Human-Intelligent Systems Integration (2019) 1:71–88
https://doi.org/10.1007/s42454-020-00012-0

RESEARCH ARTICLE

Exploring Bayesian analyses of a small-sample-size factorial design in human systems integration: the effects of pilot incapacitation

Daniela Schmid[1] · Neville A. Stanton[2]

Received: 13 January 2020 / Accepted: 17 September 2020 / Published online: 5 October 2020

Abstract

In contrast to many other areas of science, the highly-topical Bayesian statistics have not been adopted widely to Human Systems Integration (HSI). These methods overcome the weaknesses of the frequentist null-hypothesis significance testing. Bayesian probabilities reflect a direct and quantifiable evidence for either the null- or the alternative hypothesis given the data. A case study on the effects of pilot incapacitation on workload and technology acceptance during different flight phases for single pilot operations (SPO) compared to the contemporary dual crewing configuration demonstrates how Bayesian statistics produce much more transparent and unambiguous results. For example, workload was higher during incapacitation than in normal flight as well as during arrival than in other flight phases. These effects were independent from crew distribution encouraging further research regarding SPO. Finally, Bayesian statistics remain robust against the sample size which is why they provide interpretable results even for small-sample-size designs prevalent in many application domains of HSI.

Keywords Bayesian inference · Single pilot operations · Mental workload · Experimental design · Hypothesis testing

1 Introduction

During the past decade, the Bayesian approach to statistical data analysis and hypothesis testing has become a reasonable alternative to the classical frequentist null-hypothesis significance testing (NHST) in several areas of science (Aczel et al. 2020; van de Schoot et al. 2017). Bayesian statistics are used in a variety of disciplines and include several different statistical procedures. Science, Technology, Engineering, and Mathematics (so-called STEM subjects) have led this progress in statistical hypothesis testing. By way of contrast, Human Systems Integration (HSI) is still dominated by the traditional frequentist NHST in its statistical applications. HSI is defined briefly as the study of sociotechnical systems (Durso et al. 2015). It includes all interdisciplinary, technical, and management processes to integrate human considerations within and across all system elements (International Council on Systems Engineering 2011). Hence, HSI spans across parts of several disciplines such as engineering psychology, human factors and ergonomics (HFE), and cognitive psychology. Within those disciplines, Bayesian statistics have been applied to different extents (Boehm-Davis et al. 2015). In behavioural sciences and psychology, the use of Bayesian statistics has increased and broadened as an alternative approach to traditional frequentist NHST in the context of different statistical frameworks such as hypothesis testing, cognitive models, structural equation modelling (SEM), and meta-analysis as discussed in a special issue of the *Psychonomic Bulletin and Review* (Vandekerckhove et al. 2018).

The application of Bayesian statistics is less widespread in HSI than in classical scientific psychology (Boehm-Davis et al. 2015; Cooper et al. 2012; Salvendy 2012; Stanton et al. 2005; Stanton et al. 2013; van de Schoot et al. 2017). In HSI, the Bayesian framework has only been used in a very small number of cases to evaluate hypotheses independently (Karpinsky et al. 2018; Körber et al. 2018a; Körber et al. 2018b; Lee and Kolodge 2019; Roth 2015; Rubin et al. 2020; Sato et al. 2019; Tear et al. 2020; Yamani and McCarley 2016; Yamani and McCarley 2018) or as a

✉ Daniela Schmid
daniela.schmid64@icloud.com

[1] Independent Researcher, Lower Saxony, Germany

[2] Transportation Research Group, Faculty of Engineering and Physical Sciences, Boldrewood Campus, University of Southampton, Burgess Road, Southampton SO16 7QF, UK

supplement to NHST (Banducci et al. 2016; Chancey et al. 2017; Janczyk et al. 2019) during the last decade. Bayesian methods are used less in HSI compared to other subject domains. Based on these cited examples from HSI retrieved from the most impactful academic journals (for example: *Applied Ergonomics*, *Human Factors*, *Safety Science*), we conclude that the Bayesian approach to hypothesis testing has remained rather a niche methodological framework for HSI. The main driver for applications in HSI was that the Bayes factor (BF) quantifies the strength of the evidence for the null hypothesis compared to the alternative hypothesis. In doing so, Bayesian procedures represent inferential statistical procedures to estimate parameters of an underlying distribution on the basis of the observed distribution (American Psychological Association 2020). They include model specification, the descriptions of the distributions, the calculations of the models, and reporting of the related BFs. In this sense, we take a closer look at the rationale of Bayesian reasoning and how its procedures represent an advantageous addition and even alternative to frequentist NHST in HSI.

In the present paper, we demonstrate the benefits and application of selected Bayesian statistical procedures compared to the corresponding NHST procedures in a case study of HSI. In doing so, we elaborate the main advantages of the emerging trend of using Bayesian statistics in different areas of HSI for different procedures. Furthermore, we focus on the general rationale of the practical application of Bayesian statistics to analyse experimental data. Their mathematical specific procedures, foundations, and critiques are explained in detail elsewhere (Kruschke 2015; Rouder et al. 2018).

2 The benefits of Bayesian statistics for human systems integration

The Bayesian approach to statistical hypothesis testing is superior to the classical NHST in three aspects that arise from the weaknesses of the traditional use of p values (Hubbard and Lindsay 2008) and consequently from two different ways of statistical reasoning. First, in relying on traditional NHST, the researcher aims to test the H_0 given the data $P(H_0|D)$ but does the opposite. Frequentist NHST considers the probability of data given the H_0, $P(D|H_0) \neq P(H_0|D)$, which does not equal the probability of the H_0 given the data. Latter statistical statement is what the researcher aims to investigate but instead they investigate its unequal statement of the data given the H_0. Hence, the NHST's rationale of $P(D|H_0)$ contradicting the researcher's actual goal is called the inverse probability error (Cohen 1994). Frequentists retrieve the p value from an approach to probability that constitutes the data given the H_0. In other words, the sampling distribution is constructed including the H_0 by expecting hypothetical data under H_0. Hence, p values are based on data that were never observed which is why they cannot quantify statistical evidence (Wagenmakers 2007). Accordingly, the p values represent a tail-area integral over these data that have never been observed. The form of their distribution is dependent on, and determined by, the respective test statistics. This is why the rejection of the H_0 is not equivalent to the H_i becoming a more likely entity. The other integral above the remaining area of the distribution is likewise based on data that were never observed. Hence, the data are always tested by assuming a distribution of unobserved data. Against this background, p-values and their statistical significance do not measure the size of an effect or importance of a result because they are dependent upon sample size, effect size, power, and α error probability (Wasserstein and Lazar 2016). Hence, the comparability of their results suffers from these characteristics across different samples and studies also known as the *replicability problem/crisis* (Pashler and Harris 2012; Shrout and Rodgers 2018). Last but not the least, these characteristics have contributed to an exploratory procedure of investigating experimental results, in which the analysis is fine-tuned to the collected data afterwards, instead of pursuing a confirmatory procedure of testing hypotheses (Wagenmakers et al. 2012).

In contrast, Bayesian statistics approach hypothesis testing differently. First, the suggested hypotheses H_i and their corresponding H_0 are both tested on significance according to the H_0 given the data: $P(H_0|D)$. In the following, we provide a concise overview over the logic of Bayesian hypothesis testing (Masson 2011). Bayesian tests compare two hypotheses by converting prior odds (representing the probability for an entity before observation of data) to posterior odds by updating them by the probabilistic information retrieved from the observed data. The posterior odds estimated in this way represent the probability for the H_0 as well as the H_i separately given the data. Hereby, the conditional probabilities for H_0 and H_i are used. A subjective probability is updated when new information D is added:

$$p(H|D) \times p(D) = p(D|H) \times p(H) \begin{cases} p(H|D) = \dfrac{p(H\wedge D)}{p(D)} \xrightarrow{\text{yields}} p(H\wedge D) = p(H|D) \times p(D) \\ p(D|H) = \dfrac{p(H\wedge D)}{p(H)} \xrightarrow{\text{yields}} p(D\wedge H) = p(D|H) \times p(H) \end{cases} \tag{1}$$

In a next step, Bayes theorem is inserted[1]. In doing so, we set both prior odds $p(H_0) = p(H_i) = 1$ what represents common practice when the experimenter cannot or does not make any presumptions about the prior odds for a hypothesis. When previous knowledge exists, this could be incorporated into data analysis at this point. The posterior odds then represent the Bayes factor (BF). This way of calculating BF_{01} is represented in (2). The predictive performance of the H_0 divided by the predictive performance of a H_i:

[1] Bayes theorem:
$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$ (F.1)

$$\Omega = \mathrm{BF}_{01} = \frac{P(H_0|D)}{P(H_i|D)} = \frac{\frac{p(D|H_0) \times p(H_0)}{P(D)}}{\frac{p(D|H_i) \times p(H_i)}{P(D)}}$$

$$= \frac{P(D|H_0)}{P(D|H_i)} \times \frac{P(H_0)}{P(H_i)} = \frac{P(D|H_0)}{P(D|H_i)} \qquad (2)$$

Finally, the reverse of BF_{01} which is BF_{10} has become standard measure (3) likewise.

$$\mathrm{BF}_{10} = \frac{P(H_i|D)}{P(H_0|D)} \qquad (3)$$

All Bayesian statistical procedures are based on this rationale of calculating B_{01} or B_{10}. This Bayes factor represents a relative metric of a hypothesis' predictive quality what represents the second benefit of Bayesian statistics. The corresponding unambiguous categorial evaluation of BF_{01} and BF_{10} is defined in Jeffreys (1961) and cited in the relevant literature of behavioural and social sciences (e.g. Wetzels and Wagenmakers 2012).

Third, Bayesian inference is equally valid for all sample sizes and can always be trusted (Wagenmakers et al. 2018b). This benefit evolves from the fact that Bayesian estimation is coherent referring to the characteristics of probability theory that all inferential statements must be mutually consistent (Lindley 2000). Under this claim, formal logic, mathematics, and calculus of probability exclude any self-contradictions (Eagle 2011). As the observations of Bayesian analyses additionally depend only on the observed data, they are interpretable regardless from their sampling characteristics such as rigidity, flexibility, and size (Lindley 1993; Zondervan-Zwijnenburg et al. 2017). Finally, with an increase in sample size, the probability to discover the real hypothesis tends toward 1 (Goldstein and Wooff 1997). In this sense, Bayesian statistics achieve to overcome the sample size issue that represents an issue in many areas of HSI where experts are required for evaluation of a system (Boring 2006; Borsci et al. 2014; Lewis 1994; Lewis and Sauro 2006). Finally, comparisons have shown the Bayes factor calculation does not overestimate significances like the traditional NHST tends to (Hubbard and Lindsay 2008; Wagenmakers 2007).

Of course, Bayesian analyses are applicable to all statistical procedure such as the comparisons of means of ANOVAs and *t* tests (Körber et al. 2018a; Körber et al. 2018b; Roth 2015; Sato et al. 2019; Yamani and McCarley 2016; Yamani and McCarley 2018), mediation analyses (Karpinsky et al. 2018), linear (Lee and Kolodge 2019; Neyens et al. 2015; Rubin et al. 2020; Tear et al. 2020) and logistic regressions, correlations (Wetzels and Wagenmakers 2012), Bayesian networks (Regens et al. 2015), path analysis (Miranda 2018), structural equation modelling (Smid et al. 2019), meta-analysis (Senior et al. 2016), hierarchical (Bayesian) models (Zhou et al. 2014), and many more. Nonetheless, the general rationale for the calculation of probabilities, the analysis, and its interpretation remains comparable. In the present case study on pilot incapacitation for a reduced-crew in commercial aviation, we show and explain in detail how two exemplary Bayesian statistical procedures can be efficiently applied to a small-sample-size experimental design from HSI. In doing so, we demonstrate the benefits of Bayesian statistics hypothesis testing for HSI on its current state of science to show their practicability for application.

3 A case study: the effects of pilot incapacitation

Pilot incapacitation is one of the five main research challenges in reducing the flight deck crew of an airliner from two to one pilot onboard (Johnson et al. 2012). This possible future concept of operations (ConOps) for commercial air transport has been highly topical in academic research over the past decade (Schmid and Stanton 2020). These so-called single pilot operations (SPO) or reduced-crew operations (RCO), referring to long-haul flight operations that include relief pilots onboard, represent a viable option for future aviation although the concept is currently only in early stages of the design lifecycle and some way off from practical application in commercial aviation (Vu et al. 2018). As such, a remote ground-based support by a specialist operator is often included to alleviate high workload on the pilot in critical phases of flight (such as take-off, landing, and off-nominal and emergency situations). There has been less research investigating pilot incapacitation by using accident models and analyses of the role of the remote-copilot who is involved as a harbour pilot in such a *dedicated support* of the single-piloted aircraft (Schmid and Stanton 2019). *Dedicated support* refers to one remote-copilot supporting one single-pilot at a time. In the present ConOps, the remote-copilot provides mandatory flight planning and navigation support during departure and arrival whereas the single-pilot operates the aircraft on their own during cruise, and only call for support in off-nominal and emergency situations (Schmid et al. 2020; Schmid and Stanton 2019). This ConOps is based on the distribution of workload during the standard flight phases (Schmid 2017). In general, the take-off run, take-off flight path, final approach, missed approach, landing (including the landing roll), and any other phases of flight as determined by the pilot-in-command are defined as "critical" (European Commission 2015). This matches the task demand pattern retrieved from the operating procedure activities during flight. During arrival, the task requirements and workload are highest followed by departure (Federal Aviation Administration 2001). In contrast, cruise is associated with lower task activities and workload. Hence, mandatory support of the single-pilot during departure and arrival is required to distribute workload more evenly.

Pilot incapacitation is any reduction in the medical fitness of a pilot which is likely to jeopardize flight safety (International Civil Aviation Organization [ICAO], 2012, p. I-3-1). This includes any physiological or psychological state that is likely to adversely affect performance. In general, medical certification regulations and checks, the "two communication" rule of communicating health issues (ICAO 2012, p. I-3-6) as well as a specific emergency operating procedure for flight crew incapacitation (Airbus 2011, p. PRO-ABN-NECA-20) deal with this rather rare incident in current two-crew members' operational practice (Schmid and Stanton 2018; Schmid et al. 2018).

In RCO, a single-pilot incapacitation represents an emergency because the sociotechnical system must overcome the lost redundancy of the second pilot. Here, most ConOps employ a pilot health monitoring system to detect decreases in psychological health when not self-reported (Lachter et al. 2017; Schmid et al. 2020; Stanton et al. 2016). In addition, a system monitoring entries into aircraft systems will be required to detect abnormalities in system operations (such as a decrease in safe performance due to subtle incapacitations of the pilot). Schmid and Stanton (2019) suggest and discuss the application of both types of monitoring systems in RCO and worked out a dual-graded alert system to activate remote support. Either an alert for support as pilot monitoring (PM) or an alert for support as pilot flying (PF) to resume command and control is given to the remote-copilot. Such situations represent an emergency in RCO and the remote-copilot would land the aircraft as soon as it is safe to do so.

In the present study, we considered this ConOps for reducing the cockpit crew (Koltz et al. 2015; Schmid et al. 2020) in comparison to contemporary multi-crew operations (MCO) of an Airbus A320's cockpit crew. We investigated the effects of a single-pilot's incapacitation on the remote-copilot's subjective mental workload during the three main flight phases of departure, cruise, and arrival (Young et al. 2015) as well as their technology acceptance regarding the respective ConOps. The remote-copilot served as first officer (FO) and the single-pilot as captain. Taken together with the assumption that incapacitation of the captain causes higher workload on the remaining pilot we assume the following:

- H_1: Pilot incapacitation increases the perceived workload on the remaining co-pilot.
- H_2: The workload for pilot incapacitation during arrival is higher than during cruise.
- H_3: The workload for pilot incapacitation during departure is higher than during cruise.

Since previous research in RCO has shown no generalizable trends in workload differences of a copilot operating at a GS or on the MCO flight deck (Schmid and Stanton 2020), we investigate the influence of crewing configuration on workload exploratively. We used the widely established and well-acknowledged subjective measure of the raw NASA-TLX to assess subjective workload (Hart 2006; Hart and Staveland 1988). The reader interested in workload measurement is referred to the current-state-of-science reference from Young et al. (2015) since it is beyond the scope of the present paper to discuss a standard measure from the field of HSI.

Furthermore, we considered the psychological construct of technology acceptance as it represents a standard measure to continuously inform systems design at all stages of its lifecycle (Regan et al. 2014). In doing so, a human-centred perspective enables to systematically integrate the later users of the system, the pilots into evaluations of the work environment. Against this background, technology acceptance refers to making technology acceptable to its users so that they find it usable in practice. Therefore, we assessed technology acceptance as a link to usage that materialises potential safety effects by *satisfying the pilots needs and requirement* (Adell et al. 2014; Van der Laan et al. 1997). Hereby, we measured acceptance according to this most used definition via the corresponding technology acceptance scale (TAS; Van der Laan et al. 1997). The scale consists of two subscales of *usefulness* and *satisfaction*. It is partially based on the Technology Acceptance Model (TAM) which describes acceptance as an attitude toward using a system via the *perceived usefulness* and *ease of use* (Davis 1986; Davis 1989; Lee et al. 2003). In this sense, we measured the attitude of the pilots to gauge the extent to which the prototypical GS fulfils their perceived needs and requirements for RCO, when compared to MCO. Familiarity with a system and workspace should make the contemporary MCO easier to use simply due to experience (Rahman et al. 2017; Rödel et al. 2014). Thus, we assume for technology acceptance regarding RCO when compared to MCO:

- H_4: The technology acceptance of the MCO cockpit is higher than of the RCO' workplace (GS).

Facing these hypotheses, the methodological aim of the present article is to examine what can Bayesian inference statistics contribute to data analysis and interpretation in addition and comparison to the traditional NHST. In the field of HSI, Bayesian statistics have not been widely adopted yet (Boehm-Davis et al. 2015; Durso et al. 2007; Lee and Kirlik 2013; Salvendy 2012; Wickens et al. 2018). Here, they occupy a methodological niche by being applied probabilistic decision making, signal detection, and learning processes, but less to solely test hypotheses. An assessment of the actual distribution of Bayesian statistics in HSI has not been undertaken until now to the best of the authors' knowledge. Therefore, we apply them according to the recommended practices from behavioural sciences of Aczel et al. (2020), van den Berg et al.

(2019), and van Doorn et al. (2019). In psychology, Bayesian techniques are only referred to as "complex" inferential statistical procedures (American Psychological Association 2020), of which no specific reporting standards are available yet, not even in the field of statistics (Matthews et al. 2017; Stark and Saltelli 2018; Wasserstein and Lazar 2016).

4 Method

4.1 Participants

The sample consisted of N = 10 (M_{age} = 39.40; s_{age} = 12.29; 1 woman, 9 men) commercial pilots holding an active Commercial Pilots Licence (CPL) of which 6 additionally held an Aircraft Transport Pilot Licence (ATPL). All of them possessed either a type rating on an aircraft from Airbus or expertise in Airbus-specific flight operations and procedures. Their flying experience (M_{t} = 7,117.00 h, s_{t} = 6,651.62 h) ranged from 80 to 18,000 h. The participants took part in the study as FO who took over the role of the Pilot Monitoring (PM) during normal flight. In contrast, the captain was a confederate of the experimenters as pilot flying (PF) during normal operations. They acted out the incident of a pilot incapacitation in half of the trials in all three flight phases according to script (laser attack during departure, gastrointestinal illness in cruise, heart attack during arrival) as well as he mocked up the verbal alert to the copilot of the pilot health monitoring system in RCO.

4.2 Design

We employed a complete 2 × 2 × 3 within-subject design by varying the crewing configuration (MCO × RCO), the type of situation (normal × incapacitation), and the flight phase (departure × cruise × arrival). The dependent variables were subjective mental workload of each flight phase (raw NASA-TLX; Hart 2006; Hart and Staveland 1988) and technology acceptance of the two crewing conditions (Technology Acceptance Scale [TAS]; Van der Laan et al. 1997).

4.3 Materials

The MCO were set up in a Generic Experimental Cockpit (GECO) representing an Airbus A320 in its main systems and operating panels except the overhead panel. In contrast, the GS consisted of a mock-up based on an unmanned aerial vehicle's control station called U-Fly. Its main control panels and displays were kept whereas the Primary Flight Display (PFD), the Navigation Display (ND), and the Electronic Centralised Aircraft Monitor (ECAM) status display were displayed on a second screen. Functions unavailable at the current version of the software U-Fly at the GS were mocked up by verbal command into the speakers (to be realized by the confederate captain). One of the experimenters acted as air traffic controller (ATCo) providing the verbal communications with air traffic control (ATC) services via headsets while controlling the whole simulation set up. Captain and (remote-)copilot were interconnected by a hot-mike headset channel whereas the communication with the ATCo required a push-to-talk. Figure 1 depicts the whole experimental setting: the remote-copilot at the GS and the single-pilot on the mocked-up single-pilot GECO. All interactions were recorded and synchronized. The technical details of the modular simulation environment are described elsewhere (Lenz and Schmid 2019).

Each flight was flown from take-off from the departure runway to the touch-down at the destination's runway. We altogether used six different flight scenarios, three for each ConOps (MCO/RCO). One scenario represented a practice run while the two remaining scenarios represented the experimental conditions of the situation (normal/incapacitation). Accordingly, the pilots received all aeronautical charts related to the flight on their current state. The four experimental scenarios were made up comparably. Each of them lasted about 30 min.

4.4 Procedure

The participation in the study lasted about 6 h. The subjects, who took part on voluntary basis, were reimbursed financially for their participation and were informed about the general time course of the study. The flight plan of each scenario was pre-entered into the flight management system (FMS). The runs of MCO and RCO were divided into blocks and were counterbalanced in sequence. The pilots were instructed to fly safely from the departure (take-off) to the destination airport (touch-down on the runway). In RCO, they supported the departure flight up to 10,000 ft. (FL100) and were then instructed to read a neutral newspaper of their choice (either news or travel magazine) to keep them active in a standardized realistic low-level task that is neutral to their flying activity. When instructed by the captain, they should support the same as either PF or PM.

At the end of each flight phase, the simulation was paused and the copilot completed the NASA-TLX (Hart 2006; Hart and Staveland 1988). Then, the crew were instructed to continue the flight under normal conditions. After each crewing condition, the copilot completed the TAS (Van der Laan et al. 1997). Both questionnaires were administered on an iPad via the web application LimeSurvey (LimeSurvey GmbH 2003).

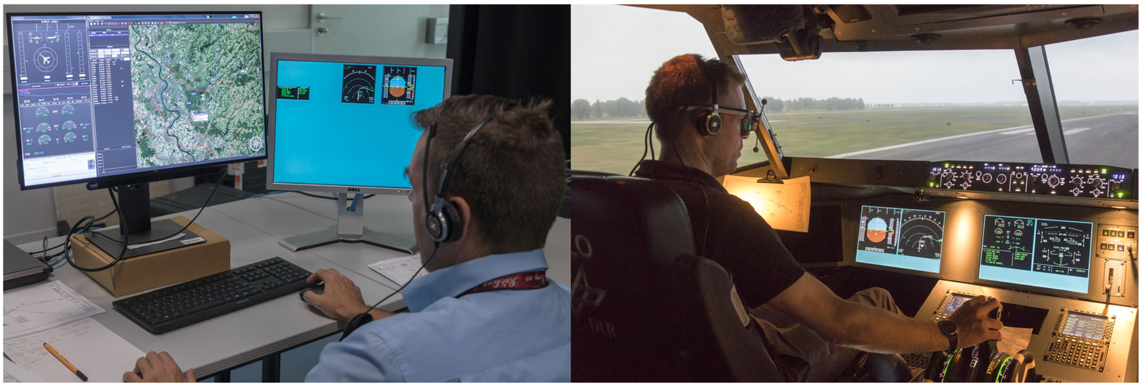

Fig. 1 The two experimental conditions: (left) the remote-copilot at the ground station and (right) the Captain on the multi-crew flight deck (only one pilot on the flight deck in RCO)

5 Results

We analysed the respective scale-sum values of the NASA-TLX and the TAS ratings. To do so, we used the statistic software JASP that provides a range of frequentist NHST procedures and their Bayesian counterparts (JASP Team 2019). Figure 2 summarises the means and standard errors for workload and technology acceptance nested by experimental condition. Both scales showed a very good internal consistency as indicated by their averaged Cronbach's α: the NASA-TLX's α = 0.876 and the TAS' α = 0.931. In general, the Shapiro-Wilk test did not show a deviation from the normal distribution for the workload as well as for the technology acceptance data under all conditions. Bayesian analyses require approximately normally distributed data. The residual's Q-Q-plots confirmed this requirement. Here, we will analyse the data using traditional frequentist NHST in comparison to their respective Bayesian procedures. Parallel analyses of the workload data and the technology acceptance data compared the results of each statistical rationale and explored their benefits and limitation.

5.1 The effects on mental workload: NASA-TLX

5.1.1 The frequentist repeated-measures ANOVA

Table 1 represents the results yielded by applying traditional null hypothesis testing to these data. A repeated measures ANOVA for the complete experimental design confirmed H_1 in all aspects (Table 1). Workload is higher on the copilot when the captain incapacitates during flight. In contrast, the significant main effect of flight phase can only be interpreted because a first-order interaction of crewing × flight phase only tends toward significance. Hence, situation and flight phase influence workload. The post-hoc comparisons confirmed following differences between the levels of the factors. Arrival increased workload when compared to cruise ($t(9)$ = 3.543; p_{holm} = 0.007**; Cohen's d = 1.120) whereas workload during cruise and departure ($t(9)$ = 1.771; p_{holm} = 0.187; Cohen's d = – 0.56) as well as during departure and arrival did not differ ($t(9)$ = – 1.771; p_{holm} = 0.187; Cohen's d = – 0.56). This supports an acceptance of H_2 and a rejection of H_3. The tendency of the interaction

Fig. 2 Subjective ratings for workload and technology acceptance for multi- and reduced-crew operations

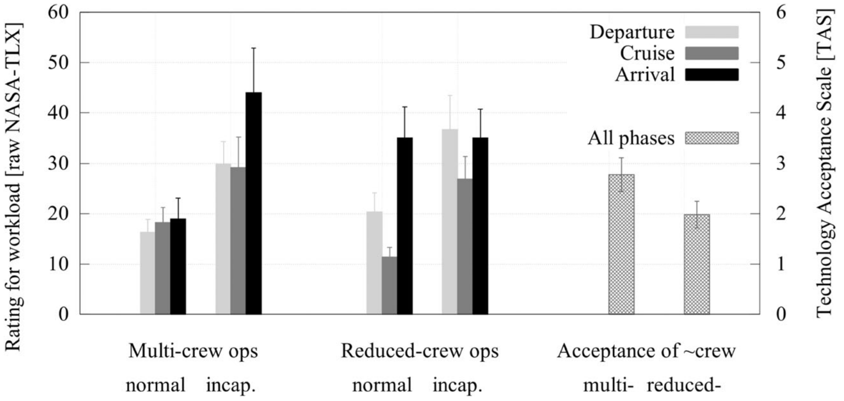

Table 1 The results calculated according to the traditional frequentist approach of statistical inference testing

Measure	Statistical procedure	Inference statistics				
		Effect	Test (df)	p	Effect size	
					η_P^2	η_G^2
Workload rating [raw NASA-TLX][a]	ANOVA (2 × 2 × 3) for repeated measures	Situation	$F(1,9) = 8.98$	.015*	.50	.15
		Flight phase	$F(1,9) = 6.28$	.009**	.41	.05
		Crewing × flight phase	$F(2,18) = 2.78$	.089	.24	.01
		Crewing × situation × flight phase	$F(2,18) = 12.65$	< .001**	.58	.05
	Durbin Test (2 × 2 × 3)[b]	Situation	$X^2(1,10) = 11.39$ $F(11,109) = 12.60$	< .001** < .001**	–	–
		Flight phase	$X^2(2,10) = 7.07$ $F(11,108) = 3.71$	.029* < .001**	–	–
Technology acceptance [TAS]					Cohen's ***d***	
	Paired-sample t test	MCO > RCO	$t(9) = 2.358$	0.021*	.746	
					r_{rb}	
	Wilcoxon signed-rank test	MCO > RCO	$Z = 40.000$	.022**	.778	

Tendencies toward significance were included into the table starting at $p < .10$

[a] The assumption of normality is met (the distributions of the workload and acceptance rating data showed all non-significant $p_i > .05$. according to the Shapiro-Wilk test)

[b] None of the Conover's post-hoc tests yielded significant results ($p_i > .05$)

crewing × flight phase indicates that crewing might affect in combination with flight phase workload, but a conclusion cannot be drawn because of the small sample size. We do not interpret this tendency further.

At the end, when we meaningfully interpret the size of the significant effects, we have to consider their effect sizes. The recommended effect size for a repeated measures ANOVA is η_G^2 for two reasons. The η_P^2 excludes the subject variance which is included into its generalised measure (η_G^2) and consequently neglects design features that affect the effect sizes (Bakeman 2005; Olejnik and Algina 2003). This leads to an overestimation of the effect size when using η_P^2 only and prevents comparability and general classification. Hence, the η_G^2 (Table 1) gives the following interpretation of the ANOVA's effects according to Cohen (1988). Accordingly, all ANOVA's effects can be labelled as small effects. When neglecting interindividual differences by looking at η_P^2, the main effect of situation (H_1) turns out to be medium (Funder and Ozer 2019). In contrast, the η_P^2 of the two-way interaction as medium represents an overestimation. Nonetheless, the confirmation of H_2 stating workload in arrival was higher than during cruise is a medium effect. Since previous statistical knowledge on the effect size of a pilot incapacitation and its effects in reduced-crewing are absent, we deemed the broad categorisation as sufficient to label the statistical effects in size. Its practical relevance cannot be retrieved from that categorisation (Ellis 2009).

5.1.2 The Bayesian Repeated-measures ANOVA

The data from the NASA-TLX were analysed using a Bayesian repeated-measures ANOVA (van den Berg et al. 2019). The effect of pilot incapacitation on workload has not been investigated before which is why no informative prior effect sizes can be proposed. Both hypotheses are estimated as equally likely in terms of their prior odds. Hence, the default JASP prior for fixed effects were used (r scale prior width = 0.5). Table 2 summarises the results of the Bayesian ANOVA. We tested the predictive performance of a particular hypothesis against the null model (H_0). The categorical predictor variables were crewing, situation, and flight phase according to the experimental design. Hereby, all plausible models were considered in JASP. We excluded all models that showed substantial evidence for H_0 (Jeffreys 1961). Furthermore, we excluded models that contain interactions without the corresponding main effects because they are considered as implausible[2] (Rouder et al. 2016). Table 2 summarizes the 14 models that are left over. We are interested in these models because they show at least a substantial evidence (Jeffreys 1961) for the hypothesis that the combination of effects is true. The best of these 14 models shown at the top consists of the main effect of situation and flight phase ($BF_{10} = 132{,}832.778$). A separate comparison of all other 13 models' predictive performance in

[2] Implausible models include an interaction but not the corresponding main effects (Rouder et al. 2016). They are considered implausible because they rely on picking the exact levels so that the true main effects perfectly cancel. Implausible effects are not considered in JASP as well as in the present paper.

Table 2 Model comparison of the Bayesian ANOVA for the NASA-TLX

No.	Models	P(M)	P(M|data)	BF_M	BF_{10}	Error %
1	Situation + Flight phase	0.053	0.500	17.979	132,832.778	9.148
2	Situation	0.053	0.138	2.881	36,672.798	0.627
3	Crewing + Situation + Flight phase	0.053	0.121	2.484	32,232.365	2.388
4	Situation + Flight phase + Situation × Flight phase	0.053	0.077	1.499	20,434.904	1.766
5	Crewing + Situation	0.053	0.038	0.704	10,000.065	2.897
6	Crewing + Situation + Flight phase + Crewing × Situation	0.053	0.034	0.641	9,134.365	4.254
7	Crewing + Situation + Flight phase + Crewing × Flight phase	0.053	0.031	0.567	8,118.091	2.873
8	Crewing + Situation + Flight phase + Situation × Flight phase	0.053	0.023	0.417	6,023.819	4.269
9	Crewing + Situation + Crewing × Situation	0.053	0.010	0.188	2,750.336	2.856
10	Crewing + Situation + Flight phase + Crewing × Situation + Crewing × Flight phase	0.053	0.008	0.152	2,226.171	2.571
11	Crewing + Situation + Flight phase + Crewing × Situation + Crewing × Flight phase + Situation × Flight phase + Crewing × Situation × Flight phase	0.053	0.008	0.145	2,129.507	6.771
12	Crewing + Situation + Flight phase + Crewing × Situation + Situation × Flight phase	0.053	0.005	0.096	1,412.642	1.927
13	Crewing + Situation + Flight phase + Crewing × Flight phase + Situation × Flight phase	0.053	0.005	0.095	1,394.805	2.578
14	Crewing + Situation + Flight phase + Crewing × Situation + Crewing × Flight phase + Situation × Flight phase	0.053	0.002	0.031	452.204	12.633

According to Jeffreys (1961) categorisation, $BF_{10} = [0.3;1]$ are excluded because they show anecdotal evidence for H_0 or H_i. Those models with $BF_{10} <$ 0.3 are excluded as well because they show substantial to decisive evidence for H_0. The order is "compare to null model"

terms of BFs to this best model showed at least a substantial evidence for lacking in predictive value when compared to it (their $BF_{10} \leq 0.33$; not presented here for reasons of space). Thus, we chose the models that are interpretable and further analysed the single effects' contributions to Bayesian significance.

In a second step, we analysed the Bayesian ANOVA's effects in Table 3. The results were averaged across all models to examine for each predictor the prior and posterior inclusion probabilities represented by the inclusion Bayes factor (BF_{incl}). In doing so, we are able to evaluate the effect of main effects and interactions on the calculation of BF_{10} of Table 2. Based on a categorial analysis (Jeffreys 1961), we only interpret the main effect of situation and flight phase because all other effect substantially support H_0 (their BF_{incl} [0.123;0.295]). They have not changed the odds in favour of the model that include them as possible predictor. The main effect of situation decisively changes the posterior odds in favour for the alternative hypothesis whereas the main effect of flight phase only anecdotally contributes to it.

As third step, we analysed the Bayesian ANOVA's effects across all matched models which is referred to as the *Baws factor* (Mathôt 2017). This strengthens the claim of analysing only plausible effects. Hereafter, we call the BF for inclusion of matched models' effects BF_{Baws} to avoid confusion. In general, the BF_{Baws} presented in Table 4 lead to the slightly

Table 3 Effect analysis of the Bayesian ANOVA for the NASA-TLX across all models

Effects	P(incl)[a]	P(incl|data)[b]	BF_{incl}[c]
Crewing	0.737	0.285	0.143
Situation	0.737	1.000	30,451.033
Flight phase	0.737	0.814	1.564
Crewing × Situation	0.316	0.068	0.158
Crewing × Flight phase	0.316	0.054	0.123
Situation × Flight phase	0.316	0.120	0.295
Crewing × Situation × Flight phase	0.053	0.008	0.145

[a] Probability that a predictor is included into the model before seeing the data

[b] Probability that a predictor is included into the model after seeing the data

[c] A quantification of the change from the prior inclusion odds to the posterior inclusion odds. It can be interpreted as the evidence from the data for including a predictor

Table 4 Baws factor (= inclusion probabilities for matched models only) analysis of the concerned effects of the Bayesian ANOVA for the NASA-TLX

Term	*P*(incl)a	$P(\text{incl}\mid\text{data})$b	Baws factorc
Crewing	0.263	0.182	0.254
Situation	0.263	0.827	70,521.142
Flight phase	0.263	0.655	3.525
Crewing × Situation	0.263	0.060	0.277
Crewing × Flight phase	0.263	0.046	0.250
Situation × Flight phase	0.263	0.112	0.161
Crewing × Situation × Flight phase	0.053	0.008	4.709

[a] Models that contain the effect of interest, but no interactions with the effect of interest

[b] Models that contain the effect of interest and interactions with the effect of interest.

[c] The Baws factor is the probability for inclusion of matched models only. Hereby, matched models refer to models that include the effect of interest but no interactions with the effect of interest calculated by JASP

different conclusions as when considering all models as above. Here, the situation (BF_{Baws} = 70,521.142) decisively influences workload as well but flight phase ($BF_{\text{Baws}} = 3.525$) influences workload anecdotally to decisively. Therefore, we have to consider single comparisons regarding the flight phases to examine this hint for a possible predictive value of some flight phase on workload.

In a fourth step, the post-hoc tests' results specified the influence of flight phase on workload. Table 5 represents those results which show at least an anecdotal evidence for an alternative hypothesis. The results that show no evidence ($BF_{10,U}$) or at least anecdotal evidence for H_0 were omitted because their interpretation is that there is no evidence for a difference between the posterior odds' distributions of the predictor's levels. According to Table 5, we conclude that pilot incapacitation decisively increases workload compared to normal flight. Above all, there is very strong evidence for H_2 that workload is higher during arrival than during cruise. Nonetheless, there remains only anecdotal evidence that workload during departure is higher than during cruise (H_3).

In the end, we investigated the magnitude of the relations by considering the posterior distributions of the effects. The model-averaged posterior distributions of each level of the effects can be described by using four statistics: posterior mean, posterior standard deviation, and the lower and upper bound of the 95% credible interval. The posterior distributions for the main effect of situation on workload are shown in Fig. 3. Consequently, the effect of situation is for normal flight M_{Normal} = 5.776 (95% CI [– 7.990; – 3.641] and for incapacitation $M_{\text{Incapacitation}}$ = – 5.776 (95% CI [3.542; 7.868]). The effect of arrival within flight phase is only about M_{Arrival} = 3.434 (95% CI [0.471; 6.407]) whose posterior distribution is displayed in Fig. 4. For example, a posterior estimate for the subjective workload ratings of a particular condition can be calculated by adding the posterior mean of the intercept $M_{\text{intercept}}$ = 23.706 to the M_i of the particular no-effect condition. For example, the posterior estimate for the incapacitation condition is 29.482 whereas the posterior estimate for the normal condition is 17.93. In this way, we can estimate the subjective workload of each condition under the given model obtained from the data. This leads us to question the model fit in general. Overall, the model averaged posterior R^2 = 0.485 with a 95% CI [0.357; 0.584] represents a strong model fit. R^2 is an accurate measure in the Bayesian context although the predictive value for new data is less accurate than for the data that were used to fit the model (Gelman et al. 2019).

In sum, the Bayesian repeated-measures ANOVA revealed that the situation independently affected the remote-copilot's

Table 5 Post-hoc comparisons for the repeated-measures ANOVA that contain an evidence for a H_i ($BF_{10,U} \geq 1$)

Comparisons		Prior odds	Posterior odds	$BF_{10,U}$	Error %
Normal	Incapacitation	1.000	9304.094	9304.094	$1.688e^{-7}$
Departure	Cruise	0.587	1.300	2.213	$1.443e^{-7}$
Departure	Arrival	0.587	0.473	0.805	$1.424\ e^{-6}$
Cruise	Arrival	0.587	0.473	35.706	$3.157e^{-5}$

The posterior odds have been corrected for multiple testing by fixing to 0.5 the prior probability that the null hypothesis holds across all comparisons (Westfall et al. 1997). Individual comparisons are based on the default *t* test with a Cauchy (0, *r* = 1/sqrt(2)) prior. The "U" in the Bayes factor denotes that it is uncorrected

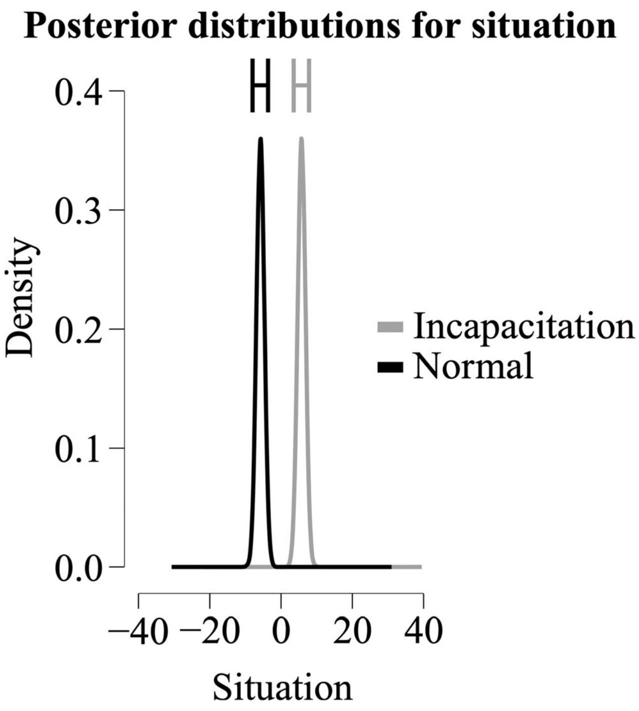

Fig. 3 The posterior distributions and confidence intervals (95%) for the main effect of situation on workload

workload (BF_{incl} = 30,451.033; BF_{Baws} = 70,521.142; $BF_{10,U}$ = 9304.094) which is higher when an incapacitation occurred. Flight phase on its own just had a marginal influence (BF_{incl} = 1.564; BF_{Baws} = 3.525) as further elaborated in post-hoc tests. The flight phase of arrival very strongly increased workload when compared to cruise ($BF_{10,U}$ = 35.706) but not when compared to departure ($BF_{10,U}$ = 0.805). In contrast, departure only anecdotally increased workload when compared to cruise ($BF_{10,U}$ = 2.213) which is why we do not interpret this particular difference further. In this sense, H_1 and H_2 are supported whereas H_3 is rejected.

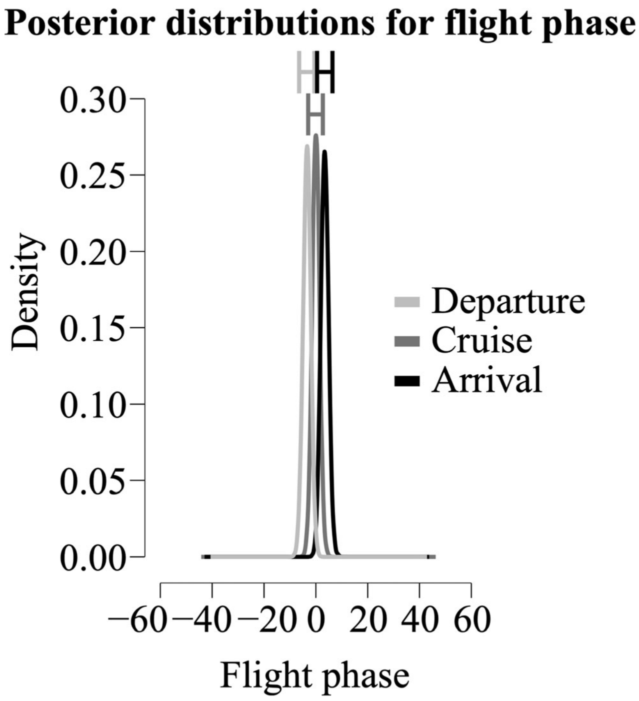

Fig. 4 The posterior distributions and confidence intervals (95%) for the main effect of flight phase on workload

5.1.3 A comparison of the Frequentist and Bayesian repeated-measures ANOVA

Table 6 shows the summary on the frequentist NHST repeated-measures ANOVA. Although there are no differences in the final conclusion regarding the hypotheses, both types of ANOVAs differ in their reliability and validity. NHST yielded simple and all-or-nothing results on significances of workload by the use of *p*-values: accept H_1 (incapacitation > normal) and H_2 (arrival > cruise), reject H_3 (departure > cruise). The effect size cannot be interpreted on their own in the present study due to the lack of reference for practice. Hence, we remain uncertain about the robustness of the results. This is mainly due to the characteristics of NHST judging the test statistics on the base of whether they fall into a prespecified region defined by two error rates (type 1, α, false positive; type 2, β, false negative). This estimation does not represent the data that were observed because it is constructed on the base of assuming hypotheses.

Hence, the Bayesian ANOVA outperformed the frequentist equivalent. The Bayes factors directly quantified how much each factor's level contributed to the posterior distributions of each model of the possible hypotheses. Thus, they directly provide an evidence for the size of an effect by quantifying an updated state of belief by the observed data regarding a hypothesis. Accordingly, H_1 and H_2 are decisively assumed as true whereas H_3 merely shows anecdotal evidence that cannot be interpreted further.

5.2 The effects on technology acceptance: the technology acceptance scale

5.2.1 The Frequentist paired-sample *t* test

A paired-sample *t* test and the Wilcoxon signed-rank test were conducted to examine the effects of crewing on the technology acceptance measured by the TAS. Table 1 shows the results that technology acceptance of MCO is higher than for RCO ($t(9) = 2.358$; $p = 0.021$*; Cohen's $d = .746$) which is why the H_4 can be accepted. Crewing influences technology acceptance. This effect is interpreted as medium (Cohen 1988; Funder and Ozer 2019).

Table 6 Conclusions drawn to interpret an effect found profoundly significant after completion of the parametric NHST statistical procedures compared to their corresponding Bayesian equivalents (N = 10)

Statistical rationale	Null-hypothesis significance testing (NHST)				Bayesian statistics			
Psychological construct	Workload [raw NASA-TLX]			Technology acceptance (scale) [TAS]	Workload [raw NASA-TLX]			Technology acceptance (scale) [TAS]
Hypothesis	H_1: Pilot incapacitation increases the perceived workload on the remaining copilot.	H_2: The workload for pilot incapacitation during arrival is higher than during cruise.	H_3: The workload for pilot incapacitation during departure is higher than during cruise.	H_4: The technology acceptance of the MCO cockpit is higher than of the RCO' workplace (GS).	H_1: Pilot incapacitation increases the perceived workload on the remaining copilot.	H_2: The workload for pilot incapacitation during arrival is higher than during cruise.	H_3: The workload for pilot incapacitation during departure is higher than during cruise.	H_4: The technology acceptance of the MCO cockpit is higher than of the RCO' workplace (GS).
Formal statement	Normal < incapacitation	Cruise < arrival	Departure > cruise	MCO > RCO	Normal < incapacitation	Cruise < arrival	Departure > cruise	MCO > RCO
Theoretical basis	$P(D \mid H_i)$: The traditional frequentist "Fisherian" p value estimates how much of evidence is there that the true effect is different from zero. An effect is tested on significance given a hypothesis on whose base the test values are generated, not on base of the observed data.				$P(H_i \mid D)$: The parameters of the posterior distribution are estimated on the basis of the observed distribution. Default priors were used. The strength of evidence for H_i and H_0 are estimated by the BF as odds ratio of relative support for one model over another model.			
Statistical procedure	Frequentist parametric repeated-measures ANOVA			Frequentist t test	Bayesian repeated-measures ANOVA			Bayesian t test
Main effects: test statistic/Bayesian model comparison and effect analysis	Situation* $\eta_p^2 = .50$ $\eta_G^2 = .15$	Flight phase** $\eta_p^2 = .41$, $\eta_G^2 = .05$		Crewing*	Situation $BF_{10} = 36{,}672.798$ $BF_{incl.} = 30{,}451.033$	Situation + Flight phase: $BF_{10} = 132{,}832.778$ $BF_{incl.} = 0.295$ Baws factor = 0.161		Crewing
		Cruise < arrival** Cohen's $d = 1.120$	Departure > cruise** Cohen's $d = – 0.56$	MCO > RCO Cohen's $d = .746$	Baws factor = 70,521.142	Cruise < arrival $BF_{10,U} = 35.706$	Departure > cruise $BF_{10,U} = 2.213$	MCO > RCO $BF_{10} = 3.848$
Interactions: test statistic, conclusion		Crewing × Flight phase (tendency) $\eta_p^2 = .24$, $\eta_G^2 = .01$				Situation + Flight phase + Crewing × Flight phase: $BF_{10} = 20{,}434.904$ $BF_{incl.} = 0.158$; Baws factor = 0.277		
	Crewing × Situation × Flight phase** $\eta_p^2 = .58$, $\eta_G^2 = .05$ (not interpreted due to higher-order interaction)				Crewing × Situation × Flight phase ($BF_{incl.} = 0{,}145$; Baws factor = 4,709), inclusion into a model is not supported			
Significance of effect (size)	$p < .05$ Small to medium	$p < .001$ Medium	None ($p > .05$) No effect	$p < .05$ Medium	Decisive evidence in favour H_1	Decisive evidence in favour H_2	Anecdotal evidence in favour H_3	Moderate evidence for H_4
Formal statement regarding variable	Normal < incapacitation	Departure ≠ cruise ≠ arrival Cruise < arrival	Departure = cruise	MCO > RCO	Normal < incapacitation	Departure ≠ cruise ≠ arrival Cruise < arrival	Departure = cruise	MCO > RCO
Verbal statement regarding variable	The situation influences workload	Flight phase affects workload … … being higher in arrival than in cruise	… not(!) between departure and cruise	Crewing influences technology acceptance	The situation *decisively* influences workload	Workload is *decisively* higher in arrival than cruise.	Departure and cruise differ *anecdotally* in workload terms	Crewing *moderately* influences the technology acceptance
Decision	Accepted	Accepted	Rejected	Accepted	Accepted	Accepted	Rejected	Accepted
Reliability	The effect sizes depend on the model's parameters of N, effect size, type 1 error, and type 2 error. The effect size does not allow a definite conclusion and is ambiguous in interpretation.				The strength of support for a hypothesis is quantified in detail. The results are independent from sample size issues and accurate in terms of model fit and parameter estimation.			
Validity	Overestimations in the significance of effects as well as in their size are present.				The Bayes factors precisely quantify the evidence for a hypothesis.			

The non-parametric measures are omitted due to providing the same conclusions like their parametric counterparts

5.3 The Bayesian paired-sample *t* test

A Bayesian directional paired-sample *t* test using default priors[3] for the TAS' sum values of MCO and RCO was also conducted. The BF_{10} is used to describe the effect whose prior and posterior distribution are represented in Fig. 5. We found low substantial evidence for H_4 that technology acceptance is higher for MCO than for RCO (BF_{H4} = 3.848; error % = ~ 7.099e^{-5}, 95% CI [0.086; 1.313]; Mdn = 0.618).

A robustness check of the BF was performed to assess the impact of possible variations in the prior distribution on the posteriors and the BF_{H4}. Figure 6 shows the extent to which the results are affected by variations of the prior distributions. The Bayes factor is less robust to changes in the priors (max BF_{H4} = 3.948 at r = 0.5379; wide prior BF_{H4} = 3.498; ultrawide prior BF_{H4} = 2.963). Accordingly, if we had measured technology acceptance of both flight ConOps before the study and updated the priors by the state of beliefs, we had received similar results. Choosing an ultrawide prior leading ($BF_{H4ultrawide}$ = 2.964) had not altered the conclusions of a moderate to anecdotal evidence for H_4. In the end, the interpretation of a moderate evidence for H_4 remains.

5.3.1 A comparison of the Frequentist and Bayesian *t* test

Table 6 also summarizes the two different *t* tests and their results for technology acceptance sorted by statistical rationale. The frequentist NHST *t* test yielded a significant effect of technology acceptance being higher for MCO than RCO. According to this statistical rationale, we can accept H_4 without further concerns. In contrast, the Bayesian approach only found a moderate evidence for the same effect. Therefore, Bayesian reasoning benefits in application from quantifying the evidence of a particular hypothesis. As a consequence, we must not overestimate the influence of MCO leading to a higher technology acceptance when compared to RCO. The effect still remains moderate at the lower bound to anecdotal evidence when the priors would be updated by pre-knowledge on technology acceptance (Fig. 4). Hence, we assume that other factors which we have not included into the model influence the variable. This is why the Bayesian *t* test has outperformed its frequentist counterpart in estimating the strength of the effect of acceptance as well as when updating the prior data by additional evidence.

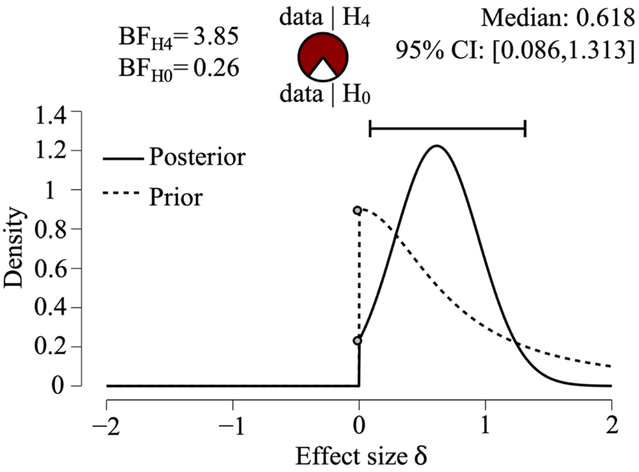

Fig. 5 The inferential plot of the prior odds and the posterior odds for the Bayesian *t* test of H_4 regarding technology acceptance would be MCO > RCO

6 Discussion

Pilot incapacitation increased the copilot's workload when compared to normal flight operations (H_1). This effect was independent from flight phase and crewing. Nonetheless, the flight phase arrival resulted in a higher workload on the copilot than cruise (H_2). When compared to departure, the flight phase cruise just revealed a slightly higher workload based on anecdotal evidence which is why we do not interpret this effect further (rejection of H_3). In contrast, the higher technology acceptance of MCO than RCO by the pilots (H_4) revealed to be of moderate evidential quality.

At first, we integrate these results into their applicational context of RCO of an airliner before we proceed with the beneficial characteristics the Bayesian approach has added to the data analysis. In general, the effects of pilot incapacitation on workload of the copilot coincide with the manufacturer's (e.g. Airbus 2011) and regulator's (ICAO 2012) classification of the event as an emergency for MCO. The event requires an immediate landing at an adjacent airport in MCO and RCO for two reasons. In both crew compositions, the loss of the second pilot as human redundancy and support during flight operations increased the workload on the remaining pilot. This finding is a consequence of more tasks simultaneously loading on the copilot without another pilot as back-up (Schmid et al. 2020; Stanton et al. 2016). Thus, the reduced-crew concept of the remote harbour pilot is based on the workload differences during the standard flight phases (Koltz et al. 2015; Schmid 2017). The *dedicated* support makes sense because workload is in fact higher during arrival while incapacitation increased workload in all flight phases.

Surprisingly, crewing and the setup of the workstation did not influence the level of perceived workload. This finding is worthy to consider further. We employed a medium-fidelity prototype as the remote-copilot's GS at early stages in the user-centred design lifecycle. In this setting, as already

[3] In the calculation of the *t* test, the default Cauchy is centred on 0 and has a scale factor called *r* that determines its width. The scale factor can equal the interquartile range as in the present example. Here, *r* = 0.707 refers to that 50% of the prior mass lies in the interval [– 0.707, 0.707]. Hence, the *Cauchy* (center = 0, *r* = 0.707) prior is used. The statistical formalities are described in detail in Gronau et al. (2020).

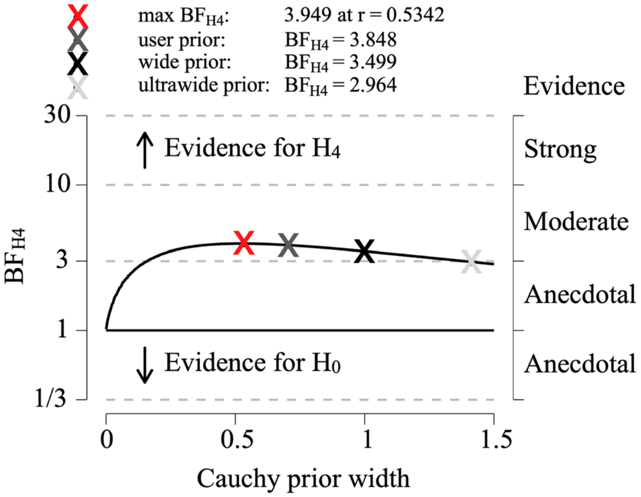

Fig. 6 The Bayes factor robustness check for the Bayesian *t* test of H_4 regarding that technology acceptance would be MCO > RCO

confirmed by previous studies, the remote-copilot can quickly enough get into the situation by being provided with the main information on aircraft systems status, flight status, and the environment (Brandt et al. 2015). Each copilot was able to land the aircraft safely at the destination airport (which was the airport of choice due to its adjacent location in the 30-min flight scenarios). The absence of an increase in workload during the incapacitation event in RCO (and in MCO) is worthy of note and should be investigated further to explore the contributing design factors. Our data suggest that it seems possible to overcome a single-pilot incapacitation by employing a ground-based *dedicated* support of a remote-copilot to take over command and control. It is unclear if the unambiguous alert to the copilot on an incapacitation by either the single-pilot, or the pilot health monitoring system, in RCO contributed to keeping the workload level comparable to normal flight or if it compensated issues with the GS' prototypical design. We cannot be sure which factors contributed to the remaining copilot's workload because the interfaces' and alerts' effects are confounded in RCO. In general, pilot incapacitation in RCO remains under-explored in terms of which factors positively affect recovering the aircraft by a remote GS (Schmid and Stanton 2019; Schmid and Stanton 2020).

The pilots' preference for MCO, in terms of technology acceptance, must not be overinterpreted because it is moderate (Wagenmakers et al. 2018b). The BF_{H4} reliably demonstrated an effect in that range even when altering the priors. Crewing partially affected their technology acceptance of MCO over protoypical RCO, but there are also other factors at play. Possible examples are age, gender, and flight experience. It is beyond the scope of the present article to investigate the TAS' two sub-scales of *usefulness* and *satisfaction* which lead to this subjective attitude toward both workstation setups. The pilots' familiarity with the setup of controls and interfaces of the standard Airbus cockpit in MCO might have caused this moderate preference.

Applying Bayesian statistics to the case study enabled us to draw more specific conclusions that are advantageous over frequentist NHST in following terms (Wasserstein and Lazar 2016). Firstly, the evidence for H_0 as well as for the alternative hypothesis H_i are quantifiable and give more information on the dimensional characteristics of an effect. It means that the Bayes factor represents a relational odds ratio that either supports one or another alternative model. The model quantifies a current state of belief under the given information. Bayesian statistics model $P(H_i|D)$ by updating the prior distribution by the observed data to estimate the posterior distribution.

In contrast, the *p* value merely describes the data in relation to a specified hypothetical explanation. As such, it does not permit a statement about this explanation. The model which is examined regarding the *p* value's level is constructed by making pre-assumptions like including the null-hypothesis. This is why $P(D|H_0)$ is tested based on data that were not observed. More precisely, the *p* value only measures how incompatible the data are with the null-hypothesis which is pre-defined by a distribution under specific assumptions. In the same, the *p* value represents a threshold to balance between two error rates. It is currently at $p = .05$ to for the type I, α-error. Hence, the distribution as well as its identified test, and related *p* value, are based on data that were never observed. Neither the size of an effect nor the importance of the result is calculated whereas only a complete report on all details can add a framework for transparency. Nonetheless, latter practices do not solve the issues related to the *p* values. Therefore, the conclusions that are drawn present questionable scientific practice (Goodman 2016).

Secondly, NHST often overestimates the significance of an effect (Hubbard and Lindsay 2008; Wagenmakers 2007). In the present study, the interaction of "crewing × flight phase" ($p = .089$; $BF_{incl} = 0.123$), the second order interaction of "crewing × situation × flight phase" ($p < .001^{**}$; $BF_{incl} = 0.145$) for workload show a tendency for significance and a high significance in NHST. By means of contrast, they yielded anecdotal evidence for H_0 close to no evidence viewed in the Bayesian approach. The significant effect of crewing on acceptance in NHST was partially overestimated as well when compared to its $BF_{H4} = 3.848$ remaining robust as low moderate evidence for H_4. This pattern of overestimating Bayesian anecdotal evidence is common when comparing frequentist NHST to the corresponding Bayesian procedures as found for 70% of a sample of 855 studies reporting *t* tests in psychology (Wetzels et al. 2011).

Thirdly, Bayesian statistics are robust against small sample sizes as demonstrated in the case study presented within this paper (Aczel et al. 2020). The strength of evidence is interpreted against the research context in which each sample size can meaningfully contribute. Thresholds are not available such as the *p* value's 0.05 because the Bayes factor represents an interpretable dimensional measure. As a consequence, Bayesian statistics are a powerful tool to synthetise research and conduct meta-analyses. The Bayes factors represent a consistent measure of evidence for different sample sizes that can be compared when considering differences in designs (Kruschke and Liddell 2018).

In these ways, the Bayesian approach currently extends the repertoire of statistical evaluation methods in HSI by providing a real and quantifiable evidence for H_0 or an H_i even for small sample size designs. An overestimation of significant results from NHST is prevented, which is a well-known issue, but to major part neglected for reasons of poor statistical education and practice in science (Matthews et al. 2017; Stark and Saltelli 2018). In doing so, the transparency in communicating and interpreting Bayesian results is superior to frequentist methods, which merely can report all complete data analyses, their effect sizes, and decisions in this procedure. This is how NHST can avoid cherry-picking piecemeal significance findings but still produce less reliable and less valid data analyses than their Bayesian equivalents, as described previously (Nuzzo 2014). Hence, Bayesian statistics improve the overall reliability and validity of results, not only in HSI, and catch up with other areas of science in methodological terms (Stanton and Young 1999).

Furthermore, the well-known *replicability problem* of empirical science is addressed in HSI as well (Kelling et al. 2017). This issue refers to results that were obtained under traditional frequentist NHST being often not replicable. In general, only between < 30 and 50% of studies in psychology, and occasionally up to 67% in prestigious journals such as *Nature* and *Science*, are replicable (Camerer et al. 2018; Lewandowsky and Oberauer 2020). This phenomenon is due to the issues related to frequentist NHST as presented here. Bayesian inference is only one suggestion on how the replicability problem could be solved by understanding why replication varies in its results (Shrout and Rodgers 2018). Thereto, the Bayesian approach updates the prior knowledge quantified as the prior distribution by empirical data to generate revised knowledge whose odds are described by the posterior distribution. In this way, the Bayesian procedure includes further emerging data into past results to assess the current evidence for a hypothesis in a replication study. Verhagen and Wagenmakers (2014) developed a methodology for a specific Bayesian replication test to compare the adequacy of 2 competing hypotheses in a replication attempt. The interested reader is referred to the literature regarding the replicability issues in psychology (Pashler and Harris 2012) and science in general (Peng 2015).

Due to all these advantages over frequentist NHST, Bayesian statistics have become a true alternative for testing hypotheses on significance to the classical frequentist methods not only for psychology during the last two decades (van de Schoot et al. 2017). Because this subject is closely related to HSI, we would appreciate the statistical evaluation of not only experimental results becoming more Bayesian here, no matter if in comparison or in use on their own as current examples including ours demonstrate (Banducci et al. 2016; Chancey et al. 2017; Karpinsky et al. 2018; Körber et al. 2018a; Körber et al. 2018b; Lee and Kolodge 2019; Neyens et al. 2015; Roth 2015; Rubin et al. 2020; Sato et al. 2019; Tear et al. 2020; Yamani and McCarley 2016; Yamani and McCarley 2018). Bayesian statistics in HSI essentially improve the reliability and validity of the statistical data analysis which is why they contribute to the overall worth of the field's research results (Stanton and Young 1999) by keeping pace with current advances in statistics (Wasserstein and Lazar 2016).

The practical implementation and applicability of Bayesian statistics has become more elaborated during the last decade and is currently promising (Kruschke 2015; Wagenmakers et al. 2018b). Several up-to-date software tools and packages are available to calculate the Bayes factor for a series of statistical procedures. We analysed the data with the well-established open-source software JASP providing a graphical and intuitive user interface, a spreadsheet and drag-and-drop layout, as well as a dynamic update of all results (JASP Team 2019; Marsman and Wagenmakers 2017). JASP offers the standard analysis procedure for both the frequentist and Bayesian methods to test hypotheses. Furthermore, it is simultaneously based on the R-package *BayesFactor* that provides the calculation of the BF for common research designs (Morey and Rouder 2018). The system JAGS implements automatic Markov chain Monte Carlo (MCMC) samplers for complex hierarchical models (Kruschke 2015; Plummer 2017). The software package Stan does the same and creates representative samples of parameter values from a posterior distribution for complex hierarchical models (Kruschke 2015; Stan Development Team 2019). IBM has recently begun to integrate seven native Bayesian procedures into its standard edition of IBM SPSS Statistics (IBM 2018a; IBM 2018b). These efforts in implementing Bayesian methods into software application have contributed to the growing popularity of the methods (van de Schoot et al. 2017).

We have identified the crucial issues of Bayesian statistics that are relevant for HSI. Since the main aim of the present work is to demonstrate the main advantages of the Bayesian statistics in an exemplary small-sample-size experimental study, we have reached our goal to raise awareness for a promising development in statistical methods of the empirical sciences (Ashby 2006; König and van de Schoot 2018; Kruschke and Liddell 2018; Lynch and Bartlett 2019). Nevertheless, HSI is slow to adopt the Bayesian approaches (Wasserstein et al. 2019). Since Bayesian procedures represent a different approach to statistical thinking, a wide range of literature on its foundations and application exists (Vandekerckhove et al. 2018). The systematic reviews from psychology (van de Schoot et al. 2017) as well as from other fields such as medicine (Ashby 2006), physics (von Toussaint

2011), and sociology (Lynch and Bartlett 2019) show this variety. It is beyond the scope of the present article to systematically assess the HSI literature on its use of Bayesian statistics.

We hope to have encouraged the reader to include Bayesian statistics into their own research. The scientific literature provides very good introductions to Bayesian foundations, reasoning, and calculation of the Bayes factor (Etz et al. 2018; Kruschke 2015; Matzke et al. 2018; Rouder et al. 2018; Wagenmakers et al. 2018a; Wagenmakers et al. 2018b) together with practicable software tools for data analysis such as JASP (Marsman and Wagenmakers 2017). Extensive work on establishing Bayesian statistics as a firm method in social and behavioural sciences has shaped the last two decades. Thereafter, we have shown how Bayesian statistics foster interpreting quantitative results unambiguously to achieve a higher reliability and validity. These benefits favour a more transparent communication of research results and their interpretation in HSI.

Funding This research received an internal institutional funding for the PhD project "Remote-Copilot-Center: Concept, Development, and Validation" from the German Aerospace Center (DLR) from 10/2015–12/2019.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Abbreviations ATC, air traffic control; ATCo, air traffic controller; ATPL, Aircraft Transport Pilot Licence; BF, Bayes factor; ConOps, concept of operations; CPL, Commercial Pilots Licence; FMS, Flight Management System; FO, first officer; GECO, generic experimental cockpit; GS, ground station; HFE, human factors and ergonomics; HSI, human systems integration; MCO, multi-crew operations; NHST, (Frequentist) null-hypothesis significance testing; PF, pilot Fflying; PM, Pilot Monitoring; RCO, reduced-crew operations; SEM, structural equation modelling; SPO, single pilot operations; TAM, Technology Acceptance Model; TAS, Technology Acceptance Scale

References

Aczel B, Hoekstra R, Gelman A, Wagenmakers EJ, Klugkist IG, Rouder JN, Vandekerckhove J, Lee MD, Morey RD, Vanpaemel W, Dienes Z, van Ravenzwaaij D (2020) Discussion points for Bayesian inference. Nat Hum Behav 4:561–563. https://doi.org/10.1038/s41562-019-0807-z

Adell E, Várhelyi A, Nilsson L (2014) The definition of acceptance and acceptability. In: Regan MA, Horberry T, Stevens A (eds) Driver acceptance of new technology: theory, measurement and optimisation. Ashgate, Farnham, pp 11–21

Airbus (2011) Airbus A380 Flight crew operating manual. Airbus S.A.S, Blagnac Cedex

American Psychological Association (2020) Publication manual of the American Psychological Association, 7th edn. American Psychological Association, Washington, D.C.

Ashby D (2006) Bayesian statistics in medicine: a 25 year review. Stat Med 25:3589–3631. https://doi.org/10.1002/sim.2672

Bakeman R (2005) Recommended effect size statistics for repeated measures designs. Behav Res Methods 37:379–384. https://doi.org/10.3758/BF03192707

Banducci SE, Ward N, Gaspar JG, Schab KR, Crowell JA, Kaczmarski H, Kramer AF (2016) The effects of cell phone and text message conversations on simulated street crossing. Hum Factors 58:150–162. https://doi.org/10.1177/0018720815609501

Boehm-Davis DA, Durso FT, Lee JD (eds) (2015) APA handbook of human systems integration. APA handbooks in psychology. American Psychological Association, Washington, D.C., USA. https://doi.org/10.1037/14528-000

Boring RL (2006) Statistical considerations for the number of participants in human factors scaling studies. Proc Hum Factors Ergon Soc Annu Meet 50:1949–1953. https://doi.org/10.1177/154193120605001750

Borsci S, Macredie RD, Martin JL, Young T (2014) How many testers are needed to assure the usability of medical devices. Expert Rev Med Dev 11:513–525. https://doi.org/10.1586/17434440.2014.940312

Brandt SL, Lachter J, Battiste V, Johnson W (2015) Pilot situation awareness and its implications for single pilot operations: analysis of a human-in-the-loop study. Procedia Manuf 3:3017–3024. https://doi.org/10.1016/j.promfg.2015.07.846

Camerer CF, Dreber A, Holzmeister F, Ho TH, Huber J, Johannesson M, Kirchler M, Nave G, Nosek BA, Pfeiffer T, Altmejd A, Buttrick N, Chan T, Chen Y, Forsell E, Gampa A, Heikensten E, Hummer L, Imai T, Isaksson S, Manfredi D, Rose J, Wagenmakers EJ, Wu H (2018) Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. Nat Hum Behav 2:637–644. https://doi.org/10.1038/s41562-018-0399-z

Chancey ET, Bliss JP, Yamani Y, Handley HAH (2017) Trust and the compliance-reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence. Hum Factors 59:333–345. https://doi.org/10.1177/0018720816682648

Cohen J (1988) Statistical power for the behavioral sciences. Erlbaum, Hillsdale

Cohen J (1994) The earth is round (p < .05). Am Psychol 49:997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Cooper H, Camic PM, Long DL, Panter AT, Rindskopf D, Sher KJ (eds) (2012) APA handbook of research methods in psychology. Research designs: Quantitative, qualitative, neuropsychological, and biological, vol 2. APA handbooks in psychology. American Psychological Association, Washington, D.C. https://doi.org/10.1037/13620-000

Davis FD (1986) A technology acceptance model for empirically testing new end-user information systems: Theory and results. Doctoral dissertation, Massachusetts Institute of Technology

Davis FD (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q 13:319–340. https://doi.org/10.2307/249008

Durso FT, Boehm-Davis DA, Lee JD (2015) A view of human systems integration from the academy. In: Boehm-Davis DA, Durso FT, Lee JD (eds) APA handbook of human systems integration. American Psychological Association, Washington, D.C., USA, pp 5–19. https://doi.org/10.1037/14528-001

Durso FT, Nickerson RS, Dumais ST, Lewandowsky S, Perfect TJ (eds) (2007) Handbook of applied cognition, 2nd edn. Wiley, Sussex

Eagle A (ed) (2011) Philosophy of probability: Contemporary readings. Routledge, New York

Ellis KKE (2009) Eye tracking metrics for workload estimation in flight deck operations. Master thesis, University of Iowa

Etz A, Gronau QF, Dablander F, Edelsbrunner PA, Baribault B (2018) How to become a Bayesian in eight easy steps: an annotated reading list. Psychon Bull Rev 25:219–234. https://doi.org/10.3758/s13423-017-1317-5

European Commission (2015) Commission Regulation (EU) No 965/2012. Off J Eur Union 55

Federal Aviation Administration (2001) Instrument flying handbook. FAA-H-8083-15 edn. U.S. Department of Transportation, Washington, D.C., USA

Funder DC, Ozer DJ (2019) Evaluating effect size in psychological research: Sense and nonsense. Adv Methods Pract Psychol Sci 2:156–168. https://doi.org/10.1177/2515245919847202

Gelman A, Goodrich B, Gabry J, Vehtari A (2019) R-squared for Bayesian regression models. Am Stat 73:307–309. https://doi.org/10.1080/00031305.2018.1549100

Goldstein M, Wooff DA (1997) Choosing sample sizes in balanced experimental designs: a Bayes linear approach. J R Stat Soc: Ser D (The Statistican) 46:167–183. https://doi.org/10.1111/1467-9884.00074

Goodman SN (2016) Aligning statistical and scientific reasoning. Sci 352:1180–1181. https://doi.org/10.1126/science.aaf5406

Gronau QF, Ly A, Wagenmakers E-J (2020) Informed Bayesian t-tests. Am Stat 74:137–143. https://doi.org/10.1080/00031305.2018.1562983

Hart SG (2006) NASA-Task Load Index (NASA-TLX): 20 years later. Proc Hum Factors Ergon Soc Annu Meet 50:904–908. https://doi.org/10.1177/154193120605000909

Hart SG, Staveland LE (1988) Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: Hancock PA, Meshkati N (eds) Human mental workload. Advances in Psychology, vol 52. North Holland, Amsterdam, pp 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Hubbard R, Lindsay RM (2008) Why p values are not a useful measure of evidence in statistical significance testing. Theor Psychol 18:69–88. https://doi.org/10.1177/0959354307086923

IBM (2018a) IBM SPSS Advanced Statistics 25. IBM, Armonk

IBM (2018b) IBM SPSS Statistics, 25.0 edn. IBM, Armonk, NY, USA

International Civil Aviation Organization (2012) Manual of civil aviation medicine. Doc 8984. Author, Montréal, Canada

International Council on Systems Engineering (2011) Systems engineering handbook – a guide for system life cycle processes and activities. Author, San Diego

Janczyk M, Xiong A, Proctor RW (2019) Stimulus-response and response-effect compatibility with touchless gestures and moving action effects. Hum Factors 61:1297–1314. https://doi.org/10.1177/0018720819831814

JASP Team (2019) JASP, 0.11.1 edn., Amsterdam, Netherlands

Jeffreys H (1961) Theory of probability, 3rd edn. Oxford University Press, Oxford

Johnson WW, Lachter J, Feary M, Comerford D, Battiste V, Mogford R (2012) Task allocation for single pilot operations: a role for the ground. In: Proceedings of the International Conference on Human-Computer Interaction in Aerospace. HCI-Aero ‘12. ACM, New York, NY, USA,

Karpinsky ND, Chancey ET, Palmer DB, Yamani Y (2018) Automation trust and attention allocation in multitasking workspace. Appl Ergon 70:194–201. https://doi.org/10.1016/j.apergo.2018.03.008

Kelling N, Ward C, Malin D, Buras W, Hetherington S (2017) The use of human factors to address medical research replicability through the development of software based solution. Proc Hum Factors Ergon Soc Annu Meet 61:597–601. https://doi.org/10.1177/1541931213601633

Koltz MT, Roberts ZS, Sweet J, Battiste H, Cunningham J, Battiste V, Vu KPL, Strybel TZ (2015) An investigation of the harbor pilot concept for single pilot operations. Procedia Manuf 3:2937–2944. https://doi.org/10.1016/j.promfg.2015.07.948

König C, van de Schoot R (2018) Bayesian statistics in educational research: a look at the current state of affairs. Educ Rev 70:486–509. https://doi.org/10.1080/00131911.2017.1350636

Körber M, Baseler E, Bengler K (2018a) Introduction matters: manipulating trust in automation and reliance in automated driving. Appl Ergon 66:18–31. https://doi.org/10.1016/j.apergo.2017.07.006

Körber M, Prasch L, Bengler K (2018b) Why do I have to drive now? Post hoc explanations of takeover requests. Hum Factors 60:305–323. https://doi.org/10.1177/0018720817747730

Kruschke JK (2015) Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan, 2nd edn. Elsevier, London

Kruschke JK, Liddell TM (2018) The Bayesian new statistics: hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. Psychon Bull Rev 25:178–206. https://doi.org/10.3758/s13423-016-1221-4

Lachter J, Brandt SL, Battiste V, Matessa M, Johnson WW (2017) Enhanced ground support: lessons from work on reduced crew operations. Cogn Tech Work 19:279–288. https://doi.org/10.1007/s10111-017-0422-6

Lee JD, Kirlik A (eds) (2013) The Oxford handbook of cognitive engineering. Oxford Library of Psychology. Oxford University Press, New York

Lee JD, Kolodge K (2019) Exploring trust in self-driving vehicles through text analysis. Hum Factors 62:1–18. https://doi.org/10.1177/0018720819872672

Lee Y, Kozar KA, Larsen KRT (2003) The technology acceptance model: past, present, and future. Commun Assoc Inf Syst 12:50. https://doi.org/10.17705/1CAIS.01250

Lenz H, Schmid D (2019) Simulation platform for reduced crew operations – a case study. Paper presented at the IEEE/AIAA 38th Digital Avionics Systems Conference (DASC), San Diego, CA, USA, September 8–12

Lewandowsky S, Oberauer K (2020) Low replicability can support robust and efficient science. Nat Commun 11:358. https://doi.org/10.1038/s41467-019-14203-0

Lewis JR (1994) Sample sizes for usability studies: additional considerations. Hum Factors 36:368–378. https://doi.org/10.1177/001872089403600215

Lewis JR, Sauro J (2006) When 100% really isn't 100%: improving the accuracy of small-sample estimates of completion rates. J Usability Stud 1:136–150

LimeSurvey GmbH (2003) LimeSurvey, 3.17.15 edn. LimeSurvey GmbH, Hamburg, Germany

Lindley DV (1993) The analysis of experimental data: The appreciation of tea and wine. Teach Stat 15:22–25. https://doi.org/10.1111/j.1467-9639.1993.tb00252.x

Lindley DV (2000) The philosophy of statistics. J R Stat Soc: Ser D (The Statistican) 49:293–337. https://doi.org/10.1111/1467-9884.00238

Lynch SM, Bartlett B (2019) Bayesian statistics in sociology: past, present, and future. Annu Rev Sociol 45:47–68. https://doi.org/10.1146/annurev-soc-073018-022457

Marsman M, Wagenmakers E-J (2017) Bayesian benefits with JASP. Eur J Dev Psychol 14:545–555. https://doi.org/10.1080/17405629.2016.1259614

Masson MEJ (2011) A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. Behav Res Methods 43:679–690. https://doi.org/10.3758/s13428-010-0049-5

Mathôt S (2017) Bayes like a baws: Interpreting Bayesian repeated measures in JASP. https://www.cogsci.nl/blog/interpreting-bayesian-repeated-measures-in-jasp. Accessed 24 June 2020

Matthews R, Wasserstein R, Spiegelhalter D (2017) The ASA's p-value statement, one year on. Significance 14:38–41. https://doi.org/10.1111/j.1740-9713.2017.01021.x

Matzke D, Boehm U, Vandekerckhove J (2018) Bayesian inference for psychology, part III: Parameter estimation in nonstandard models. Psychon Bull Rev 25:77–101. https://doi.org/10.3758/s13423-017-1394-5

Miranda AT (2018) Understanding human error in naval aviation mishaps. Hum Factors 60:763–777. https://doi.org/10.1177/0018720818771904

Morey RD, Rouder JN (2018) Package ‘BayesFactor’:computation of Bayes factors for common designs, 0.9.12-4.2 edn. CRAN,

Neyens DM, Boyle LN, Schultheis MT (2015) The effects of driver distraction for individuals with traumatic brain injuries. Hum Factors 57:1472–1488. https://doi.org/10.1177/0018720815594057

Nuzzo R (2014) Scientific method: Statistical errors. Nature 506:150–152

Olejnik S, Algina J (2003) Generalized eta and omega squared statistics: measures of effect size for some common research designs. Psychol Methods 8:434–447. https://doi.org/10.1037/1082-989X.8.4.434

Pashler H, Harris CR (2012) Is the replicability crisis overblown? Three arguments examined. Perspect Psychol Sci 7:531–536. https://doi.org/10.1177/1745691612463401

Peng R (2015) The reproducibility crisis in science: a statistical counterattack. Significance 12:30–32. https://doi.org/10.1111/j.1740-9713.2015.00827.x

Plummer M (2017) JAGS, 4.3.0 edn.,

Rahman MM, Lesch MF, Horrey WJ, Strawderman L (2017) Assessing the utility of TAM, TPB, and UTAUT for advanced driver assistance systems. Accid Anal Prev 108:361–373. https://doi.org/10.1016/j.aap.2017.09.011

Regan MA, Horberry T, Stevens A (2014) Driver acceptance of new technology: theory, measurement and optimisation. Human factors in road and rail transport. Ashgate, Farnham

Regens JL, Mould N, Jensen CJ, Graves MA, Edger DN (2015) Probabilistic graphical modeling of terrorism threat recognition using Bayesian networks and monte carlo simulation. J Cogn Eng Decis Mak 9:29–5311. https://doi.org/10.1177/1555343415592730

Rödel C, Stadler S, Meschtscherjakov A, Tscheligi M (2014) Towards autonomous cars: The effect of autonomy levels on acceptance and user experience. Paper presented at the Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Seattle, WA, USA,

Roth W-M (2015) Cultural practices and cognition in debriefing: The case of aviation. J Cogn Eng Decis Mak 9:263–278. https://doi.org/10.1177/1555343415591395

Rouder JN, Engelhardt CR, McCabe S, Morey RD (2016) Model comparison in ANOVA. Psychon Bull Rev 23:1779–1786. https://doi.org/10.3758/s13423-016-1026-5

Rouder JN, Haaf JM, Vandekerckhove J (2018) Bayesian inference for psychology, part IV: Parameter estimation and Bayes factors. Psychon Bull Rev 25:102–113. https://doi.org/10.3758/s13423-017-1420-7

Rubin M, Giacomini A, Allen R, Turner R, Kelly B (2020) Identifying safety culture and safety climate variables that predict reported risk-taking among Australian coal miners: an exploratory longitudinal study. Saf Sci 123:104564. https://doi.org/10.1016/j.ssci.2019.104564

Salvendy G (ed) (2012) Handbook of human factors and ergonomics, 4th edn. Wiley, Hoboken. https://doi.org/10.1002/9781118131350

Sato T, Yamani Y, Liechty M, Chancey ET (2019) Automation trust increases under high-workload multitasking scenarios involving risk. Cogn Tech Work 22:399–407. https://doi.org/10.1007/s10111-019-00580-5

Schmid D (2017) A workload-centered perspective on reduced crew operations in commercial aviation. Paper presented at the H-Workload 2017: The first international symposium on human mental workload, Dublin, Ireland, June 28-30

Schmid D, Korn B, Stanton NA (2020) Evaluating the reduced flight deck crew concept using cognitive work analysis and social network analysis: Comparing normal and data-link outage scenarios. Cogn Tech Work 22:109–124. https://doi.org/10.1007/s10111-019-00548-5

Schmid D, Stanton NA (2018) How are laser attacks encountered in commercial aviation? A hazard analysis based on systems theory. Saf Sci 110:178–191. https://doi.org/10.1016/j.ssci.2018.08.012

Schmid D, Stanton NA (2019) A future airliner’s reduced-crew: modelling pilot incapacitation and homicide-suicide with systems theory. Hum-Intell Syst Integr 1:27–42. https://doi.org/10.1007/s42454-019-00001-y

Schmid D, Stanton NA (2020) Progressing toward airliners’ reduced-crew operations: A systematic literature review. Int J Aerosp Psychol 30:1–24. https://doi.org/10.1080/24721840.2019.1696196

Schmid D, Vollrath M, Stanton NA (2018) The System Theoretic Accident Modelling and Process (STAMP) of medical pilot knock-out events: pilot incapacitation and homicide-suicide. Saf Sci 110:58–71. https://doi.org/10.1016/j.ssci.2018.07.015

Senior AM, Grueber CE, Kamiya T, Lagisz M, O’Dwyer K, Santos ESA, Nakagawa S (2016) Heterogeneity in ecological and evolutionary meta-analyses: its magnitude and implications. Ecol 97:3293–3299. https://doi.org/10.1002/ecy.1591

Shrout PE, Rodgers JL (2018) Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. Annu Rev Psychol 69:487–510. https://doi.org/10.1146/annurev-psych-122216-011845

Smid SC, McNeish D, Miočević M, van de Schoot R (2019) Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. Struct Equ Model Multidiscip J 27:1–31. https://doi.org/10.1080/10705511.2019.1577140

Stan Development Team (2019) Stan, 2.21 edn. Stan Development Team,

Stanton NA, Harris D, Starr A (2016) The future flight deck: Modelling dual, single and distributed crewing options. Appl Ergon 53:331–342. https://doi.org/10.1016/j.apergo.2015.06.019

Stanton NA, Hedge A, Brookhuis K, Salas E, Hendrick H (eds) (2005) Handbook of human factors and ergonomics methods. CRC Press, Boca Raton

Stanton NA, Salmon PM, Rafferty LA, Walker GH, Baber C, Jenkins DP (2013) Human factors methods: a practical guide for engineering and design, 2nd edn. Ashgate, Farnham

Stanton NA, Young MS (1999) What price ergonomics. Nature 399:197–198. https://doi.org/10.1038/20298

Stark PB, Saltelli A (2018) Cargo-cult statistics and scientific crisis. Significance 15:40–43. https://doi.org/10.1111/j.1740-9713.2018.01174.x

Tear MJ, Reader TW, Shorrock S, Kirwan B (2020) Safety culture and power: Interactions between perceptions of safety culture, organisational hierarchy, and national culture. Saf Sci 121:550–561. https://doi.org/10.1016/j.ssci.2018.10.014

van de Schoot R, Winter SD, Ryan O, Zondervan-Zwijnenburg M, Depaoli S (2017) A systematic review of Bayesian articles in psychology: The last 25 years. Psychol Methods 22:217–239. https://doi.org/10.1037/met0000100

van den Berg M et al. (2019) A tutorial on conducting and interpreting a Bayesian ANOVA in JASP. OSF Storage (Germany – Frankfurt). doi:10.31234/osf.io/spreb

Van der Laan JD, Heino A, De Waard D (1997) A simple procedure for the assessment of acceptance of advanced transport telematics. Transp Res Part C Emerg Technol 5:1–10. https://doi.org/10.1016/S0968-090X(96)00025-3

van Doorn BA et al. (2019) The JASP guidelines for conducting and reporting a Bayesian analysis. PsyArXiv. doi:10.31234/osf.io/yqxfr

Vandekerckhove J, Rouder JN, Kruschke JK (2018) Editorial: Bayesian methods for advancing psychological science. Psychon Bull Rev 25:1–4. https://doi.org/10.3758/s13423-018-1443-8

Verhagen J, Wagenmakers E-J (2014) Bayesian tests to quantify the result of a replication attempt. J Exp Psychol Gen 143:1457–1475. https://doi.org/10.1037/a0036731 von Toussaint U (2011) Bayesian inference in physics. Rev Mod Phys 83:943–999. https://doi.org/10.1103/RevModPhys.83.943

Vu K-PL, Lachter J, Battiste V, Strybel T (2018) Single pilot operations in domestic commercial aviation. Hum Factors 60:755–762. https://doi.org/10.1177/0018720818791372

Wagenmakers E-J (2007) A practical solution to the pervasive problems of p values. Psychon Bull Rev 14:779–804. https://doi.org/10.3758/bf03194105

Wagenmakers E-J, Love J, Marsman M, Jamil T, Ly A, Verhagen J, Selker R, Gronau QF, Dropmann D, Boutin B, Meerhoff F, Knight P, Raj A, van Kesteren EJ, van Doorn J, Šmíra M, Epskamp S, Etz A, Matzke D, de Jong T, van den Bergh D, Sarafoglou A, Steingroever H, Derks K, Rouder JN, Morey RD (2018a) Bayesian inference for psychology. Part II: Example applications with JASP. Psychon Bull Rev 25:58–76. https://doi.org/10.3758/s13423-017-1323-7

Wagenmakers E-J, Marsman M, Jamil T, Ly A, Verhagen J, Love J, Selker R, Gronau QF, Šmíra M, Epskamp S, Matzke D, Rouder JN, Morey RD (2018b) Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. Psychon Bull Rev 25:35–57. https://doi.org/10.3758/s13423-017-1343-3

Wagenmakers E-J, Wetzels R, Borsboom D, van der Maas HLJ, Kievit RA (2012) An agenda for purely confirmatory research. Perspect Psychol Sci 7:632–638. https://doi.org/10.1177/1745691612463078

Wasserstein RL, Lazar NA (2016) The ASA statement on p-values: Context, process, and purpose. Am Stat 70:129–133. https://doi.org/10.1080/00031305.2016.1154108

Wasserstein RL, Schirm AL, Lazar NA (2019) Moving to a world beyond "p < 0.05". Am Stat 73:1–19. https://doi.org/10.1080/00031305.2019.1583913

Westfall PH, Johnson WO, Utts JM (1997) A Bayesian perspective on the Bonferroni adjustment. Biometrika 84:419–427. https://doi.org/10.1093/biomet/84.2.419

Wetzels R, Matzke D, Lee MD, Rouder JN, Iverson GJ, Wagenmakers E-J (2011) Statistical evidence in experimental psychology: an empirical comparison using 855 t tests. Perspect Psychol Sci 6:291–298. https://doi.org/10.1177/1745691611406923

Wetzels R, Wagenmakers E-J (2012) A default Bayesian hypothesis test for correlations and partial correlations. Psychon Bull Rev 19:1057–1064. https://doi.org/10.3758/s13423-012-0295-x

Wickens CD, Hollands JG, Banbury S, Parasuraman R (2018) Engineering psychology and human performance, 4th edn. Pearson, London

Yamani Y, McCarley JS (2016) Workload capacity: a response time-based measure of automation dependence. Hum Factors 58:462–471. https://doi.org/10.1177/0018720815621172

Yamani Y, McCarley JS (2018) Effects of task difficulty and display format on automation usage strategy: a workload capacity analysis. Hum Factors 60:527–537. https://doi.org/10.1177/0018720818759356

Young MS, Brookhuis KA, Wickens CD, Hancock PA (2015) State of science: Mental workload in ergonomics. Ergonomics 58:1–17. https://doi.org/10.1080/00140139.2014.956151

Zhou F, Ji Y, Jiao RJ (2014) Prospect-theoretic modeling of customer affective-cognitive decisions under uncertainty for user experience design. IEEE Trans Hum-Mach Syst 44:468–483. https://doi.org/10.1109/THMS.2014.2318704

Zondervan-Zwijnenburg M, Peeters M, Depaoli S, Van de Schoot R (2017) Where do priors come from? Applying guidelines to construct informative priors in small sample research. Res Hum Dev 14:305–320. https://doi.org/10.1080/15427609.2017.1370966

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.