zealand forestry industry (Lilley et al., 2002). In their study, 367 forestry workers completed a self-administered questionnaire that focused on work/rest patterns, sleep patterns, fatigue experience, the perceived impact of fatigue, and any injury or near injury experiences. The results of the study found fatigue to be common in the forestry industry, with 78% of workers reporting experiencing fatigue at least "sometimes". The results of the survey also showed that some workers reported long working hours and compromised recovery time, and that the number of breaks taken during the day was associated with high levels of fatigue.

Finally, we are aware of one study involving in-situ fatigue tests in the forestry industry (Anonymize, 2019). However, it is a study that we conducted in 2019. In this study we collected reaction time data from forestry workers using two objective fatigue tests. The study was conducted on-site with 15 workers, where reaction test data was collected in the morning before work commenced, around midday after lunch, and in the afternoon after the workers had finished for the day. The study showed high variability in the results of the fatigue tests across days, and across participants, and highlighted some of the difficulties of conducting research with subjective and objective fatigue tests. It is that study which motivated the work described in this paper.

2.4. Fatigue in hazardous industries

The purpose of this paper is to investigate the use of fatigue tests within a forestry context. However, as highlighted in Section 2.3, little has been done in this area. Instead, majority of fatigue test research focuses on other hazardous industries. As an example of this, we conducted two literature scoping reviews: one for fatigue in hazardous industries, and one for fatigue in forestry.

The following search terms were used in Google Scholar to select papers on fatigue based studies in hazardous work environments.

- Subjective fatigue test + hazardous industry
- Objective fatigue test + hazardous industry

Using these search terms, the first five papers (sorted by relevance) that included participant-based studies, hazardous industries, and subjective and/or objective fatigue tests were selected using each search term. This resulted in the selection of ten studies on fatigue in hazardous work environments (Good et al., 2020; Hu and Lodewijks,

[4] The concentration of studies by a single research group (Bentley et al.) highlights not only their significant contribution to the field, but also a broader gap in diversified research efforts. This suggests that fatigue in forestry remains an underexplored area, warranting further empirical investigation using validated testing methods.

Table 1
Review of papers by industry and fatigue test.

Year	Paper	Industry	Fatigue tests
2020	Good et al. (2020)	Military	Subjective & Objective
2020	Hu and Lodewijks (2020)	Driving & Aviation	Subjective
2020	Darbandy et al. (2020)	Construction	Subjective
2017	Fan and Smith (2017)	Rail	Subjective
2004	Hursh et al. (2004)	Military	Subjective & Objective
2000	Balkin et al. (2000)	Driving	Subjective & Objective
1992	Mascord and Heath (1992)	Driving	Objective
1985	Angus and Heslegrave (1985)	Military	Subjective & Objective
1983	Storm (1983)	Aviation	Subjective
1974	Johnson and Naitoh (1974)	Aviation	Subjective & Objective

2020; Darbandy et al., 2020; Fan and Smith, 2017; Hursh et al., 2004; Balkin et al., 2000; Mascord and Heath, 1992; Angus and Heslegrave, 1985; Storm, 1983; Johnson and Naitoh, 1974)

Likewise, the following search terms were used with the same database to select papers on fatigue based studies in the forestry industry.

- Subjective fatigue test + forestry industry
- Objective fatigue test + forestry industry

In contrast to the earlier search for hazardous industries, the forestry industry returned primarily non-related papers under two main categories: surveys on worker fatigue (Lilley et al., 2002; Nakata et al., 2022) and the structural fatigue strength of wood (Yildirim et al., 2015; Ratnasingam et al., 2010). The only paper we were able to identify using these search terms, and that included participant-based studies, the forestry industry, and subjective and/or objective fatigue tests was a paper we wrote involving in-situ fatigue tests in the forestry industry (described earlier in Section 2.3). Given that it describes our own work, we will reference the study within this paper, but focus primarily on the ten papers that were identified for other hazardous industries.

Table 1 shows the ten hazardous industry based papers, including the industries they focus on and whether the studies involved subjective or objective methods for measuring fatigue. Of the ten papers, three were centered on driver-based studies (Balkin et al., 2000; Mascord and Heath, 1992), three on military studies (Hursh et al., 2004; Angus and Heslegrave, 1985; Good et al., 2020), three on aviation (Storm, 1983; Johnson and Naitoh, 1974; Hu and Lodewijks, 2020), one on construction (Darbandy et al., 2020), and one on fatigue in the rail industry (Fan and Smith, 2017). Four papers used only subjective methods of measuring fatigue (Hu and Lodewijks, 2020; Darbandy et al., 2020; Fan and Smith, 2017; Storm, 1983), one used only objective methods of measuring fatigue (Mascord and Heath, 1992), and the remaining five used a combination of both (Good et al., 2020; Hursh et al., 2004; Balkin et al., 2000; Angus and Heslegrave, 1985; Johnson and Naitoh, 1974).

3. The challenges of fatigue testing

Based on a review of the ten papers described above, we have identified four main challenges when working with fatigue tests: (1) variability in test selection, (2) the use of multiple fatigue tests, (3) the introduction of technology, and (4) fatigue testing across high-risk industries.

3.1. Variability in test selection

First, in addition to the two main categories of fatigue test — subjective versus objective tests — there is also significant diversity in the approaches used for each. For example, subjective tests can range from simply ranking your sleepiness in a score from 1 to 7 (Hoddes et al., 1972), to a 29-item positive and negative response (Johnson and Naitoh, 1974). Likewise, objective tests vary from the simplest device, like catching a falling stick or ruler, to a 10 min computerized vigilance test (Reifman et al., 2018). Different researchers have used a variety of different fatigue tests. Yet none seem to address their reasoning for selecting one test over another.

By considering the review studies, we can see this to be the case. For example, Balkin et al. (2000) conducted a study with one subjective fatigue test—the Stanford Sleepiness Scale—and several objective fatigue tests—4-Choice Reaction Time, 10-Choice Reaction Time, Serial Addition and Subtraction, and Psychomotor Vigilance Task. While they describe in detail how each test works, they do not provide any reasoning for why these specific tests were chosen. Angus and Heslegrave (1985) are the same. They also used the Stanford Sleepiness Scale and the 4-Choice Reaction Time test, along with two other subjective tests — the USAF checkcard and the NHRC Mood Scale — and three other objective tests — Logical Reasoning, Encoding/Decoding, and Auditory Vigilance. Again, they gave detailed descriptions of each test, but did not give any indication for the reasoning behind their selection. Finally, Fan and Smith (2017) also used a selection of fatigue tests, including the Psychomotor Vigilance Task, a visual search test, and a logical reasoning test. While they provide their rationale for the study as a whole, like the others they do not give any reasoning behind their selection of the specific fatigue tests that they chose to use. In fact, we have done the same. During our 2019 study, we used the Simple Reaction Time and 4-Choice Reaction Time tests but gave no reasoning for why these tests were chosen over others (Anonymize, 2019). This type of ambiguity amongst researchers raises the question, does it matter which fatigue test we use, and if it does, how do we pick the right one?

3.2. Using multiple tests

Second, researchers tend to use more than one subjective and/or more than one objective test in their study. Angus and Heslegrave (1985) used three different subjective tests in the course of one study, while Balkin et al. (2000) and Fan and Smith (2017) both used multiple objective tests. If subjective tests measure perceived fatigue, while objective tests measure performance, what use would a researcher have in using more than one of each, unless each test measures perceived fatigue and performance to a different degree? While each of these researchers used multiple tests, none specified their reasoning for it. Do different tests measure different levels of fatigue? The study conducted by Balkin et al. (2000) suggests this may be the case. They found that all objective fatigue tests were sensitive to changes in sleep duration, but to varying extents. This further complicates the process of test selection.

3.3. Technological challenges

Third, the emergence of technology has complicated this process even further. Not only are there several types of subjective and objective fatigue tests, but each type now has several different digitized versions. The Simple Reaction Time test, for example, can be conducted using numerous different mobile applications and computer programs, or in its original form where the participant catches a stick. This presents additional challenges, such as questioning the accuracy of different tests, or the reaction speed when performed on different hardware.

3.4. Fatigue testing across high-risk industries

While the reviewed literature provides valuable insights into the use of fatigue tests in hazardous industries, there are important contextual differences that should be noted between aviation, construction, military, transport and forestry operations. Aviation and military contexts often involve structured environments with predictable schedules and centralized control, allowing for tightly managed testing conditions and access to advanced technologies such as eye-tracking and wearable EEG devices. Similarly, transport studies, particularly those involving driving simulators, offer a controlled setting that enables frequent and minimally intrusive assessments. In contrast, forestry operations are typically decentralized, physically intensive, and take place in remote, dynamic, and environmentally challenging conditions. Workers are often engaged in highly physical tasks with limited downtime, tight performance pressures, and elevated safety risks. These constraints make it difficult to implement time-consuming or obtrusive testing procedures. As such, while general findings—such as the variability and sensitivity of fatigue tests—may be relevant, the operational realities of forestry demand more adaptable, low-burden, and context-aware fatigue monitoring solutions. This highlights the need for forestry-specific validation of fatigue tools that have primarily been tested in more controlled or technologically enabled environments.

4. Our investigation

Based on the challenges highlighted above, the purpose of this paper is to investigate a variety of subjective and objective fatigue tests, providing evidence for or against the use of multiple subjective and objective fatigue tests within one study.

As such, we have designed an investigation that evaluates subjective and objective tests. The investigation includes three main focuses, as follows.

1. Test selection
2. In-depth study
3. Larger-scale study

First, when considering the use of fatigue tests in hazardous industries, we noted that while researchers tend to use a variety of tests (and multiple tests within one study) they do not tend to discuss their reasoning behind this choice (discussed in Section 3). As such, we suggest that the first step in our investigating should involve describing and reasoning about our test selection (Section 5).

Second, we conducted an in-depth study with a single participant (Section 6). The study involved the participant conducting a set of fatigue tests every two hours for five consecutive days. Here, we recognize that a single-participant case study is not generalizable. However, we argue that it provides rich detail that can be used to glean insights into this participant's experience, and guide the creation of a larger-scale study.

Third, we conducted a larger scale study with 31 participants (Section 7). This study allowed us to build on the findings from the single-participant study in a more generalizable way, and allowed us to evaluate the use of multiple fatigue tests, drawing comparisons between different subjective tests, different objective tests, and comparing subjective and objective tests with each other (discussed in Section 8).

5. Test selection

In this section, we first outline our test selection criteria. Following this, we give a brief overview of each of the tests that were selected.

Table 2
An evaluation of available subjective and objective fatigue tests.

Test	Type	Valid	Usage	Duration
Stanford Sleepiness Scale	Sub	Y[a]	3	<1 min
The USAF checkcard	Sub	Y[b]	2	<1 min
NHRC Mood Scale	Sub	Y[c]	2	<1 min
Visual Analogue Mood Rating	Sub	Y[d]	1	<1 min
Psychomotor Vigilance Test	Ob	Y[e]	6	10 min
4-Choice Reaction Time	Ob	Y[f]	3	<5 min
Auditory Vigilance	Ob	Y[g]	2	10–30 min
Simple Reaction Time	Ob	Y[h]	2	<1 min
10-Choice Reaction Time	Ob	Y[i]	1	<5 min
Serial Addition and Subtraction	Ob	Y[j]	1	1–10 min
Encoding/Decoding	Ob	Y[k]	1	1–10 min
Visual Search Test	Ob	Y[l]	1	<5 min
Logical Reasoning	Ob	Y[m]	1	1–10 min

[a] Hoddes et al. (1972).
[b] Pearson and Byars (1956).
[c] Johnson and Naitoh (1974).
[d] Luria (1975).
[e] Khitrov et al. (2014).
[f] Deary et al. (2011).
[g] Angus and Heslegrave (1985).
[h] Deary et al. (2011).
[i] Balkin et al. (2000).
[j] Balkin et al. (2000).
[k] Angus and Heslegrave (1985).
[l] Fan and Smith (2017).
[m] Fan and Smith (2017).

5.1. Selection criteria

One of the challenges that we identified earlier centers on the selection of fatigue tests. As such, we propose a set of selection criteria for use in case studies. These criteria have been developed based on the specific requirements of our study. However, the general principle of developing selection criteria can be applied to the requirements of any study. It is our hope that highlighting the challenges in, and importance of, test selection for hazardous work environments, and providing an example set of selection criteria, will encourage other researchers to formally outline their test selections as well.

Our requirements and selection criteria are as follows.

1. Validating the accuracy of a fatigue test does not sit within the scope of this study. As such, any tests that we select should already be shown to reliably measure levels of fatigue.
2. Any tests that we select should be used in at least one (preferably more than one) fatigue study for hazardous work environments.
3. This study will involve the use of multiple fatigue tests, one after the other. As such, the combined time to sit all tests should be conservative, i.e., 15 to 20 min.
4. We would like to include an equal number of subjective and objective tests.

We have used our selection criteria, and the literature discussed in Section 2.4, to select a series of subjective and objective fatigue tests for use in our study. The literature discussed in Section 2.4 centered on ten studies on fatigue in hazardous work environments (Good et al., 2020; Hu and Lodewijks, 2020; Darbandy et al., 2020; Fan and Smith, 2017; Hursh et al., 2004; Balkin et al., 2000; Mascord and Heath, 1992; Angus and Heslegrave, 1985; Storm, 1983; Johnson and Naitoh, 1974). Table 2 lists the fatigue tests that were used in each of these studies, along with information about their validity, usage, and test duration.

To begin, as outlined in our first criterion, validating the accuracy of a fatigue test does not sit within the scope of this study. As such, any tests that we select should already be shown to reliably measure levels of fatigue. The third column in Table 2 shows that all tests have been validated previously, either for the original versions of the

tests (Pearson and Byars, 1956; Johnson and Naitoh, 1974; Hoddes et al., 1972; Luria, 1975), for specific versions of the tests (Deary et al., 2011; Khitrov et al., 2014), or through findings in supported literature (Balkin et al., 2000; Angus and Heslegrave, 1985; Fan and Smith, 2017).

Our second criterion outlines the importance of usage. Any tests that we select should be used in (preferably) more than one fatigue study for hazardous work environments. In Table 2 the tests have been ordered, first by whether they are subjective or objective tests, then by their usage in the ten studies that we are considering. By eliminating tests that were used in a single study only, we can narrow our test selection down to seven fatigue tests: the Stanford Sleepiness Scale, the USAF checkcard, the NHRC Mood Scale, the Psychomotor Vigilance Task, the 4-Choice Reaction Time test, the Auditory Vigilance test, and the Simple Reaction Time test.

Third, this study will be repeated multiple times. As such, the combined time to sit all of the tests should be conservative, i.e., 15 to 20 min. As shown in Table 2, all of the subjective tests take less than one minute to complete. However, some of the objective tests run much longer. The Psychomotor Vigilance Task runs for 10 min and the Auditory Vigilance test runs for 10 to 30 min. Combining all seven tests would result in a test duration of approximately 29 to 49 min, which is too long to meet our third criteria. This suggests that one of the longer tests should be excluded. Luckily the Psychomotor and Auditory tests are both vigilance tests, meaning that the (longer) auditory test can be excluded without affecting the variability of the test selection.

The fourth selection criteria specifies a balance between the number of subjective and objective tests that are selected. This criteria is not essential. However, it would allow us to put equal focus on each type of test during our study and analysis. The previous three selection criteria have already narrowed our test selection down to three subjective and three objective fatigue tests. Therefore these will be our final set. The selected subjective tests are the USAF checkcard, the Stanford Sleepiness Scale (SSS), and the NHRC Mood Scale. The three selected objective tests are the Simple Reaction Time test (SRT), the 4-Choice Reaction Time test (CRT), and the Psychomotor Vigilance Task (PVT). Each of these tests are outlined in detail in Appendix A.

6. Case study 1

In order to investigate subjective and objective fatigue tests, we conducted two case studies. The first case study was an in-depth single-participant study, where the participant conducted fatigue tests every two hours for five consecutive days. This section outlines the methodology and results of this study, while a comparison between this and the second study is discussed in Section 8.

6.1. Methodology

6.1.1. Location

The single-person study involved the participant sitting a selection of fatigue tests every two hours for five consecutive days (discussed further in Section 6.1.4). Given the duration of the study, not all tests could be conducted in the same location. The location of the single-person study was primarily based around a university campus. Most of the tests were conducted in the participant's office. However, the participant also conducted part of the study in a shared computer lab, at home, and on site in a forestry workplace.

6.1.2. Participant

Our participant was a female in her early 30 s who works as an academic researcher at a university. We recognize that this demographic is different to those that would be found on-site in the forestry industry. However, measuring fatigue in a hazardous work environment introduces major challenges (Anonymize, 2019). Although not a forestry worker, our participant had ties to the forestry industry through a current research project, and was visiting forestry workers on-site near the end of the study.

6.1.3. Test selection

As discussed in Section 5, we have used our selection criteria to select three subjective tests: the USAF checkcard (USAF), the Stanford Sleepiness Scale (SSS), and the NHRC Mood scale (NHRC), and three objective tests: the Simple Reaction Time test (SRT), the 4-Choice Reaction Time test (CRT), and the Psychomotor Vigilance Task (PVT). A full description of the tests can be found in Appendix A.

6.1.4. Schedule

The study began on a Monday morning at 8 am, ran for five consecutive days, and ended at 6 pm on Friday evening. The participant performed three subjective tests and three objective tests every two hours (8 am, 10 am, 12 pm, 2 pm, 4 pm, and 6 pm) for all five consecutive days (Monday to Friday). The tests were completed in-between the participant's normal day-to-day activities.

Day-to-day activities included interactive activities, such as lab supervision, assignment marking, and workshops; office based activities, such as research and paper writing; recreational activities such as going for lunch or coffee; and domestic activities, such as school drop offs, cooking dinner and cleaning. Finally, on the last day of the study, the participant traveled out of town to visit forestry workers on site, where she spent time interacting with the forestry workers while also working remotely and attending virtual Zoom meetings. Although the participant is not a forestry worker and was not undertaking forestry tasks, the variability in her daily activities, and changes in activity level, should act as a good indicator of fatigue measurement in situations that are more complex and less controlled than a laboratory environment.

6.1.5. Order of tests

In each sitting, the participant completed the subjective tests first (USAF, SSS, NHRC), followed by the objective tests (SRT, CRT, PVT). The tests were completed in this order so that the latter could not affect the former. For example, it can be imagined that, should the participant feel they performed poorly on the objective tests they may subconsciously carry this through to their answers for the subjective tests. The PVT test was completed last for a similar reason; the PVT test has the longest duration (10 min) and could therefore frustrate the participant and consequently affect any subsequent tests. Finally, the participant was not shown any results until the completion of the study.

6.1.6. Test versions

Each test was completed electronically, on a Dell Latitude 5300 laptop. Here, we list the versions of each test that were used.

- USAF, SSS, and NHRC: the subjective tests were filled out in a Google Sheet document. Each test conformed with the content from the original, the formats of which are shown in Appendix A.
- SRT and CRT: the Simple Reaction Time and 4-Choice Reaction Time tests were conducted using the Deary–Liewald Reaction Time Task software. The Simple Reaction Time test had 20 iterations, with an inter-stimulus interval of 1000 to 3000 ms (default for Deary–Liewald). The 4-Choice Reaction Time test had 40 iterations, again with an inter-stimulus interval of 1000 to 3000 ms (default for Deary–Liewald).
- PVT: the Psychomotor Vigilance Task was conducted using the PC-PVT 2.0 software. The test was run for 10 min with an inter-stimulus interval of 2000 to 10,000 ms.

6.1.7. Diary entry

Finally, the participant was asked to keep a diary for the duration of the study. In this diary, the participant was asked to record the day, time, and location of each test session. They were also asked to briefly describe how they were feeling at the time of the sessions, and record the activities that were undertaken in the two hour time-frame between each test session. Table 3 shows a sample of the participant's diary entries, while the full diary is shown in Appendix B.

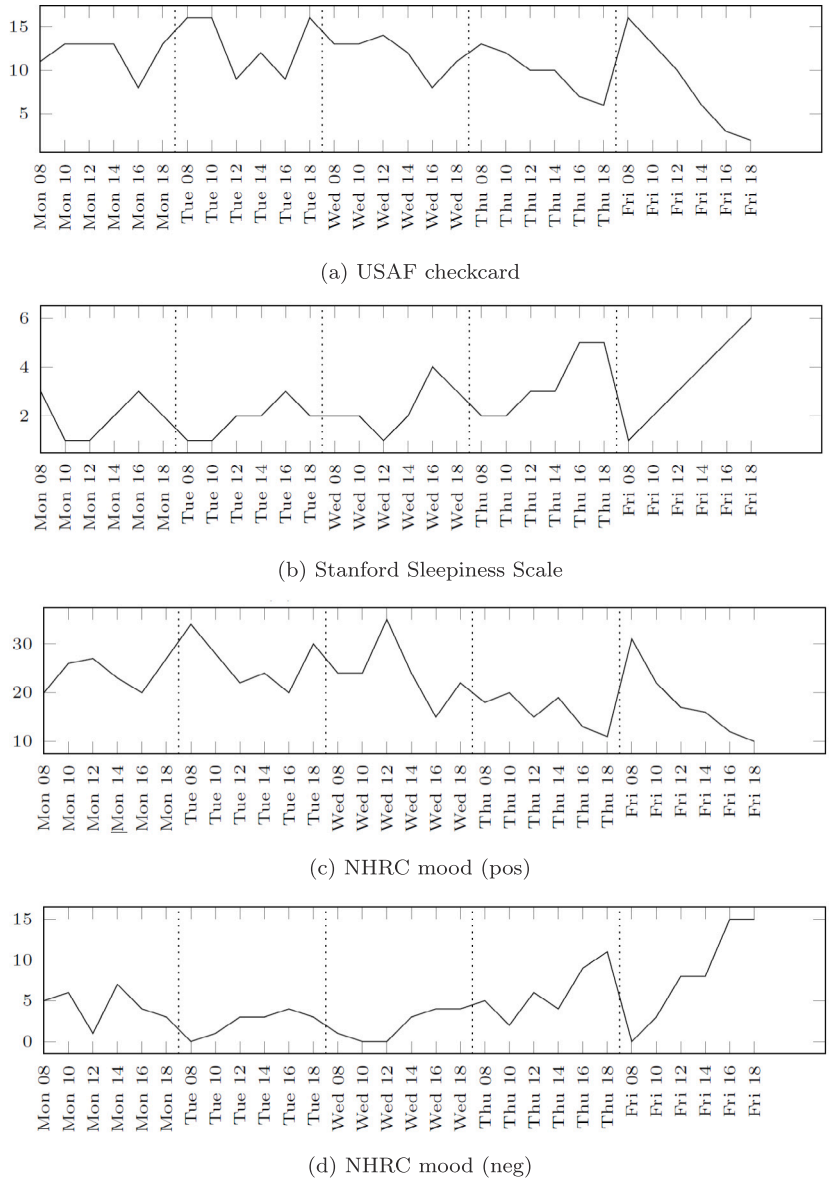


Fig. 1. Subjective fatigue tests.

6.2. Results

6.2.1. Subjective results

Fig. 1 illustrates trends in the subjective results across day and time. Here it should be noted that some tests indicate fatigue with a lower score while others indicate fatigue with a higher score. For example, when the scores of the USAF checkcard and NHRC mood (pos) increase, the scores of SSS and NHRC (neg) decrease.

Although there is some variability between the tests, there is a weak trend on Monday, Tuesday, and Wednesday where fatigue increases throughout the day, then decreases slightly in the evening. This pattern is particularly clear on Monday, Tuesday, and Wednesday for the USAF checkcard, SSS, and NHRC (pos). The participant's diary entries show that the first five test sessions each day (8 am–4 pm) are conducted at work, while the final session is completed at home. On Monday, Tuesday, and Wednesday, the diary entries indicate that the participant spent the two hour gap between the 4 pm and 6 pm sessions at home performing domestic tasks, or out and about shopping. For example, the diary entry at 6 pm on Monday says *"Feeling okay. Finished work, did dishes and rubbish. Cooked dinner"*. The trend of fatigue decreasing between 4 pm and 6 pm suggests that this time away from work is giving the participant time to rest or recover slightly at the end of the day.

A different pattern is apparent on Thursday, where all four tests show a slight but steady increase in fatigue throughout the day and evening. This is reflected in the participant's diary entries, where every entry on Thursday from 8 am to 6 pm mentions being tired: 8 am *"Feeling okay – a little foggy …"*, 10 am *"Quite tired already …"*, 12 pm *"Feeling really tired …"*, 2 pm *"Feeling a little tired .."*, 4 pm *"Feeling really tired again …"*, and 6 pm *"Absolutely exhausted …"*. The final 6 pm diary entry on Thursday says *"Absolutely exhausted! Did workshop then came home, did the dishes and ordered dinner"*. This illustrates that the participant worked later than usual, i.e., "did workshop", which could have attributed to the continued increase in fatigue on Thursday evening, as opposed to the decrease that was identified on Monday, Tuesday, and Wednesday.

Finally, Friday shows a sharp increase in fatigue throughout the day and evening. This trend is apparent in all four test measurements (USAF, SSS, NHRC (pos), and NHRC (neg)). In fact, Friday at 6 pm, is the day and time with the highest level of fatigue for all four tests. This

Table 3
A subset of the dates, times, and details of the study.

Day	Time	Location	Description
Mon	8:00	Office	Conducted in office. Feeling a little foggy. Morning went well, woke up on time, no stress with getting ready etc.
Mon	10:00	<lab>	Conducted in <lab>with distractions. Organizing from 8 to 9, marking from 9 to 10. Feeling time pressured and a little stressed but more awake than this morning.
Mon	12:00	Office	Conducted in office. Feeling pretty good. Marking from 10–11, emails and student zoom meetings from 11–12. Feeling some time pressure.
Mon	14:00	Office	Conducted in office. Feeling okay. Supervised a lab, met with <name>, had a late lunch. Frustrated.
Mon	16:00	Office	Conducted in office. Starting to feel drained. Uploaded week 9 material and assignment 7 to Moodle, and assigned students to groups and answered emails.
Mon	18:00	Home	Conducted at home. Feeling okay. Finished work, did dishes and rubbish. Cooked dinner.

Table 4
Case study 1: the correlation between subjective fatigue tests.

	USAF	SSS	NHRC (pos)	NHRC (neg)
USAF	1.00			
SSS	−0.92	1.00		
NHRC (pos)	0.88	−0.89	1.00	
NHRC (neg)	−0.84	0.85	−0.81	1.00

corresponds with the participant's diary entry for that date and time, stating *"It appears that I was so tired that I forgot to fill in this diary entry. I did conduct the tests though."*. Although the diary entry for the 6 pm test is not overly helpful, the entry from 4 pm shows that the participant was already home and exhausted *"Absolutely exhausted! 3:30 am wake up has made me feel exhausted. Drove back to Hamilton (<name> driving) and headed home."*.

As illustrated above, there appears to be a correlation between subjective fatigue test scores. If one test indicates an increase in fatigue, so do the others, and vice versa. To test this, we have conducted a correlation analysis, where we have calculated the correlation co-efficient between the results of each subjective tests. Table 4 shows the results, where any results that have a correlation co-efficient greater than 0.5 or less than −0.5 have been underlined to show a strong correlation (whether positive or negative).

As was suggested by the raw results, there is a strong correlation between each of the subjective tests. USAF shows a strong positive correlation with NHRC pos (0.88), and a strong negative correlation with SSS (−0.92) and NHRC neg (−0.84). SSS shows a strong positive correlation with NHRC neg (0.85) and a strong negative correlation with USAF (−0.92) and NHRC P (−0.89). NHRC pos shows a strong positive correlation with USAF (0.88) and a strong negative correlation with SSS (−0.89) and NHRC neg (−0.81). NHRC neg shows a strong positive correlation with SSS (0.85) and a strong negative correlation with USAF (−0.84) and NHRC pos (−0.81). This combination of positive (greater than 0.5) and negative (less than 0.5) correlations maps back to the difference in scoring techniques. For example, since USAF indicates fatigue by a higher score, but SSS indicates fatigue by a lower score, they show a negative correlation.

Table 5
Case study 1: the correlation between objective fatigue tests.

	SRT mrt	SRT fs	CRT mrt	CRT fs	PVT mrt	PVT fs
SRT mrt	1.00					
SRT fs	−0.17	1.00				
CRT mrt	0.31	−0.21	1.00			
CRT fs	0.43	0.34	−0.11	1.00		
PVT mrt	0.43	0.10	0.47	0.21	1.00	
PVT fs	0.08	0.01	−0.12	−0.13	−0.30	1.00

6.2.2. Objective results

Fig. 2 shows the mean reaction time (mrt) results for each of the individual objective fatigue tests. In this figure we can see that, where the subjective tests showed consistency in their results, the objective tests were significantly more varied. Nevertheless, they show some associations with the participants diary entries. The slowest mean reaction time (indicating high fatigue) for SRT was recorded on Friday at 12 pm (350.70). We have already discussed that Friday was a difficult day for the participant. The diary entry from this time states *"Parked by a lake. Feeling really tired. The early morning has definitely caught up to me. Wrote a marking schedule and had the staff meeting."*. The slowest for CRT was recorded on Thursday at 6 pm (494.38), when the diary entry states *"Absolutely exhausted! Did workshop then came home, did the dishes and ordered dinner."*. The slowest for PVT was recorded on Friday at 6 pm (439.05), when as already discussed, the participant was fatigued enough not to enter a diary entry *"It appears that I was so tired that I forgot to fill in this diary entry. I did conduct the tests though."*.

Similarly, the fastest mean reaction time (indicating low fatigue) for SRT was recorded on Wednesday at 12 pm (281.45) when the participant stated *"Feeling okay. Less stressed now. Spent the last hour and a half answering student queries and setting up devices"*.; the fastest for CRT (402.92) was recorded on Wednesday at 8 am (*"Feeling good - a little foggy. Morning went well, woke up on time, no stress with getting ready etc"*.); and the fastest for PVT (297.94) was recorded on Monday at 10 am (*"Distractions. Organizing from 8 to 9, marking from 9 to 10. Feeling time pressured and a little stressed but more awake than this morning"*.).

There are also some trends evident between each of the tests. For example, on Monday, both SRT and PVT show a slight peak at 6 pm; on Tuesday they both show a rise and fall from 10 am to 6 pm; on Wednesday they show a peak (one significantly stronger than the other) in the afternoon around 4 pm to 6 pm; on Thursday all three tests (SRT, CRT, and PVT) show two peaks, the first around 12 pm and the second around 6 pm; and on Friday SRT and CRT show a peak between 12 pm and 2 pm.

Although there are some trends evident between the tests, there is no statistical correlation between them. Table 5 shows the result of a correlation co-efficient comparison, where any results that have a correlation co-efficient greater than 0.5 or less than −0.5 have been underlined to show a strong correlation (whether positive or negative). As illustrated in this table, no objective test shows a strong correlation with any another. Although the objective tests showed some similarities in peaks and drops, and showed some associations with the participants diary entries, they were still more varied then we were expecting (as was illustrated with the correlation comparison).

6.2.3. Subjective versus objective results

Finally, we conducted a correlation co-efficient comparison to compare the subjective and objective test results. Table 6 shows the result of a correlation co-efficient comparison, where any results that have a correlation co-efficient greater than 0.5 or less than −0.5 have been underlined to show a strong correlation (whether positive or negative). The top left section of the table shows the correlation co-efficients when comparing the subjective tests with each other (a replication from earlier in Table 4, where we found all subjective tests showed a strong correlation with each other). Likewise, the bottom right shows

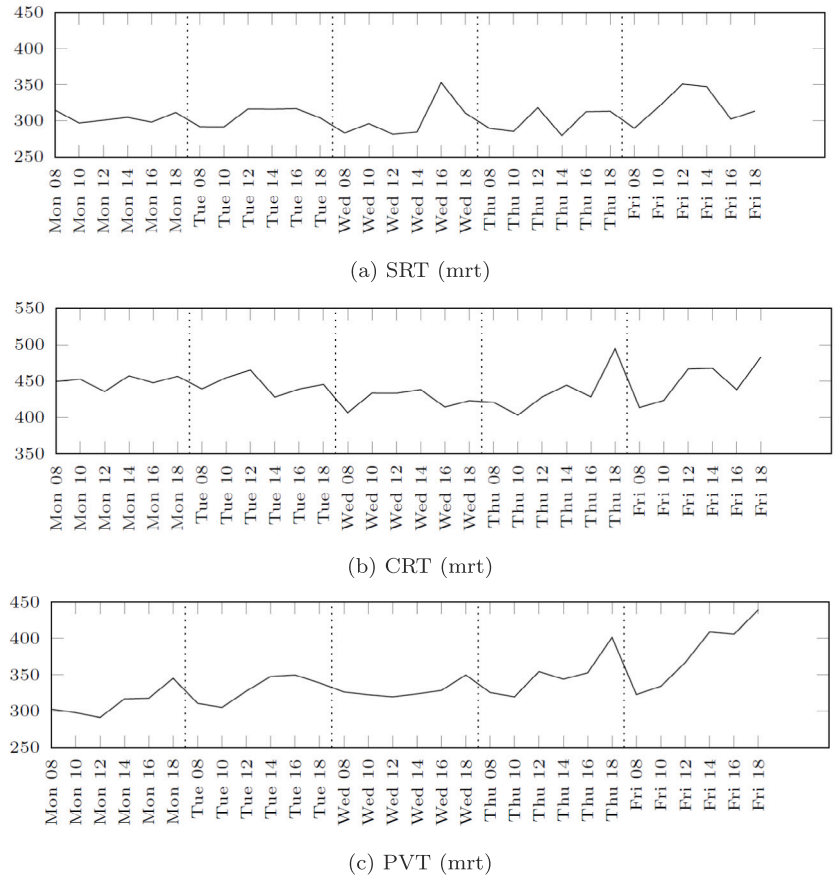


Fig. 2. Objective fatigue tests.

Table 6
Case study 1: the correlation between subjective and objective fatigue tests.

	USAF	SSS	NHRC pos	NHRC neg	SRT mrt	SRT fs	CRT mrt	CRT fs	PVT mrt	PVT fs
USAF	1.00									
SSS	−0.92	1.00								
NHRC pos	0.88	−0.89	1.00							
NHRC neg	−0.84	0.85	−0.81	1.00						
SRT mrt	−0.46	0.47	−0.49	0.39	1.00					
SRT fs	0.10	−0.10	0.22	−0.09	−0.17	1.00				
CRT mrt	−0.41	0.37	−0.27	0.53	0.31	−0.21	1.00			
CRT fs	−0.20	0.13	−0.24	0.17	0.43	0.34	−0.11	1.00		
PVT mrt	−0.79	0.83	−0.69	0.80	0.43	0.10	0.47	0.21	1.00	
PVT fs	0.11	−0.09	0.13	−0.22	0.08	0.01	−0.12	−0.13	−0.30	1.00

the correlation co-efficients when comparing the objective tests with each other (a replication from earlier in Table 5, where we found no objective test shows a strong correlation with any another). The bottom left of the table shows the correlation co-efficients when comparing the subjective tests with the objective tests. As can be seen in the table, the PVT mean reaction time shows a strong correlation with all of the subjective tests (USAF, SSS, NHRC pos, NHRC neg). While no other objective test (SRT or CRT) shows a correlation with any subjective tests, the correlation between PVT and the subjective tests is promising. This could be an artifact from this one participant. However, it could also indicate that PVT provides a comparative measure to subjective fatigue tests. This is discussed further in Section 8.

7. Case study 2

As mentioned earlier, while we argue that the single-participant study provides rich detail that can be used to glean insights into a single participant's experience, a larger-scale study was required in order to build on these findings in a more generalizable way. As such, a second case study was conducted which involved 31 participants completing the same fatigue tests, three times on three different occasions. This section outlines the methodology and results of this study, while a comparison between this and the first study is discussed in Section 8.

7.1. Methodology

7.1.1. Location

This study involved participants sitting a selection of fatigue tests three times on three different occasions. All tests were conducted in the same location, in a quiet room on the university campus. Each participant undertook the study individually, with only the participant and the researcher in the room at the time of the session. This allowed for a quiet space with little to no distraction.

7.1.2. Participants

The study was conducted with 31 university students. Similar to the first case study, we recognize that this demographic is different to

those that would be found on-site in the forestry industry. However, as previously discussed, measuring fatigue in a hazardous work environment introduces major challenges. Although not forestry workers, our participants came in at different times of the day, and after completing a variety of different physical and mental tasks.

7.1.3. Test selection

The same tests were used in both the first and second case study (see Section 6). See Section 5 for the rationale behind the test selection.

7.1.4. Schedule

Each participant was asked to come in to complete the tests on three separate occasions. Test sessions were booked over a three-week period, with some participants completing their sessions within one week, and others taking multiple weeks to complete them. Two participants have been excluded from the study as they did not complete all three sessions.

7.1.5. Order of tests

The tests were completed in the same order for both the first and second case study. In each sitting, participants completed the subjective tests first (USAF, SSS, NHRC), followed by the objective tests (SRT, CRT, PVT). See Section 6 for the rationale for test ordering.

7.1.6. Test versions

The subjective tests were completed using pen and paper, while the objective tests were completed electronically, on a Dell Latitude 5300 laptop. Here, we list the versions of each test that were used.

- USAF, SSS, and NHRC: the subjective tests were filled out in a paper form. Each test matched the original, the formats of which are shown in Appendix A.
- SRT and CRT: the Simple Reaction Time and 4-Choice Reaction Time tests were conducted using the Deary–Liewald Reaction Time Task software. The Simple Reaction Time test had 20 iterations, with an inter-stimulus interval of 1000 to 3000 ms (default for Deary–Liewald). The 4-Choice Reaction Time test had 40 iterations, again with an inter-stimulus interval of 1000 to 3000 ms (default for Deary–Liewald).
- PVT: the Psychomotor Vigilance Task was conducted using the PC-PVT 2.0 software. The test was run for 10 min with an inter-stimulus interval of 2000 to 10,000 ms.

7.2. Results

7.2.1. Subjective results

Table 7 shows a subset of the subjective fatigue test results (for P1 to P5), while Appendix C shows the results for all participants. Here it should be noted that, during the analysis of the first case study, we were able to consider trends in fatigue over time. This is because the single-participant undertook the study regularly (every two hours for five consecutive days) totaling 30 time-series data points, including diary entries for each of these times. In contrast, the larger-scale study only required individual participants to complete the study three times in total. As such, analysis for the second case study will focus on the correlation between scores, rather than any trend in fatigue over time.

As can be seen in the table, some participants gained similar scores across all three sessions (e.g. P1, P3, P4), while others' scores were variable (P2, P5). This could indicate a difference in fatigue for some participants across their sessions. For example, P2 appears to have been the most fatigued during their second session (9 out of 20 for their USAF score, *"A little foggy; not at peak; let down"* for their SSS score, and a higher NHRC negative score). In comparison, during their third session, they had a much higher NHRC score (20 out of 20), selected the highest SSS score (*"Feeling active and vital; alert; wide awake"*.), and had a NHRC negative score of 0.

Table 7
A subset of the subjective fatigue test results (P1–P5)

Participant	Session	USAF	SSS	NHRC pos	NHRC neg
P1	1	17	2	53	2
P1	2	17	1	51	2
P1	3	17	1	53	3
P2	1	13	3	38	5
P2	2	9	4	45	9
P2	3	20	1	52	0
P3	1	14	2	33	2
P3	2	14	1	37	0
P3	3	14	1	35	0
P4	1	16	2	30	0
P4	2	17	2	33	0
P4	3	17	3	29	1
P5	1	10	4	30	9
P5	2	11	3	33	7
P5	3	9	3	30	8

Table 8
The correlation between subjective fatigue tests.

	USAF	SSS	NHRC pos	NHRC neg
USAF	1			
SSS	−0.60	1		
NHRC pos	0.41	−0.59	1	
NHRC neg	−0.59	0.56	−0.31	1

While the above shows variability across sessions, it also shows consistency across subjective fatigue tests for each participant. As discussed, when P2 indicated lower fatigue for one subjective measure, the same was reflected in all tests. Likewise, when P2 indicated higher fatigue for one measure, the same was reflected in all tests. This pattern can be seen for each of the participants included in Table 7, and is supported by Table 8.

Table 8 shows the correlation co-efficient between the results of each subjective test. Any results that have a correlation co-efficient greater than 0.5 or less than −0.5 have been underlined to show a strong correlation (whether positive or negative). As was suggested by the raw results, there is a strong correlation between most of the subjective tests. USAF shows a strong negative correlation with SSS (−0.60) and NHRC neg (−0.59). SSS shows a strong positive correlation with NHRC neg (0.56) and a strong negative correlation with NHRC pos (−0.59). NHRC pos shows a strong negative correlation with SSS (−0.59). NHRC neg shows a strong positive correlation with SSS (0.56) and a strong negative correlation with USAF (−0.59). This combination of positive (greater than 0.5) and negative (less than 0.5) correlations maps back to the difference in scoring techniques. For example, since USAF indicates fatigue by a higher score, but SSS indicates fatigue by a lower score, they show a negative correlation. Finally, NHRC pos and neg do not show a strong correlation (−0.31), and neither do NHRC pos and USAF (0.41).

7.2.2. Objective results

Table 9 shows a subset of the objective fatigue test results (for P1, P2, P3, P4, and P5), while Appendix D shows the results for all participants. This includes the mean reaction time (mtr) for each of the objective fatigue tests (SRT, CRT, and PVT) as well as any false starts (fs). As shown in the table, participants tended to take longer (have a higher mean reaction time) for the CRT test than the SRT and PVT. This is understandable as the CRT test required participants to select one stimuli from a choice of four boxes, rather than simply indicate when a stimuli appeared on the screen. This provision of providing a choice tends to slow down peoples' reaction times.

While familiar trends were able to be seen across the objective tests (i.e. the increased reaction time of CRT, when compared to SRT and

Table 9
A subset of the objective fatigue test results (P1–P5)

Participant	Session	SRT mrt	SRT fs	CRT mrt	CRT fs	PVT mrt	PVT fs
P1	1	281.90	0	477.63	0	315.80	0
P1	2	313.35	0	448.97	1	327.72	1
P1	3	296.60	1	464.95	1		
P2	1	310.55	1	595.80	0		
P2	2	300.65	0	532.58	0	327.42	1
P2	3	348.50	0	544.58	0	405.51	0
P3	1	406.90	0	561.75	0	336.10	1
P3	2	372.25	1	478.92	1	386.11	0
P3	3	340.30	1	429.35	1	393.21	1
P4	1	329.00	0	395.51	1	314.16	0
P4	2	317.80	0	394.84	1	350.10	1
P4	3	340.25	0	416.17	3	355.72	0
P5	1	285.10	0	414.53	1		
P5	2	321.40	0	419.86	1	319.51	0
P5	3	335.15	0	445.08	0	328.77	1

Table 10
The correlation between objective fatigue tests.

	SRT mrt	SRT fs	CRT mrt	CRT fs	PVT mrt	PVT fs
SRT mrt	1.00					
SRT fs	0.13	1.00				
CRT mrt	0.44	0.29	1.00			
CRT fs	−0.17	−0.01	−0.34	1.00		
PVT mrt	0.73	0.15	0.39	−0.10	1.00	
PVT fs	−0.08	0.12	0.08	−0.19	−0.12	1.00

PVT), the results proved much more varied when considering the test across multiple sessions. For P1 for example, their fastest SRT and PVT times occurred during session 1, while their fastest CRT time occurred during session 2. For P2, their fastest SRT, CRT, and PVT times all occurred during session 2. For P3, their fastest SRT and CRT times occurred during session 3, while their fastest PVT time occurred during session 1. There is no observable pattern to these results. Only two participants (P9 and P12) achieved consistent reaction times when considering the tests across multiple sessions. For P9, their fastest SRT, CRT, and PVT times all occurred during their second session. Their second fastest SRT, CRT, and PVT times all occurred during their third session, and their slowest times for all three tests occurred during their first session. Similarly, for P12, their fastest SRT, CRT, and PVT times all occurred during their second session. Their second fastest SRT, CRT, and PVT times all occurred during their first session, and their slowest times for all three tests occurred during their third session. With the exception of these two participants, all others showed variation in speeds across sessions. This is supported by Table 10.

Table 10 shows the correlation co-efficient between the results of each objective test. Any results that have a correlation co-efficient greater than 0.5 or less than −0.5 have been underlined to show a strong correlation (whether positive or negative). As was suggested by the raw results, there is little correlation between the objective tests. SRT mean reaction time shows a strong correlation with PVT mean reaction time (0.73). However, there are no other correlations shown. This is discussed further in Section 8.

7.2.3. Subjective versus objective results

Table 11 shows the correlation co-efficient between the results of each subjective and objective test. Any results that have a correlation co-efficient greater than 0.5 or less than −0.5 have been underlined to show a strong correlation (whether positive or negative). The top left section of the table shows the correlation co-efficients when comparing the subjective tests with each other (a replication from earlier in Table 8, where we found most subjective tests showed a strong correlation with each other). Likewise, the bottom right shows the correlation co-efficients when comparing the objective tests with each other (a replication from earlier in Table 10, where we found only one set of objective tests showed a strong correlation with each other). The bottom left of the table shows the correlation co-efficients when comparing the subjective tests with the objective tests. As can be seen in the table, there are no strong correlations that span between the subjective and objective fatigue tests. In fact, the average correlation between subjective and objective tests (±0.12) is lower than both the average correlation between subjective tests (±0.51) and the average correlation between objective tests (±0.22). This suggests that the earlier finding of a correlation between PVT and the subjective tests (from the single-participant study) may have been an artifact from that one participant.

Finally, Table 12 shows the p-values for the subjective and objective tests. Here, it can be seen that majority of the correlations are statistically significant (p-value < 0.05). In some cases, this refers to strong correlations (greater than 0.5 or less than −0.5) that are statistically significant (p-value < 0.05). A strong and statistically significant correlation suggests a close connection between two variables, meaning that changes in one are very likely to be linked to changes in the other, and this association is unlikely to be due to random chance. This is the case for all of the strong correlations, with the exception of one (SRT (mrt) and PVT (mrt), discussed below). In other cases, this refers to weak correlations (less than 0.5 or greater than −0.5) that are statistically significant (p-value < 0.05). A weak but statistically significant correlation indicates a modest, yet real, association between two variables, with the observed relationship unlikely to have occurred by chance. This is the case for all of the weak correlations, with the exception of one (CRT (fs) and PVT (fs), discussed below). There are two cases that are not statistically significant; one with a strong correlation, and one with a weak correlation. First, SRT (mrt) and PVT (mrt) have a strong correlation (0.73) but this correlation is not statistically significant (p-value 0.8670). A strong correlation between two variables can occur even when it is not statistically significant. This often arises in cases where the sample size is small or the effect size is too modest to be confidently detected amid data variability. Although the correlation coefficient may suggest a strong association, a p-value above the significance threshold indicates that the observed relationship might be the result of random chance. Here, it should be noted that this was the only strong correlation found between any of the objective fatigue tests. Second, CRT (fs) and PVT (fs) have a weak correlation (−0.19) that is not statistically significant. A weak correlation reflects a minimal connection between two variables, whereas a non-significant correlation means there is insufficient evidence to determine that the relationship is statistically meaningful. So, while the correlation between CRT (fs) and PVT (fs) is weak, this may be due to chance.

8. Discussion

This study set out to explore the use of subjective and objective fatigue tests by reviewing existing literature across hazardous industries and conducting two targeted case studies. Together, these studies explored the relationships and discrepancies between different fatigue measurement tools and raise important questions about test selection, technological reliability, and cross-test comparability. In this section, we discuss the findings in relation to measuring subjective fatigue, measuring objective fatigue, and the use of both subjective and objective fatigue tests.

8.1. Subjective measures of fatigue

As discussed in Section 2.2, subjective tests measure *perceived* fatigue, which usually involves some form of self-reporting. These types of tests involve participants indicating their perceived level of tiredness, weariness, exhaustion, etc. When investigating the use of subjective

Table 11
The correlation between subjective and objective fatigue tests.

	USAF	SSS	NHRC pos	NHRC neg	SRT mrt	SRT fs	CRT mrt	CRT fs	PVT mrt	PVT fs
USAF	1.00									
SSS	−0.60	1.00								
NHRC pos	0.41	−0.59	1.00							
NHRC neg	−0.59	0.56	−0.31	1.00						
SRT mrt	−0.05	0.09	−0.17	0.22	1.00					
SRT fs	−0.04	0.01	0.01	0.25	0.13	1.00				
CRT mrt	−0.06	0.13	0.04	0.36	0.44	0.29	1.00			
CRT fs	0.10	−0.08	−0.09	−0.14	−0.17	−0.01	−0.34	1.00		
PVT mrt	−0.11	0.11	0.04	0.23	0.73	0.15	0.39	−0.10	1.00	
PVT fs	0.19	−0.14	0.12	−0.03	−0.08	0.12	0.08	−0.19	−0.12	1.00

Table 12
The p-values between subjective and objective fatigue tests.

	USAF	SSS	NHRC pos	NHRC neg	SRT mrt	SRT fs	CRT mrt	CRT fs	PVT mrt	PVT fs
USAF										
SSS	<0.01									
NHRC pos	<0.01	<0.01								
NHRC neg	<0.01	<0.01	<0.01							
SRT mrt	<0.01	<0.01	<0.01	<0.01						
SRT fs	<0.01	<0.01	<0.01	<0.01	<0.01					
CRT mrt	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01				
CRT fs	<0.01	<0.01	<0.01	<0.01	<0.01	0.011	<0.01			
PVT mrt	<0.01	<0.01	<0.01	<0.01	**0.867**	<0.01	<0.01	<0.01		
PVT fs	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	**0.376**	<0.01	

tests in existing literature, we found that researchers tend to use a variety of subjective tests, and often include more than one test within the same study. As such, we conducted two participatory studies to compare a series of subjective fatigue tests.

The results revealed a high degree of internal consistency across the three subjective tests—USAF Checkcard, Stanford Sleepiness Scale (SSS), and NHRC Mood Scale—suggesting that these tools reliably capture perceived fatigue over time and across individuals. The strong alignment across subjective tests indicates that participants were generally consistent in their perception of fatigue. This suggests that self-reported measures can be a reliable tool for assessing perceived fatigue. From these findings, we suggest that these subjective fatigue tests can be used to measure perceived fatigue interchangeably, and that only one such test would be required within a single study.

8.2. Objective measures of fatigue

As discussed in Section 2.2, objective fatigue tests measure quantifiable performance, which usually involves performance-based tests such as reaction speed. Similar to subjective tests, when investigating the use of objective tests in existing literature, we found that researchers tend to use a variety of objective tests, and often include more than one test within the same study. As such, we compared a series of objective fatigue tests during our two participatory studies.

In contrast to the results from the subjective tests, the objective tests—Simple Reaction Time (SRT), 4-Choice Reaction Time (CRT), and the Psychomotor Vigilance Task (PVT)—showed greater variability, both between participants and within repeated measures. While some patterns were observed (e.g., slower reaction times during periods of high self-reported fatigue), overall correlations between objective tests were weak or inconsistent.

One correlation was found, between SRT and PVT during the larger-scale case study. While the correlation was found not to be statistically significant, the relationship between SRT and PVT makes sense, as both SRT and PVT require the participant to respond to a single stimulus as quickly as possible. One simply runs for a longer duration than the other. Still, we would have expected to see a correlation between SRT, PVT, and CRT as well. Each of the objective fatigue tests (including CRT) measures the same variable, i.e., mean reaction time, albeit using different techniques.

While objective measures are often assumed to provide a more quantifiable assessment of fatigue, our findings suggest that they may each be measuring slightly different aspects of performance. For instance, while both SRT and CRT measure reaction time, the cognitive load and complexity differ, possibly explaining their weak correlation. This would align with the study conducted by Balkin et al. (2000), who found that, while the objective fatigue tests were all sensitive to changes in fatigue, the extent of sensitivity was varied. Still, even with different sensitivity levels, we would have expected enough consistency across the reaction times recorded by each of these tests to show a strong correlation. Nevertheless, this was not the case. This suggests that further research needs to be conducted, this could include investigating different versions of each objective test, and the technology used in the studies.

8.3. Subjective vs. objective measures of fatigue

When comparing the results of subjective fatigue tests with the results of objective fatigue tests, we found very little/weak correlations. During the single-participant study, a strong correlation was found between one of the objective tests (PVT) and all of the subjective tests (USAF, SSS, NHRC). While we theorized at the time that this could be a promising sign, no correlation was found between subjective and objective tests during the larger-scale study. This suggests that the correlation between the subjective tests and PVT was an artifact from the single-participant study.

This is not surprising. As already discussed, subjective tests measure *perceived* fatigue, while objective fatigue tests measure quantifiable performance. The lack of any strong correlation between the two types of tests highlight this — that there is a difference between a participant's perception, and their objective performance. This does not suggest that one is more important than the other, only that it is important to

recognize the difference and include both measures (discussed more in the following section).

8.4. The underlying mechanisms of fatigue measurement

The observed variability in objective test results—both across individuals and within repeated sessions—raises important questions about the underlying mechanisms that differentiate these tools from subjective assessments. One explanation may lie in the nature of what each type of test measures. Subjective fatigue tests capture a participant's holistic self-perception of tiredness, which can remain relatively stable and internally consistent, particularly when prompted by structured scales. In contrast, objective tests measure discrete aspects of cognitive and motor performance (e.g., reaction time, sustained attention), each of which may be differentially affected by fatigue, distraction, motivation, or task engagement at any given moment.

Furthermore, individual differences in how fatigue manifests physiologically and behaviorally can also influence objective test performance. For example, one participant may maintain consistent reaction times despite feeling tired, while another may show immediate performance decline. Environmental factors—such as slight distractions, variations in test setting, or mental load from prior activities—may also disproportionately affect objective test results, especially in ecologically valid but less controlled conditions. Additionally, the temporal sensitivity of each objective test varies; some may detect acute fatigue effects more readily than others, contributing to inconsistencies when multiple objective tests are used in combination.

These findings suggest that while subjective assessments offer stable insight into perceived fatigue, objective tests may require more precise matching to the context, timing, and cognitive demands of the work environment. Future research should investigate which performance domains are most reliably affected by fatigue and explore how contextual factors modulate objective test sensitivity.

8.5. The role of context in fatigue testing

The contrast between our two case studies highlights the importance of context in fatigue testing. In the single-participant longitudinal study, clear patterns emerged across the day and week, with both subjective and objective data aligning with diary entries about daily workload and fatigue. These temporal trends—particularly fatigue peaks around mid-morning and late afternoon—echo previous findings in forestry incident reports. In the larger cross-sectional study, however, variability across test sessions was more pronounced. Without the rich contextual detail provided by the participant diary in Case Study 1, interpreting fluctuations in performance becomes more challenging. This illustrates the limitations of isolated test sessions and underscores the value of longitudinal and context-aware methodologies when studying fatigue in applied settings.

8.6. Challenges in fatigue test selection

Our review and empirical investigation confirm a broader issue within the literature: a lack of transparency and consistency in fatigue test selection. Many studies adopt a suite of subjective and objective tools without clearly articulating the rationale behind their choices. Our proposed selection criteria—centered on validity, prior use in hazardous contexts, test duration, and balance between test types—offer one approach to addressing this gap. However, even with careful selection, the results of this study suggest that combining multiple tests may not necessarily enhance measurement clarity. Instead, researchers must weigh the benefits of redundancy against the potential for participant fatigue and test burden, particularly in applied or field-based research.

In addition, the growing diversity of digital test implementations adds further complexity. Variations in device type, software platform, and test parameters (e.g., inter-stimulus interval) may affect reliability and comparability across studies. This technological heterogeneity raises concerns about the generalizability of findings and reinforces the need for greater standardization or at least clearer reporting of test setup and conditions.

8.7. Implications for future research and practice

These findings have several implications for both research and practice. First, subjective tests—despite concerns over self-report biases—appear to offer stable and consistent insights into fatigue perception, particularly when multiple tools are used together. Second, objective tests should not be assumed to be interchangeable or inherently superior; rather, their utility may depend on task demands and research goals. Third, future studies should report their rationale for test selection, ideally referencing validation studies and explaining how tests align with their operational context.

In applied domains such as forestry, where workers operate in remote locations with limited access to digital infrastructure, the burden of testing must be carefully considered. Short, well-validated subjective tools may be more feasible than more time-consuming objective tests—particularly when coupled with contextual data like work schedules, rest breaks, or task load. However, it is important to recognize that subjective and objective tests measure different aspects of fatigue (perception versus performance) and that one cannot be substituted with the other.

8.8. Limitations and future work

This study has several limitations. While the single-participant case study offers rich longitudinal insight, its findings are not generalizable. Furthermore, both studies involved university staff and students rather than forestry workers, limiting the ecological validity of the results for the target industry. We also acknowledge that sleep quality, caffeine intake, and other lifestyle factors that influence fatigue were not controlled for, particularly in the multi-session study. These limitations were partly due to the practical constraints of this early-stage research and the need to ensure participant safety and feasibility before transitioning to high-risk field settings.

Future work will address these limitations by combining the methodologies of the two studies into an in-depth, longitudinal investigation with forestry workers on-site in operational environments. While laboratory-based studies allow for unrestricted participant access, in-situ studies in hazardous industries are inherently more constrained. Access to workers will be limited by operational schedules, safety protocols, and industry-mandated break times, which may restrict the number and timing of fatigue assessments during the workday. Ideally, fatigue testing would occur at multiple points throughout the day—before and after breaks, at the start and end of shifts—but this must be balanced with the practical realities of the industry context.

In addition to shifting to a more ecologically valid participant population, future work will explore the integration of subjective self-assessments and objective fatigue tests with physiological metrics to enhance sensitivity and relevance. Blink detection and eye-tracking, for example, offer promise for real-time fatigue monitoring. However, their implementation in the forestry sector is currently impractical due to concerns over worker safety, visibility, and the intrusive nature of such technologies. Nevertheless, we will investigate these and alternative, low-profile sensing approaches that are compatible with the demands of physically intensive and safety-critical work.

Overall, this work highlights both the promise and pitfalls of fatigue testing in hazardous industries. Subjective tools provide consistent insight into perceived fatigue, while objective tools offer the potential for nuanced performance tracking—albeit with challenges in consistency and interpretation. To advance fatigue research in real-world environments like forestry, we advocate for transparent test selection practices, balanced test designs, and greater attention to context. Fatigue measurement must move beyond lab-based paradigms to account for the complexity and constraints of applied settings.

		Time/Date	
Make one, and only one, tick for each of the items. Think carefully about how you feel RIGHT NOW.			
Statement	Better than	Same as	Worse than
1. Very lively			
2. Extremely tired			
3. Quite fresh			
4. Slightly pooped			
5. Extremely peppy			
6. Somewhat fresh			
7. Petered out			
8. Very refreshed			
9. Fairly well pooped			
10. Ready to drop			

Fig. A.3. The USAF Checkcard, modeled on the original checkcard seen in Pearson and Byars (1956) and Storm (1983).

9. Conclusion

Fatigue is a major contributor to accidents in hazardous industries like forestry, yet fatigue testing remains inconsistent in both application and interpretation. This study reviewed existing approaches and evaluated six common fatigue tests through two case studies. Subjective tests showed strong internal consistency and aligned closely with participant experiences, while objective tests displayed high variability and weak cross-test correlations. These findings challenge assumptions about the reliability of objective measures and highlight the importance of context in fatigue assessment. In particular, subjective tools—often undervalued—proved effective and practical, especially for use in complex, real-world settings. We recommend that researchers clearly justify their test selections, consider contextual constraints, and balance test depth with participant burden. Our proposed selection criteria offer a starting point for more deliberate fatigue research. Future work should focus on simplifying protocols and validating tools in applied environments to improve safety outcomes in high-risk industries.

CRediT authorship contribution statement

J.L. König: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **J. Bowen:** Writing – original draft, Supervision, Funding acquisition, Conceptualization. **A. Hinze:** Writing – original draft, Supervision, Conceptualization.

Ethics approval

For the single-participant study, ethics was applied for and approved by the University of Waikato Human Research Ethics Committee on 31st March 2020 (HREC(Health)2020#21). For the larger participant study, ethics was applied for and approved by the University of Waikato Human Research Ethics Committee on 4th March 2024 (HREC(HECS)2024#02).

Funding

Our submission describes work that is part of a larger project on safety in the forestry industry, which was funded by the New Zealand Ministry of Business, Innovation & Employment (UOWX1806).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Fatigue tests

A.1. USAF checkcard

The United States Air Force Checkcard (USAF checkcard) was developed by the USAF School of Aerospace Medicine in the 1950s (Pearson and Byars, 1956). It has been validated as collecting "self-ratings of subjective fatigue" from USAF aircrew (Storm, 1983). The checkcard consists of ten statements, such as "very lively", "quite fresh", and "slightly pooped", each of which the user responds to as feeling "better than", "same as", or "worse than". For example, if you feel a little tired, you may rank yourself as feeling the "same as slightly pooped". The ten statements are listed below, while a formatted checkcard can be found in Fig. A.3.

1. Very lively
2. Extremely tired
3. Quite fresh
4. Slightly pooped
5. Extremely peppy
6. Somewhat fresh
7. Petered out
8. Very refreshed
9. Fairly well pooped
10. Ready to drop

The USAF checkcard is scored by summing each response. Any "better than" responses are worth 2 points each, "same as" responses are worth 1 point each, and "worse than" responses are worth zero points. This scoring system results in a score somewhere between 0 and 20, with lower scores indicating higher levels of perceived fatigue (Storm, 1983).

A.2. Stanford Sleepiness Scale

The Stanford Sleepiness Scale (SSS), developed by Hoddes et al. (1972) in 1972, has been validated as measuring perceived (subjective) sleepiness (Hoddes et al., 1972, 1973). In this case, the participant is given a scale with seven levels of 'sleepiness', where they must select the level that best reflects their current state. The scale is shown below, while a formatted scale is shown in Fig. A.4.

Degree of Sleepiness	Scale Rating
1. Feeling active and vital; alert; wide awake.	
2. Functioning at a high level, but not at peak; able to concentrate.	
3. Relaxed; awake; not at full alertness; responsive.	
4. A little foggy; not at peak; let down.	
5. Fogginess; beginning to lose interest in remaining awake; slowed down.	
6. Sleepiness; prefer to be lying down; fighting sleep; woozy.	
7. Almost in reverie; sleep onset soon; lost struggle to remain awake.	

Fig. A.4. The Stanford Sleepiness Scale, modeled on the original scale seen in Hoddes et al. (1973).

Item	Not at all	A little	Quite a bit	Extremely	Item	Not at all	A little	Quite a bit	Extremely
Active					Good natured				
Alert					Grouchy				
Annoyed					Happy				
Carefree					Jittery				
Cheerful					Kind				
Able to concentrate					Lively				
Considerate					Pleasant				
Defiant					Relaxed				
Dependable					Satisfied				
Drowsy					Sleepy				
Dull					Sluggish				
Efficient					Tense				
Friendly					Able to think clearly				
Full of pep					Tired				
Scores		N		P	Able to work hard				

Fig. A.5. The NHRC mood scale, modeled on the original scale seen in Johnson and Naitoh (1974) and Angus and Heslegrave (1985).

1. Feeling active and vital; alert; wide awake.
2. Functioning at a high level, but not at peak; able to concentrate.
3. Relaxed; awake; not at full alertness; responsive.
4. A little foggy; not at peak; let down.
5. Fogginess; beginning to lose interest in remaining awake; slowed down.
6. Sleepiness; prefer to be lying down; fighting sleep; woozy.
7. Almost in reverie; sleep onset soon; lost struggle to remain awake.

The SSS test is scored by assigning an integer value to the 'sleepiness' level that has been selected. The scale ranges from 1 to 7, with 1 being associated with the top-most level (*"Feeling active and vital; alert; wide awake".*) and 7 associated with the bottom-most (*"Almost in reverie; sleep onset soon; lost struggle to remain awake".*) (Hoddes et al., 1973).

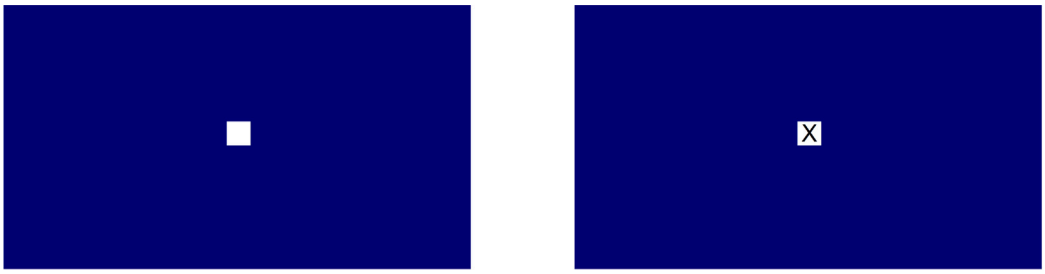
A.3. NHRC mood

The U.S. Naval Health Research Center's Mood Scale (NHRC Mood), developed by a San Diego based Navy research group, has been validated as measuring mood changes associated with lack of sleep or fatigue (Angus and Heslegrave, 1985; Johnson and Naitoh, 1974). The test contains 29 mood-related items, 19 positive (e.g. "active", "alert", "carefree") and 10 negative (e.g. "annoyed", "defiant", "drowsy"). Dependant on how the participant is feeling at the time of the test, each item must be marked as "not at all", "a little", "quite a bit", or "extremely". The mood-related items are shown in Fig. A.5.

The NHRC test is scored by associating an integer value to each of the responses. "Not at all" responses are assigned a score of 0, "a little" responses are assigned a score of 1, "quite a bit" are assigned a score of 2, and "extremely" are assigned a score of 3. The responses from the positive mood-related items should be summed to determine the positive (pos) score, while the responses from the negative items are summed to determine the negative (neg) score. Both positive and negative items were included in the test, as Johnson and Naitoh (1974) found that different subjects tended to be more or less reluctant to answer with positive or negative responses. For example, where military personnel were reluctant to admit negative feelings but willing to admit positive responses, university students showed the opposite approach (Johnson and Naitoh, 1974). As such, the inclusion of both positive and negative items accommodates for different personality types. It is recommended that the two scores not be combined but instead be used separately. A lower positive score, and/or higher negative score, indicates higher levels of fatigue. A higher positive score, and/or lower negative score, indicates lower levels of fatigue.

A.4. Simple Reaction Time

The Simple Reaction Time test (SRT) is an objective fatigue test that measures reaction speed. It requires the participant to respond to a stimulus as quickly as possible (Deary et al., 2011). There are several different versions of this test, from catching a falling ruler and measuring the distance between the start of the ruler and where it was caught (Aranha et al., 2015), to computerized tests that show a stimulus on the screen and record the time it takes for the participant to react by pushing a button. This test is now most often administered on a computer. However, this has resulted in numerous version of the test being developed, including both mobile applications and computer programs, which researchers have found makes comparisons difficult (Deary and Der, 2005; Der and Deary, 2006). As such, Deary



(a) SRT without stimulus (b) SRT with stimulus

Fig. A.6. The Deary–Liewald Simple Reaction Time (SRT) test.

(a) CRT without stimulus (b) CRT with stimulus

Fig. A.7. The Deary–Liewald Choice Reaction Time (CRT) test.

et al. (2011) developed a freely available testkit called "The Deary–Liewald Reaction Time Task".[5] This test kit has been validated in a study involving 150 participants, where the performance of the Deary–Liewald Reaction Time Task was compared with the performance of an existing, widely used reaction time device. The results found that the testkit performed reliably, with a high correlation to the existing device (Deary et al., 2011).

The Deary–Liewald Reaction Time Task includes a Simple Reaction Time test and a 4-Choice Reaction Time test. The Simple Reaction Time test consists of a simple white square placed on a blue background, as shown in Fig. A.6(a). The stimulus is a black 'X' which appears within the square, as shown in Fig. A.6(b). The participant is required to press any key as quickly as possible after the stimulus appears. The stimulus remains on screen until the participant presses a key. If the participant presses a key while there is no stimulus present, it is recorded as a 'false start'. This continues for (default) 20 iterations, with a wait (inter-stimulus interval) of between 1000 and 3000 ms between each iteration. The Deary–Liewald SRT test records the reaction time of each iteration, any false starts, and the mean reaction time across the test.

A.5. Choice reaction time

The Choice Reaction Time test (CRT) is similar to the Simple Reaction Time test, but with the further complication that the participant must respond correctly to one of a number of stimuli (Deary et al., 2011). Again, there are many versions of this test. One approach involves two stimuli, for example, one arrow on the left hand side of the screen and another on the right. In this case the participant is required to click either the left or right button, dependant on the stimuli that is shown (Sharma, 2013). Another approach involves four stimuli, where the participant is required to press one of four buttons, dependant on the stimuli that is shown. Like the SRT test, this variation in versions can make it difficult to draw comparisons (Deary and Der, 2005; Der and Deary, 2006). As such, the Deary–Liewald Reaction Time Task also includes a 4-Choice Reaction Time test.

The Deary–Liewald 4-Choice Reaction Time test consists of four simple white squares placed on a blue background, as shown in Fig. A.7(a). The stimulus is a black 'X' which appears in one of the four squares, as shown in Fig. A.7(b). The participant is required to press the correct key as quickly as possible after the stimulus appears. The stimulus remains on screen until the participant presses a key. If the participant presses a key while there is no stimulus present, it is recorded as a 'false start'. If the participant presses the wrong key while the stimulus is present, it is recorded as an 'incorrect response'. This continues for (default) 40 iterations, with an inter-stimulus interval of 1000 to 3000 ms between each iteration. The Deary–Liewald CRT test records the reaction time of each iteration, any false starts or incorrect responses, and the mean reaction time across the test.

A.6. Psychomotor Vigilance Task

The Psychomotor Vigilance Task (PVT) (Dinges et al., 1997; Dinges and Powell, 1985) is a sustained reaction time task. It is similar to the SRT test, in that it requires a participant to respond to a single stimulus as quickly as possible. However, where SRT and CRT tend to repeat for a short number of iterations (e.g. 20 or 40), the PVT test is run for a sustained period of time (5 to 10 min, depending on the version). It is understood that the sustained duration of the PVT allows for the detection of neuro-behavioral effects of fatigue (Goel et al., 2015). The PVT test originated as a physical, hand-held device, of which the "PVT-192" device was considered the gold standard (Khitrov et al., 2014). However, as with SRT and CRT, the PVT test has also been developed into multiple computer-based versions, one of which is the "PC-PVT" (Reifman et al., 2018).

PC-PVT is a PVT test that has been developed and validated for use on a personal computer (PC) (Khitrov et al., 2014). This version consists of a simple black background. The stimulus is a red number which appears in the center of the screen and counts up in milliseconds, as shown in Fig. A.8. The participant is required to click the mouse button as quickly as possible after the stimulus appears. The stimulus remains on screen until the participant clicks the mouse. If the participant clicks the mouse button while there is no stimulus present, it is recorded as a 'false start'. If the participant takes more than 500 ms to respond, it is

[5] https://datashare.ed.ac.uk/handle/10283/2085.

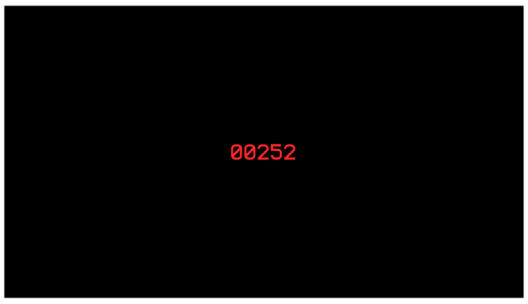


Fig. A.8. The PC-PVT Psychomotor Vigilance Task.

recorded as a 'minor lapse'. If the participant takes more than 1000 ms to respond, it is recorded as a 'major lapse' (Reifman et al., 2018). This continues for a duration of 10 min. The PC-PVT test records the reaction time of each response, any false starts, minor lapses, and major lapses, and the mean reaction time across the test.

Appendix B. Case study 1: Participant diary

Day	Date	Time	Location	Description
Mon	21/09/20	8:00	Office	Feeling a little foggy. Morning went well, woke up on time, no stress with getting ready etc.
Mon	21/09/20	10:00	Lab	Distractions. Organizing from 8 to 9, marking from 9 to 10. Feeling time pressured and a little stressed but more awake than this morning.
Mon	21/09/20	12:00	Office	Feeling pretty good. Marking from 10–11, emails and student zoom meetings from 11–12. Feeling some time pressure.
Mon	21/09/20	14:00	Office	Feeling okay. Supervised a lab, met with <name>, had a late lunch. Frustrated.
Mon	21/09/20	16:00	Office	Starting to feel drained. Uploaded week 9 material and assignment 7 to moodle, and assigned students to groups and answered emails.
Mon	21/09/20	18:00	Home	Feeling okay. Finished work, did dishes and rubbish. Cooked dinner.
Tues	22/09/20	8:00	Office	Feeling good. Morning went well, woke up on time, no stress with getting ready etc.
Tues	22/09/20	10:00	Office	Feeling good. Answered student emails, rearranged <paper>groups.
Tues	22/09/20	12:00	Office	Feeling a bit pooped. Answered student queries, attended <paper>lab.
Tues	22/09/20	14:00	Office	Feeling pretty pooped. Had lunch, ran workshop.
Tues	22/09/20	16:00	Car	Feeling very pooped. Had office hours, drove to meeting.
Tues	22/09/20	18:00	Home	Feeling amped but tired. Met with <name>, signed house papers, drove home, cooked dinner.
Wed	23/09/20	8:00	Office	Feeling good - a little foggy. Morning went well, woke up on time, no stress with getting ready etc.
Wed	23/09/20	10:00	Office	Feeling stressed and under time constraint. Hicchups. Just completed a workshop.
Wed	23/09/20	12:00	Office	Feeling okay. Less stressed now. Spent the last hour and a half answering student queries and setting up devices.
Wed	23/09/20	14:00	Office	Feeling okay - a little pooped. Last two hours spent having lunch and meeting with student
Wed	23/09/20	16:00	Office	Feeling pretty pooped. Spent the last two hours working on the paper.
Wed	23/09/20	18:00	Home	Feeling pooped but happy. Left work, drove to shops, got home and did test
Thurs	24/09/20	8:00	Office	Feeling okay – a little foggy. Morning went well, woke up on time. A little stress in taking <name>to school, Thursday is a late start for her.
Thurs	24/09/20	10:00	Office	Quite tired already! Spent the last two hours answering student emails and working on the incident data paper.
Thurs	24/09/20	12:00	Office	Feeling really tired! Have a headache. Spent the last two hours meeting with <name>and formatting data.
Thurs	24/09/20	14:00	Office	Feeling a little tired but not as bad as earlier. Had lunch and supervised a lab.
Thurs	24/09/20	16:00	Office	Feeling really tired again! Pretty pooped. Spent the last two hours meeting with <name>and working on the paper.
Thurs	24/09/20	18:00	Home	Absolutely exhausted! Did workshop then came home, did the dishes and ordered dinner.

Fri	25/09/20	8:00	Car	On site at the forest. Woke up at 3.30 so early start but I am feeling pretty fresh.
Fri	25/09/20	10:00	Car	Parked up at the lake. Feeling a bit tired now but still good. Spent the last hour and a half on site watching the guys work, then creating a marking schedule.
Fri	25/09/20	12:00	Car	Parked by a lake. Feeling really tired. The early morning has defiantly caught up to me. Wrote a marking schedule and had the staff meeting.
Fri	25/09/20	14:00	Car	Driving away from the site (<name> driving). Tired but not as bad as earlier. Spent the last 2 h having lunch and then heading back to spend some more time on site.
Fri	25/09/20	16:00	Home	Absolutely exhausted! 3:30 am wake up has made me feel exhausted. Drove back to Hamilton (<name> driving) and headed home.
Fri	25/09/20	18:00	Home	It appears that I was so tired that I forgot to fill in this diary entry. I did conduct the tests though.

Appendix C. Case study 2: Subjective results

Participant	Session	USAF	SSS	NHRC pos	NHRC neg
P1	1	17	2	53	2
P1	2	17	1	51	2
P1	3	17	1	53	3
P2	1	13	3	38	5
P2	2	9	4	45	9
P2	3	20	1	52	0
P3	1	14	2	33	2
P3	2	14	1	37	0
P3	3	14	1	35	0
P4	1	16	2	30	0
P4	2	17	2	33	0
P4	3	17	3	29	1
P5	1	10	4	30	9
P5	2	11	3	33	7
P5	3	9	3	30	8
P6	1	9	3	22	11
P6	2	14	2	41	5
P6	3	17	1	42	0
P7	1	10	4	21	5
P7	2	7	5	14	9
P7	3	5	5	9	10
P8	1	15	5	41	5
P8	2	10	3	47	24
P8	3	15	1	52	2
P9	1	3	5	26	12
P9	2	15	2	43	0
P9	3	13	2	46	0
P10	1	15	2	26	2
P10	2	14	2	27	3
P10	3	13	2	17	0
P11	1	14	3	24	0
P11	2	16	5	22	11
P11	3	14	2	31	3
P12	1	12	3	41	7
P12	2	15	3	41	7
P12	3	15	3	28	8
P13	1	11	4	19	6
P13	2	12	3		
P13	3	12	3	22	5
P14	1	6	3	27	4
P14	2	16	2	30	3
P14	3	12	6	15	5
P15	1	12	3	29	2
P15	2	12	4	11	6
P15	3	11	4	20	7
P16	1	9	3	29	3
P16	2	12	2	45	0
P16	3	9	2	41	0
P17	1	13	3	34	6
P17	2	13	3	29	6
P17	3	14	3	28	2
P18	1	11	3	27	5
P18	2	10	2	39	5
P18	3	5	3	41	5
P20	1	5	5	28	15
P20	2	12	3	39	12
P20	3	12	2	35	13
P21	1	11	3	38	7
P21	2	15	1	43	0
P21	3	15	1	52	2
P22	1	12	6	16	9
P22	2	10		22	13
P22	3	10	rule	22	9
P23	1	15	3	22	0
P23	2	15	2	37	0
P23	3	19	1	37	0
P24	1	13	2	37	6
P24	2	13	3	29	9
P24	3	7	7	25	13
P25	1	14	2	37	0
P25	2	15	1	42	0
P25	3	12	3	33	4
P26	1	7	3	29	7
P26	2	8	4	16	15
P26	3	14	2	27	7
P27	1	13	3	23	2
P27	2	10	3	21	8
P27	3	12	3	26	3
P28	1	12	3	55	14
P28	2	8		42	6
P28	3	8		41	6
P29	1	6	3	37	9
P29	2	13	2	38	11
P29	3	7	1	38	5
P30	1	8	3	21	6
P30	2	9	4	17	11
P30	3	3	5	10	16
P31	1	9	3	34	6

P31	2	13	3	33	10
P31	3	6	5	27	11
P32	1	6	5	24	14
P32	2	20	1	41	4
P32	3	13	2	14	10
P33	1	9	2	23	12
P33	2	8	3	19	15
P33	3	16	1	32	7

Appendix D. Case study 2: Objective results

Participant	Session	SRT mrt	SRT fs	CRT mrt	CRT fs	PVT mrt	PVT fs
P1	1	281.90	0	477.63	0	315.80	0
P1	2	313.35	0	448.97	1	327.72	1
P1	3	296.60	1	464.95	1		
P2	1	310.55	1	595.80	0		
P2	2	300.65	0	532.58	0	327.42	1
P2	3	348.50	0	544.58	0	405.51	0
P3	1	406.90	0	561.75	0	336.10	1
P3	2	372.25	1	478.92	1	386.11	0
P3	3	340.30	1	429.35	1	393.21	1
P4	1	329.00	0	395.51	1	314.16	0
P4	2	317.80	0	394.84	1	350.10	1
P4	3	340.25	0	416.17	3	355.72	0
P5	1	285.10	0	414.53	1		
P5	2	321.40	0	419.86	1	319.51	0
P5	3	335.15	0	445.08	0	328.77	1
P6	1	313.90	0	465.00	1	316.86	0
P6	2	320.95	0	398.87	0	314.11	1
P6	3	284.05	0	407.65	0	317.67	4
P7	1	301.65	0	454.97	0	304.40	0
P7	2	298.00	0	505.40	0	314.84	0
P7	3	319.00	0	514.34	0	331.23	0
P8	1	316.95	0	585.03	0	326.25	0
P8	2	326.55	1	608.45	0	346.44	3
P8	3	346.00	0	533.08	0	346.52	0
P9	1	300.45	0	478.54	0	290.14	0
P9	2	296.40	0	453.60	0	274.24	1
P9	3	298.35	0	465.59	0	274.50	0
P10	1	320.90	0	490.38	0	295.32	3
P10	2	290.65	0	533.16	1	300.01	0
P10	3	294.65	0	513.31	1	293.60	0
P11	1	349.60	0	565.68	0	310.26	3
P11	2	354.35	1	609.95	0	339.26	2
P11	3	342.85	0	576.88	0	308.72	2
P12	1	351.55	1	482.08	1	310.38	1
P12	2	335.70	0	459.49	0	309.81	0
P12	3	352.15	0	490.89	1	366.02	0
P13	1	282.70	0	447.15	0		
P13	2	298.00	0	473.24	1	300.33	0
P13	3	334.90	0	446.90	0	302.70	3
P14	1	417.45	0	566.88	0	409.16	0
P14	2	448.90	1	505.72	0	401.79	0
P14	3	371.05	0	515.33	0	399.46	0
P15	1	291.70	0	462.22	2	285.79	0
P15	2	326.45	1	427.16	2	295.91	0
P15	3	327.50	1	442.21	0	291.87	0
P16	1	255.25	0	389.95	0	253.99	0
P16	2	264.25	1	408.58	0	240.65	1
P16	3	258.25	0	394.54	1	245.49	1
P17	1	310.95	0	507.65	0	338.83	0
P17	2	344.95	0	499.00	0	333.72	0
P17	3	381.90	0	516.20	0	337.59	1
P18	1	336.85	0	434.05	0	303.24	0
P18	2	315.90	0	401.82	1	330.39	0
P18	3	301.20	1	446.40	2	356.49	0
P20	1	343.20	0	738.05	0	337.38	0
P20	2	360.15	1	726.95	0		
P20	3	345.75	2	688.18	0	325.04	3
P21	1	294.20	1	525.13	0	310.99	1
P21	2	345.55	0	523.10	0	311.06	0
P21	3	321.85	0	546.55	0	290.19	0
P22	1	387.15	0	491.05	0	418.51	1
P22	2	356.35	1	599.68	0	395.10	0
P22	3	395.65	0	483.25	0	375.70	1
P23	1	278.80	1	422.87	1	267.76	1
P23	2	277.20	0	404.08	4	272.90	0
P23	3	284.20	0	411.95	1	265.81	3
P24	1	335.75	0	433.72	0	326.38	0
P24	2	315.10	0	443.10	0	315.46	0
P24	3	347.90	0	485.42	0	325.59	1
P25	1	282.70	0	448.65	0		
P25	2	283.65	0	455.53	0	306.08	0
P25	3	315.35	0	434.67	0	327.07	0
P26	1	373.65	0	503.61	2	340.71	0
P26	2	344.00	0	463.53	3	344.74	0
P26	3	347.70	0	457.61	3	401.25	0
P27	1	300.40	0	458.05	0	293.90	0
P27	2	321.10	0	462.63	0	293.14	0
P27	3	284.85	0	377.86	2	263.31	1
P28	1	335.55	1	515.55	0	469.33	0
P28	2	374.05	0	616.35	0	469.60	0
P28	3	430.55	0	558.41	0	504.49	1
P29	1	304.60	0	441.59	0	305.66	2
P29	2	289.75	0	439.56	0	301.84	2
P29	3	294.60	0	446.85	0	307.07	0
P30	1	311.95	0	487.33	0	348.43	2
P30	2	333.65	0	488.45	0	280.15	0
P30	3	371.95	2	501.03	0	385.57	0
P31	1	301.80	0	388.78	0	303.86	1
P31	2	309.00	0	386.28	1	254.73	0
P31	3	302.75	0	419.66	2		
P32	1	289.30	0	502.85	0	288.52	0
P32	2	302.80	0	585.32	1	294.71	1
P32	3	326.55	1	560.34	1	293.32	0
P33	1	378.30	0	484.45	0	364.12	0
P33	2	358.05	0	483.62	0	345.86	0
P33	3	399.80	0	523.77	0	345.76	0

Data availability

Data will be made available on request.

References

Angus, R.G., Heslegrave, R.J., 1985. Effects of sleep loss on sustained cognitive performance during a command and control simulation. Behav. Res. Methods Instrum. Comput. 17 (1), 55–67.
2019. Anonymize.
2021. Anonymize.
Aranha, V.P., Joshi, R., Samuel, A.J., Sharma, K., 2015. Catch the moving ruler and estimate reaction time in children. Indian J. Med. Heal. Sci. 2 (1).

Balkin, T., Thome, D., Sing, H., Thomas, M., Redmond, D., Wesensten, N., Williams, J., Hall, S., Belenky, G., et al., 2000. Effects of Sleep Schedules on Commercial Motor Vehicle Driver Performance. Technical Report, Department of Transportation. Federal Motor Carrier Safety ..., United States.
Bell, J.L., 2002. Changes in logging injury rates associated with use of feller-bunchers in West Virginia. J. Saf. Res. 33 (4), 463–471.
Bentley, T.A., Parker, R.J., Ashby, L., 2005. Understanding felling safety in the New Zealand forest industry. Appl. Ergon. 36 (2), 165–175.
Bentley, T.A., Parker, R., Ashby, L., Moore, D., Tappin, D., 2002. The role of the New Zealand forest industry injury surveillance system in a strategic ergonomics, safety and health research programme. Appl. Ergon. 33 (5), 395–403.
Darbandy, M.T., Rostamnezhad, M., Hussain, S., Khosravi, A., Nahavandi, S., Sani, Z.A., et al., 2020. A new approach to detect the physical fatigue utilizing heart rate signals. Res. Cardiovasc. Med. 9 (1), 23.
Deary, I.J., Der, G., 2005. Reaction time, age, and cognitive ability: longitudinal findings from age 16 to 63 years in representative population samples. Aging Neuropsychol. Cogn. 12 (2), 187–215.
Deary, I.J., Liewald, D., Nissan, J., 2011. A free, easy-to-use, computer-based simple and four-choice reaction time programme: the deary-liewald reaction time task. Behav. Res. Methods 43 (1), 258–268.
Der, G., Deary, I.J., 2006. Reaction time age changes and sex differences in adulthood. results from a large, population based study: the uk health and lifestyle survey. Psychol. Aging 21, 62–73.
Dinges, D.F., Pack, F., Williams, K., Gillen, K.A., Powell, J.W., Ott, G.E., Aptowicz, C., Pack, A.I., 1997. Cumulative sleepiness, mood disturbance, and psychomotor vigilance performance decrements during a week of sleep restricted to 4–5 h per night. Sleep 20 (4), 267–277.
Dinges, D.F., Powell, J.W., 1985. Microcomputer analyses of performance on a portable, simple visual rt task during sustained operations. Behav. Res. Methods Instrum. Comput. 17 (6), 652–655.
Driscoll, T.R., Ansari, G., Harrison, J.E., Frommer, M.S., Ruck, E.A., 1995. Traumatic work-related fatalities in forestry and sawmill workers in Australia. J. Saf. Res. 26 (4), 221–233.
Fan, J., Smith, A.P., 2017. The impact of workload and fatigue on performance. In: International Symposium on Human Mental Workload: Models and Applications. Springer, pp. 90–105.
Goel, N., Basner, M., Dinges, D.F., 2015. Phenotyping of neurobehavioral vulnerability to circadian phase during sleep loss. Methods Enzymol. 552, 285–308.
Good, C.H., Brager, A.J., Capaldi, V.F., Mysliwiec, V., 2020. Sleep in the united states military. Neuropsychopharmacology 45 (1), 176–191.
Hawley, J.A., 1997. Fatigue revisited. J. Sports Sci. 15, 245–246.
Hoddes, E., Dement, W., Zarcone, V., 1972. The development and use of the stanford sleepiness scale (SSS). Psychophysiol. 9, 150.
Hoddes, E., Zarcone, V., Smythe, H., Phillips, R., 1973. Quantification of sleepiness: a new approach. Psychophysiol. 10, 431–436.
Holtzer, R., Yuan, J., Verghese, J., Mahoney, J.R., Izzetoglu, M., Wang, C., 2017. Interactions of subjective and objective measures of fatigue defined in the context of brain control of locomotion. J. Gerontol.: Ser. A 72 (3), 417–423.
Hu, X., Lodewijks, G., 2020. Detecting fatigue in car drivers and aircraft pilots by using non-invasive measures: The value of differentiation of sleepiness and mental fatigue. J. Saf. Res. 72, 173–187.
Hursh, S.R., Redmond, D.P., Johnson, M.L., Thorne, D.R., Belenky, G., Balkin, T.J., Storm, W.F., Miller, J.C., Eddy, D.R., 2004. Fatigue models for applied research in warfighting. Aviat. Space Environ. Med. 75 (3), A44–A53.
Johnson, R.C., McClearn, G.E., Yuen, S., Nagoshi, C.T., Ahern, F.M., Cole, R.E., 1985. Galton's data a century later. Am. Psychol. 40 (8), 875.
Johnson, L.C., Naitoh, P., 1974. The Operational Consequences of Sleep Deprivation and Sleep Deficit. Technical Report, ADVISORY GROUP FOR AEROSPACE RESEARCH AND DEVELOPMENT NEUILLY-SUR-SEINE (FRANCE).
Khitrov, M.Y., Laxminarayan, S., Thorsley, D., Ramakrishnan, S., Rajaraman, S., Wesensten, N.J., Reifman, J., 2014. Pc-pvt: a platform for psychomotor vigilance task testing, analysis, and prediction. Behav. Res. Methods 46 (1), 140–147.
Lilley, R., Feyer, A.-M., Kirk, P., Gander, P., 2002. A survey of forest workers in new zealand: Do hours of work, rest, and recovery play a role in accidents and injury? J. Saf. Res. 33 (1), 53–71.
Luria, R.E., 1975. The validity and reliability of the visual analogue mood scale. J. Psychiatr. Res..
Marcora, S.M., Staiano, W., Manning, V., 2009. Mental fatigue impairs physical performance in humans. J. Appl. Physiol..
Mascord, D.J., Heath, R.A., 1992. Behavioral and physiological indices of fatigue in a visual tracking task. J. Saf. Res. 23 (1), 19–25.
Ministry of Business Innovation Employment, 2012. Approved Code of Practice for Safety and Health in Forest Operations. Technical Report, WorkSafe Mahi Haumaru Aotearoa.
Nakata, C., Itaya, A., Inomata, Y., Yamaguchi, H., Yoshida, C., Nakazawa, M., 2022. Working conditions and fatigue in log truck drivers within the japanese forest industry. Int. J. For. Eng. 1–12.
Pearson, R.G., Byars, Jr., G.E., 1956. The Development and Validation of a Checklist for Measuring Subjective Fatigue. Technical Report, School of Aviation Medicine Randolph AFB TX.
Ratnasingam, J., Ioras, F., et al., 2010. Static and fatigue strength of oil palm wood used in furniture. J. Appl. Sci. 10 (11), 986–990.
Reifman, J., Kumar, K., Khitrov, M.Y., Liu, J., Ramakrishnan, S., 2018. Pc-pvt 2.0: An updated platform for psychomotor vigilance task testing, analysis, prediction, and visualization. J. Neurosci. Methods 304, 39–45.
Sharma, A., 2013. Cambridge neuropsychological test automated battery. In: Encyclopedia of Autism Spectrum Disorders. pp. 498–515.
Storm, W.F., 1983. Aircrew Fatigue During Extended Transport, Tactical and Command Post Operations. Technical Report, School Of Aerospace Medicine, Brooks AFB, TX.
Yildirim, M.N., Uysal, B., Ozcifci, A., Ertas, A.H., 2015. Determination of fatigue and static strength of scots pine and beech wood. Wood Res. 60 (4), 679–686.

Jemma König is a Lecturer (Software Engineering) in the School of Computing and Mathematical Sciences, at the University of Waikato, New Zealand. Her research centers on wearable technology for health applications. Prior to this, Jemma's postdoctoral research was part of a larger MBIE funded project called Tini o te Hakituri. Hakituri is centered on investigating technology uses in hazardous work environments. As part of this research, Jemma investigated using wearable technology for fatigue analysis in the forestry industry.

Judy Bowen is an Associate Professor (Software Engineering) in the School of Computing and Mathematical Sciences, at the University of Waikato, New Zealand. She led the MBIE-funded Hakituri project, which develops innovative, ethical and evidence-based wearable monitoring approaches for hazardous industries, and has been working with New Zealand forestry companies and contractors for this work since 2013.

Annika Hinze is a Professor in the School of Computing and Mathematical Sciences, at the University of Waikato, New Zealand, where she heads the Databases and Information Systems (ISDB) group. She is involved in the Hakituri project, which develops an innovative, ethical and evidence-based wearable monitoring approach suitable for the New Zealand workforce. Her research interests include complex event processing, location-based systems and semantic analysis.