

Article

Perceived Workload in Mixed IFR/VFR Terminal Operations: An Exploratory Pilot Simulator Study in Zagreb TMA

Bruno Antulov-Fantulin , Karolina Krajček Nikolić , Petar Andrašić  and Antonio Živko

Faculty of Transport and Traffic Sciences, University of Zagreb, Vukelićeva 4, 10000 Zagreb, Croatia; kkrajcek@fpz.unizg.hr (K.K.N.); pandrasi@fpz.unizg.hr (P.A.); antonio.zivko0@gmail.com (A.Ž.)

* Correspondence: bantulov@fpz.unizg.hr

Abstract

Visual flight rules (VFR) traffic is routinely excluded from capacity models and workload assessments in terminal airspace, despite its potential to increase air traffic controller (ATCO) workload. This exploratory pilot study examined the association between a mixed instrument flight rules (IFR) and VFR terminal scenario and perceived approach-control workload in the simulated Zagreb Terminal Manoeuvring Area (TMA). Nine ATCO trainees completed three within-subject simulator exercises in a fixed order: a mixed IFR/VFR scenario (EXP1), an operationally comparable IFR-only baseline (EXP2), and a higher-density IFR-only scenario (EXP3). Workload was measured using the NASA Task Load Index (NASA-TLX). The mixed scenario was associated with a 9.26-point higher mean weighted NASA-TLX score than the IFR-only baseline (Cohen's $d_z = 0.67$; 95% CI -1.33 to 19.85 ; two-sided $p = 0.079$); an exploratory directional analysis, specified before data collection, yielded one-sided $p = 0.039$. The omnibus analysis was not significant ($F(2, 16) = 2.56$, $p = 0.109$). Frustration and Effort were the only subscales with Holm-adjusted p -values below 0.05; these results are interpreted as exploratory. The two IFR-only scenarios showed little observed difference. The fixed scenario order, small trainee sample, and scripted degraded communication preclude causal inference, motivating a larger counterbalanced confirmatory study with operational controllers.

Keywords: air traffic control; air traffic complexity; ATCO workload; NASA-TLX; terminal manoeuvring area; mixed IFR/VFR operations; human factors; simulator study; exploratory study

1. Introduction

The global air transport industry faces a critical challenge: balancing sustained traffic growth with the cognitive limitations of human operators. According to Eurocontrol's Statistics and Forecast Service (STATFOR), air traffic in Europe is predicted to reach 12.4 million flights in 2031, an increase of approximately 12% over the record traffic of 2019 [1].

Due to such continuous increase in air traffic volume, air navigation service providers (ANSPs) operate sectors increasingly close to their theoretical capacities. While technological advancements and procedural optimisation provide some relief, the Air Traffic Controller (ATCO) remains the central decision-maker. Consequently, ATCO workload has emerged as the primary functional constraint on airspace capacity [2].

This constraint is particularly acute in the Terminal Manoeuvring Area (TMA), where airspace is constricted and operations are dynamic. The under-quantified variable in TMA complexity is the presence of mixed-mode operations, specifically, the interaction between



Academic Editor: Michael Schultz

Received: 9 June 2026

Revised: 11 July 2026

Accepted: 16 July 2026

Published: 19 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the

[Creative Commons Attribution \(CC BY\) license](https://creativecommons.org/licenses/by/4.0/).

traffic operating under Instrument Flight Rules (IFR) and Visual Flight Rules (VFR). While IFR traffic is predictable and centrally managed, VFR traffic introduces higher degrees of uncertainty, requiring additional coordination and communication loads that standard capacity models may fail to capture.

Current capacity models, such as the CAPAN (Capacity Analysis) methodology, widely used across European ANSPs, estimate workload by summing average task execution times for categories of IFR controller tasks [3]. These models do not incorporate VFR-specific interactions, meaning that declared sector capacities may underestimate actual workload during periods of mixed IFR/VFR operations, a condition common in European TMAs during summer months and at aerodromes with active general aviation.

To the best of the authors' knowledge, and based on a targeted review of the available literature, no published study has directly compared perceived workload in mixed IFR/VFR terminal operations against an operationally matched IFR-only baseline using NASA-TLX data. Existing ATC workload research has focused predominantly on IFR-only scenarios, traffic density manipulations, or automation effects [4–7]. While Kopardekar and Magyarits [5] demonstrated that traffic interactions, rather than density alone, drive workload, and Histon and Hansman [8] theorised that VFR aircraft disrupt the structural regularities on which controllers rely, neither study provided empirical NASA-TLX data comparing mixed IFR/VFR operations against an IFR-only baseline in a terminal approach context. Recent years have also produced a substantial body of work on data-driven sector complexity metrics and controller workload prediction [9,10]. This work includes graph-based [11–14], deep-learning [15,16], and machine-learning approaches [17,18]. However, these metrics are derived almost exclusively from IFR traffic data and do not model flight-rule heterogeneity (see Section 2). This gap is the primary motivation for the present work.

This paper examines perceived approach-control workload during mixed IFR/VFR operations within the Zagreb TMA using the NASA Task Load Index (NASA-TLX) [19]. The aim is to design scenarios within a high-fidelity simulator that approximate operationally comparable mixed IFR/VFR and IFR-only conditions, while explicitly acknowledging that the present fixed-order pilot design cannot isolate causal effects. The specific contributions of this work are:

1. Design of controlled Zagreb TMA simulator scenarios intended to compare mixed IFR/VFR and IFR-only terminal operations;
2. A quantitative assessment of perceived workload using the NASA-TLX, analysing both weighted overall workload and subscale-level patterns;
3. Identification of operationally relevant hypotheses for future capacity and workload assessment in mixed-mode terminal environments.

We test the following hypotheses, distinguishing between the primary and secondary analyses. Because the study employs a fixed scenario order and a small convenience sample without preregistration, both hypotheses are treated as directional predictions to be evaluated within a pilot-study framework rather than as confirmatory tests in the strict inferential sense:

H1 (Primary): The mixed IFR/VFR scenario, taken as a complete operational package (flight-rule heterogeneity together with its inherent communication, coordination, and performance-mix demands), is associated with higher perceived ATCO workload than the operationally comparable IFR-only baseline ($EXP1 > EXP2$), as measured by the overall weighted NASA-TLX score. The same directional expectation is extended, exploratorily, to the comparison of the mixed scenario with the higher-density IFR-only scenario ($EXP1 > EXP3$), since mixed-mode features rather than aircraft count are hypothesised to drive workload.

H2 (Secondary): Any workload increase is expected to be most apparent in the Effort and Frustration dimensions of the NASA-TLX scale.

The remainder of this paper is organised as follows: Section 2 reviews complexity and workload concepts with a focus on VFR-specific factors; Section 3 describes the study design, scenarios, measures, and statistical approach; Section 4 presents results; Section 5 discusses findings, limitations and implications; Section 6 concludes.

2. Background: Air Traffic Complexity and ATCO Workload

To define safe limits of airspace capacity, one must understand the relationship between the objective flight environment (complexity) and the subjective impact on the controller (workload).

2.1. Complexity, Workload, and Their Relationship

Air traffic complexity refers not merely to the number of aircraft in a sector (density) but to the dynamic interactions among them [20,21]. As Kirwan et al. [21] demonstrated, identical traffic counts can produce vastly different interaction patterns: a structured flow generates fewer interactions than a randomised flow [4]. Notably, safety occurrences often do not correlate linearly with peak traffic density and incidents frequently occur during periods of normal or low traffic volume [21].

ATCO workload, the total cognitive and physical effort required for task execution, is generally accepted as a function of sector complexity, moderated by factors such as equipment quality, individual experience, and the cognitive strategy employed [4,22]. The primary cognitive tasks include maintaining situational awareness, planning and revising traffic sequences, resolving conflicts, and managing communications [23,24]. The relationship between complexity and workload is not linear: two controllers managing identical traffic may experience different workload levels depending on their experience and strategies [4].

Research on sector complexity metrics has moved from expert-weighted linear indicators toward data-driven and learning-based models; a recent bibliometric review maps this growing body of work [9]. Graph-based approaches describe the traffic situation as a network of aircraft interdependencies: spatiotemporal graph indicators [11], graph representation learning on flight conflict networks [12], dynamical graph learning that predicts controller-reported workload [13], and attention-based graph neural networks that yield interpretable per-aircraft task-demand scores [14]. Deep-learning approaches evaluate complexity from image representations of traffic [15] or with ordered deep metrics [16], while machine-learning studies identify the most influential complexity parameters [17] and predict a dynamic complexity indicator at tactical time horizons [18]. Related simulator studies quantified how individual traffic flows shape ATCO actions [25] and conflict detection [26], and the Performance Review Unit complexity framework has been extended with a probabilistic treatment of prediction uncertainty [27].

Work on workload measurement likewise emphasises the central role of the human operator: a systematic review of seventy workload-prediction studies groups the influential factors into task complexity, airspace structure, operational limitations, and cognitive load [10]; mental-workload measures have been shown to predict performance decrements and complexity peaks better than established task factors [28]; and behavioural markers extracted from ATCO voice recordings [29], together with speech-based deep-learning assessment [30], provide unobtrusive proxies of task demand. Notably, these recent metrics are computed almost exclusively from IFR surveillance and flight-plan data: flight-rule heterogeneity is not represented as an explicit complexity factor, so the contribution of VFR traffic to sector complexity and workload remains largely unquantified.

2.2. Benchmarking Workload: The CAPAN Methodology

To manage workload operationally, ANSPs utilise the CAPAN methodology [3], which estimates workload by summing average execution times of observable task categories: flight data management, conflict search, coordination, R/T communication and radar tasks.

While CAPAN provides a necessary baseline, it relies on average task times for standardised IFR operations and does not capture the cognitive overhead associated with monitoring uncertain, non-standard elements such as VFR traffic. This limitation motivates the use of subjective measures, specifically the NASA-TLX [19], to provide a more complete picture of controller experience in mixed-mode operations. In the CAPAN methodology, ATCO workload is calculated by summing individual task durations (measured in seconds) and expressing the total as a percentage of working time per hour [3]. As shown in Table 1, a sector is classified as being in a state of Overload if the workload reaches or exceeds 70% (equivalent to 42 min of task execution per hour). This 70% limit is formally defined as the Heavy Load Threshold (HLT), representing the maximum sustainable workload beyond which safety and performance may be compromised. While the HLT serves as the primary operational guideline for capacity declarations, it is based on average task execution times and is therefore treated as a flexible threshold rather than an absolute limit.

Table 1. CAPAN methodology workload thresholds. Data from Ref. [3].

Threshold	Interpretation	Working Time/Hour
≥70%	Overload	42+ min
54–69%	Heavy	32–41 min
30–53%	Medium	18–31 min
18–29%	Light	11–17 min
0–17%	Very light	0–10 min

2.3. The Impact of Mixed-Mode Operations (IFR/VFR)

The literature has long emphasised that traffic heterogeneity is a critical complexity factor. Kopardekar and Magyarits [5] demonstrated that interactions between aircraft, not density per se, drive controller workload.

Histon and Hansman [8] proposed that ATCOs rely on airspace structure (standard routes, predictable speeds, uniform procedures, etc.) to reduce cognitive load. VFR aircraft disrupt these regularities by operating outside standard IFR trajectories, at lower and more variable speeds, and with non-standard communication, forcing a shift from pattern-based to state-based monitoring.

Netjasov [31] identifies aircraft type mix as a key variable in conflict risk assessment. In terminal environments, the speed differential between typical VFR training aircraft (~100 kt) and IFR arrivals (180–200 kt) creates sequencing challenges. The operational community has recently re-emphasised this gap. Working papers presented at the 2024 IFATCA conference argue that VFR flights are not subject to ATFM slotting and flow control and therefore add workload that current capacity figures do not capture [32]. They also propose that capacity determination criteria should explicitly include the flight rules applied within the airspace [33]. Despite these considerations, empirical data on VFR-specific workload contributions in controlled TMA environments remain scarce. In addition, as noted in Section 2.1, the recent generation of data-driven complexity metrics has not yet incorporated flight-rule heterogeneity as an input feature.

Manning et al. [34] identified R/T communication time as a reliable workload proxy in ATC operations. VFR traffic generates non-standard communication demands, including FIC coordination, special clearances, and the provision of traffic information required under Class C airspace rules, none of which are captured by CAPAN's IFR-centric task taxonomy.

This convergence of theoretical predictions and methodological gaps forms the basis for the present investigation.

3. Materials and Methods

3.1. Study Design and Procedure

The study employed a human-in-the-loop (HITL) real-time simulation methodology, in which ATCO trainees actively controlled simulated traffic in real time while interacting with a pseudopilot operating the surrounding aircraft. The study used a within-subject repeated-measures design. Each participant completed two preparatory exercises (PRP1–PRP2) followed by three research exercises (EXP1–EXP3) in the simulated Zagreb TMA.

Sessions were conducted over two days. On Day 1, participants completed PRP1 and PRP2 under the supervision of an instructor who is a Zagreb approach controller holding a valid instructor endorsement. Prior to Day 1, all participants viewed a 40-minute recorded demonstration exercise. The instructor provided individualised pre-briefing and debriefing for each preparatory exercise.

On Day 2 (approximately one week later), participants completed EXP1 → EXP2 → EXP3. The instructor was present but assumed a passive role: before each exercise, participants received a brief operational briefing comprising the number of aircraft and general traffic composition (mixed IFR/VFR or IFR-only), accompanied by general motivational remarks (e.g., encouraging focused performance). No tactical guidance, real-time intervention, corrective feedback, or instructional debriefing was provided during the research exercises, in contrast to the individualised instruction and debriefing provided during PRP1–PRP2. This change in the level of instructional support represents an additional potential confound: EXP1 was the first exercise conducted under this reduced-support regime, and any elevated workload may partly reflect the withdrawal of active guidance rather than the VFR manipulation alone. Although the instructor's continued physical presence likely attenuated acute performance anxiety, the qualitative shift in support level cannot be fully separated from the experimental condition in the present fixed-order design. Immediately after each exercise, participants completed the NASA-TLX. Breaks of 15–20 min were provided between exercises. The complete experimental protocol is summarised in Table 2, an overview of the purpose and content of each exercise is given in Table 3, and the full sequence of familiarisation, preparatory, and research phases is illustrated in Figure 1. After the final research exercise (EXP3), and only after completion of the corresponding NASA-TLX, participants were invited to provide brief spontaneous impressions of the session (Section 4.6). The discussion was non-directive and did not include tactical feedback, corrective instruction, or performance evaluation; because it took place after all exercises and questionnaires had been completed, it could not have influenced task performance or the workload ratings.

All participants completed the research exercises in a fixed order (EXP1 → EXP2 → EXP3) due to scheduling constraints. This introduces a potential confound between the experimental manipulation and sequence position that cannot be empirically resolved with the present data. A learning effect, whereby familiarity with the simulator and airspace model accumulates across exercises, would reduce workload in later scenarios (EXP2, EXP3) and thereby bias the design against H1. Conversely, fatigue and anxiety associated with the first exercise conducted under passive instructor support could inflate EXP1 workload irrespective of VFR content, biasing the design in favour of H1. Because the VFR manipulation and the sequence position are fully confounded, neither interpretation can be empirically verified. The two-day familiarisation protocol (PRP1 and PRP2 on Day 1 with instructor supervision) was designed to attenuate initial unfamiliarity, but cannot eliminate

all sequence-related effects. A Latin-square counterbalanced design is therefore a first-order priority for the planned confirmatory study (see Section 5).

Table 2. Experimental protocol timeline.

Phase	Day	Activity	Duration
Familiarisation	Pre-Day 1	Recorded demonstration video	~40 min
Preparation	Day 1	PRP1 + instructor debrief	~45 min
		Break	15–30 min
		PRP2 + instructor debrief	~45 min
Experiment	Day 2	Briefing + EXP1 (IFR+VFR) + TLX	~40 min
		Break	15–20 min
		Briefing + EXP2 (IFR-only) + TLX	~40 min
		Break	15–20 min
		Briefing + EXP3 (IFR + 2) + TLX	~40 min
		Informal post-exercise observations	~5 min

Table 3. Overview of preparatory and research exercises.

Exercise	Purpose	Traffic Composition	Aircraft (n)	Instructor support	NASA-TLX
PRP1	Familiarisation with simulator and airspace	9 IFR + 2 VFR	11	Active (briefing and debriefing)	No
PRP2	Familiarisation, increased demand	10 IFR + 2 VFR + 1 Z	13	Active (briefing and debriefing)	No
EXP1	Mixed-mode research scenario	9 IFR + 2 VFR + 1 Z (6 planned conflicts)	12	Passive	Yes
EXP2	IFR-only baseline	12 IFR (6 planned conflicts)	12	Passive	Yes
EXP3	IFR-only density control	14 IFR (6 planned conflicts)	14	Passive	Yes

Note: Z denotes a Zulu (mixed-rule) flight plan. Detailed flight plans are provided in Appendix A.

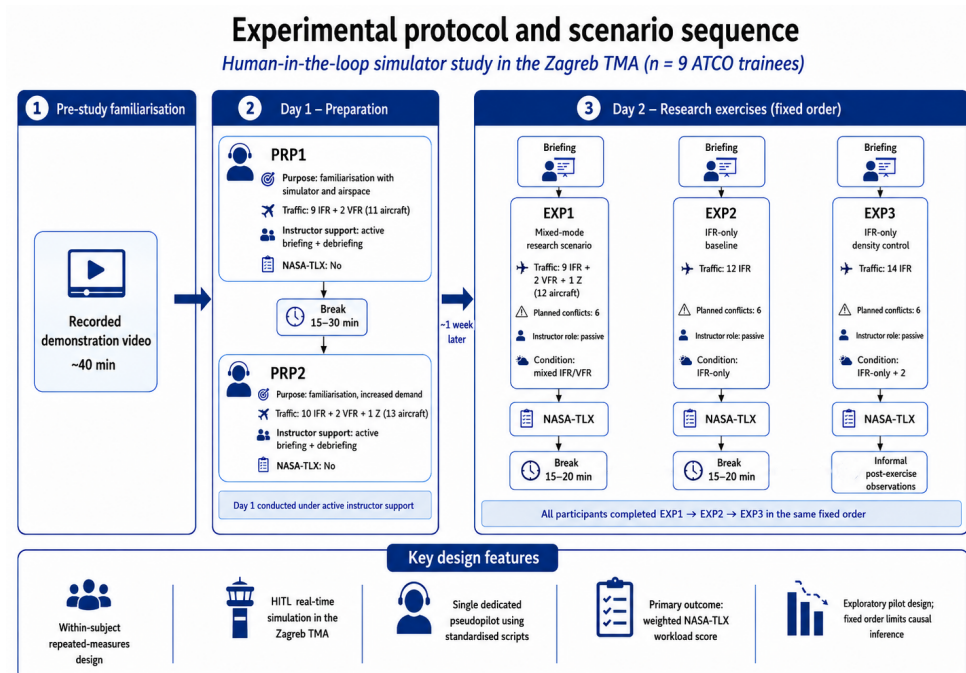


Figure 1. Flow diagram of the experimental protocol and scenario sequence. Instructor support was active (individualised briefing and debriefing) during the Day 1 preparatory exercises and passive during the Day 2 research exercises, which all participants completed in the same fixed order; the NASA-TLX was administered immediately after each research exercise.

3.2. Participants

Participants were nine air traffic controller (ATCO) trainees ($n = 9$; 4 male; 5 female; aged 21–25 years) recruited from the Croatian Air Traffic Control Education Centre (Hrvatsko učilišno središte za kontrolu zračnog prometa–HUSK). HUSK is the primary training organisation operating within the Department of Aeronautics at the Faculty of Transport and Traffic Sciences, University of Zagreb. All participants had completed the ATCO Basic Training curriculum at the Faculty and possessed approximately 30 h of previous simulator experience on the MicroNav BEST platform. Participation was voluntary, and summary demographic characteristics are presented in Table 4.

Table 4. Participant demographic characteristics ($n = 9$).

Characteristic	Value
Sample size	9
Sex (male/female)	4/5
Age range (years)	21–25
Training stage	Completed Basic ATCO Training
Simulator experience	≈30 h (MicroNav BEST platform)
Operational rating	None (pre-rating trainees)

Note: All participants were recruited from a single institution (University of Zagreb Faculty of Transport and Traffic Sciences). A formal demographic questionnaire was not administered. Values are based on institutional training records and self-reported age.

Although the sample size is small ($n = 9$), comparable sample sizes have been used in exploratory ATC simulator research. For example, Oprins et al. [35] employed $n = 10$ trainees in a competence-based assessment study, and Metzger and Parasuraman [7] used $n = 8$ controllers in a decision-aid evaluation. The present sample should therefore be interpreted as appropriate for an exploratory pilot study aimed at estimating the direction and approximate magnitude of the effect, rather than for confirmatory inference or broad generalisation. The within-subject repeated-measures design reduces between-participant variability because each participant serves as their own control. However, it does not remove the limitations associated with small-sample uncertainty, trainee participants, or the fixed scenario order.

3.3. Simulator Environment

The research was conducted within the Laboratory for Air Traffic Control at the Department of Aeronautics, Faculty of Transport and Traffic Sciences, University of Zagreb [36]. The laboratory houses the Croatian Air Traffic Control Education Centre (HUSK), certified as an ATCO Training Organisation (certificate HR.ATCOTO.01) by the Croatian Civil Aviation Agency (CCAA) in accordance with Commission Regulation (EU) 2015/340 [37]. The Laboratory is equipped with the licensed MicroNav BEST radar ATC simulator, a high-fidelity platform used for professional air traffic controller training and competency assessments. The HUSK facility provides two controller work positions (CWPs) and two pseudopilot workstations. The positions, as shown in Figure 2, have the corresponding equipment necessary for the successful performance of ATCO tasks, such as a touch display with an interface for radio communication, peripherals (mice, keyboards, and headphones), an additional screen for flight progress strips control and pertaining maps of the Croatian airspace. Only one exercise was conducted at a time for consistency. A single dedicated pseudopilot operated all aircraft throughout all exercises for all participants, following a standardised script for each scenario. For one non-IFR aircraft in EXP1, the Zulu-flight 9AZAO departing from Varaždin, the script specified degraded communication behaviour (hesitant readbacks, non-standard phraseology, delayed responses) calibrated to repli-

cate real-world VFR training flight characteristics frequently observed in Zagreb TMA operations. The remaining two VFR aircraft (9ADMB and 9ADDD) communicated using standard phraseology. The same pseudopilot behaviour was applied consistently across all nine participants.

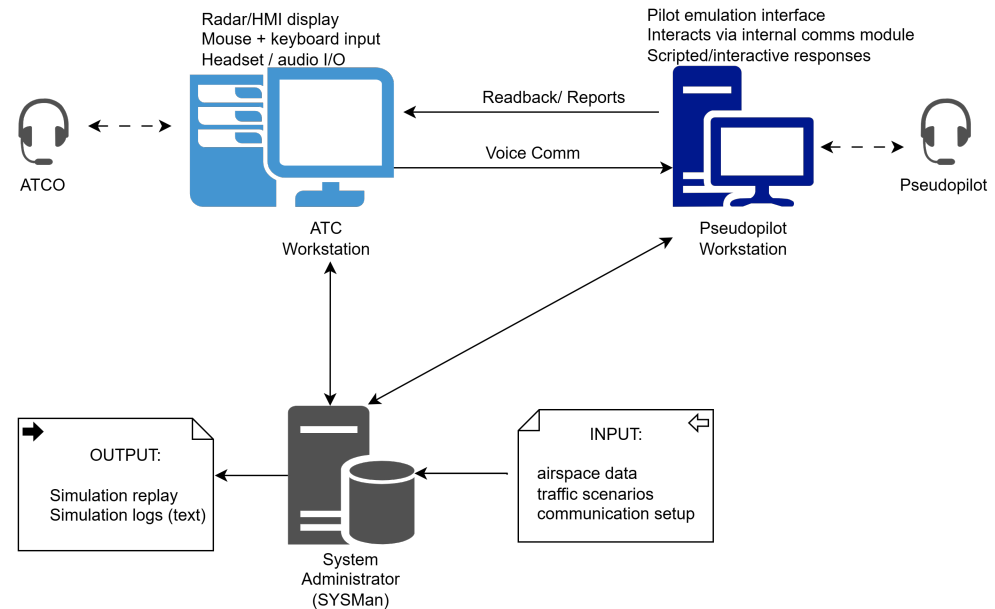


Figure 2. Simulation data collection pipeline in the HUSK/MicroNav BEST environment. A single ATC controller working position (CWP) interacted with a single pseudopilot workstation through the internal communications module, while the System Manager (SysMan) executed the scenario configuration and maintained the traffic state distributed to the radar/HMI.

3.4. Airspace Model (Zagreb TMA)

Due to the division of airspace within Croatia, area controllers rarely encounter VFR traffic. The study, therefore, focused on the approach-control environment.

3.4.1. Airspace Structure

The simulated airspace comprised the Zagreb TMA (Class C, 1000 ft AGL to 20,500 ft) and two control zones: CTR Zagreb (LDZA) and CTR Lučko (LDZL), both Class D from ground to 2500 ft AGL. Within this structure, IFR traffic received ATC separation, while VFR traffic received the ATS appropriate to the airspace class, including traffic information and, where applicable, separation relative to IFR [38].

Departures from Varaždin Airport (LDVA), modelled without ATC, required FIC coordination [39].

3.4.2. Waypoints and Navigation Fixes

The simulation incorporated published waypoints (ARGOM, ROLBA, MAGAM, PODET, PETOV, OBUTI, KOPRY, RASIN, TEBLI, RUDIK, NIVES, KOTOR), navigation-aid fixes (PIS, SK, ZAG, VBA), and VFR visual reference points (N1, N2, A1, S1, S2, E2, W1) per the Croatian AIP [39].

Figure 3 illustrates the simulated Zagreb TMA airspace, including the CTRs and aerodromes, Varaždin Airport, minimum vectoring altitude (MRVA), lateral airspace boundaries, and all waypoints used in the scenario design.

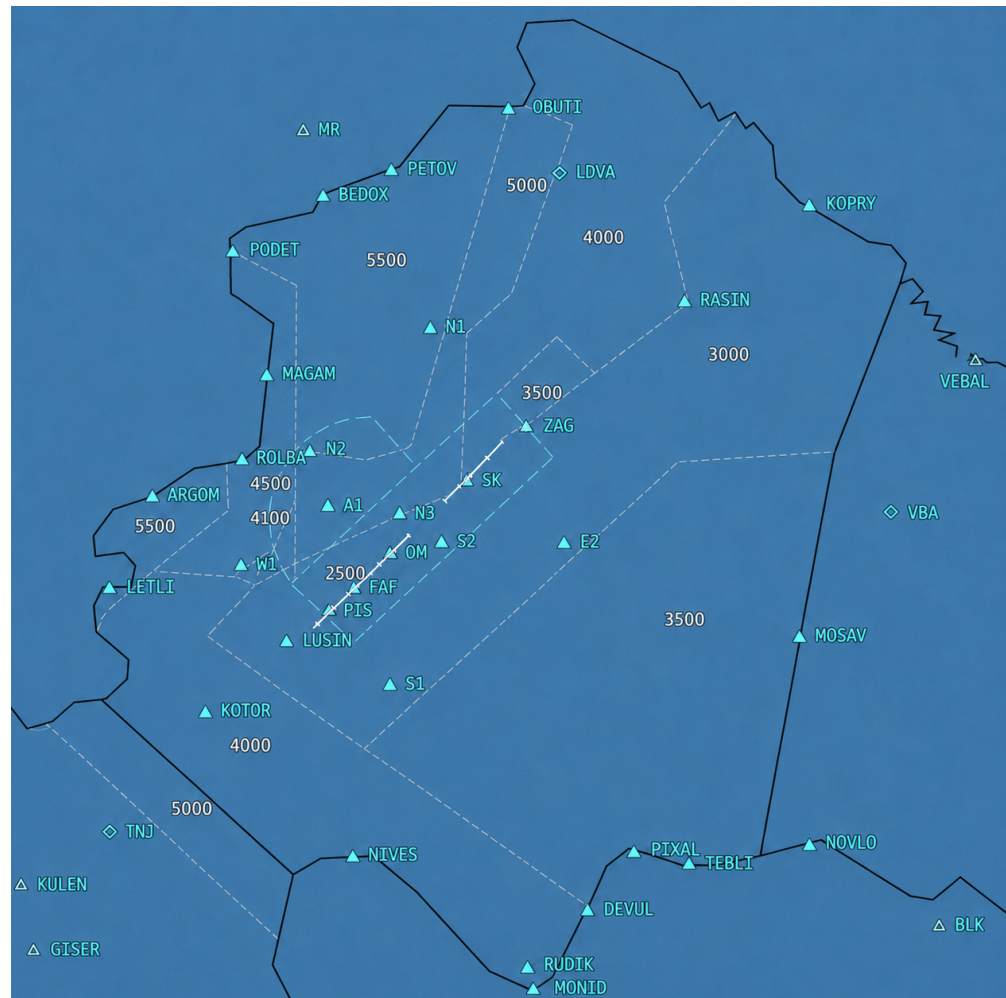


Figure 3. Simulated airspace of Zagreb TMA.

3.5. Scenario Design

Three research exercises were designed to examine perceived workload under mixed IFR/VFR and IFR-only terminal operations. EXP1 represented mixed IFR/VFR operations typical of terminal environments, including VFR training flights and a mixed-rule (Z) flight. EXP2 was an IFR-only scenario designed to match the task intent of EXP1 by replacing VFR-related tasks with IFR aircraft generating comparable sequencing, coordination, and communication demands. EXP3 increased IFR density by adding two IFR aircraft to EXP2 while keeping the operational mode homogeneous (IFR-only).

Two preparatory exercises, PRP1 and PRP2, familiarised participants with the environment using the instructional protocol detailed in Section 3.1. Detailed flight plans for all exercises are provided in Appendix A (Tables A1–A3).

3.5.1. Scenarios EXP1–EXP3 (Design Intent)

EXP1 comprised twelve aircraft in total: nine operating under IFR, two under VFR, and one under the Zulu flight rule. A Zulu (Z) flight plan denotes an aircraft departing under VFR that transitions to IFR en route, requiring the controller to manage the rule change and associated coordination [39]. For terminological consistency, 9AZAO is referred to throughout as a Zulu (mixed-rule) flight; the term VFR is reserved for the two aircraft that operated under visual flight rules for the entire flight (9ADMB and 9ADDD), and the three aircraft are collectively described as non-IFR traffic. Three aircraft ($\approx 25\%$) operated outside conventional IFR, two pure VFR and one Zulu, all C172s belonging to ICAO Wake

Turbulence Category (WTC) Light. These aircraft maintained 100 kt in level flight and 80 kt during climb and descent.

Each VFR aircraft served a specific design purpose within the scenario. Aircraft 9ADMB, departing from Lučko aerodrome (LDZL), requested two practice approaches to Zagreb Airport (LDZA), illustrating the sequencing difficulty arising from the speed differential between VFR and IFR traffic—IFR aircraft in the exercise maintained approach speeds of 180–200 kt, representing a difference of over 100 kt. Aircraft 9ADDD, also departing from LDZL, requested a climb to 7000 ft for parachute jump operations, creating a conflict with northbound traffic inbound to LDZA. Aircraft 9AZAO operated on a Zulu flight plan from Varaždin (LDVA) to Osijek (LDOS), departing VFR and requesting the change to IFR upon initial contact with Zagreb approach, shortly after leaving the LDVA traffic circuit in the vicinity of waypoint RASIN. To replicate real-world training characteristics, this flight was implemented with the degraded communication quality described in Section 3.3. The remaining two VFR aircraft (9ADMB and 9ADDD) communicated using standard phraseology. The degraded communication of flight 9AZAO introduced a specific variable of pilot unreliability that is operationally plausible in VFR training traffic but analytically important. Consequently, the present design cannot separate the effects of flight-rule heterogeneity from the effects of degraded communication quality. The resulting workload pattern is therefore interpreted as an exploratory mixed-mode scenario effect rather than as a pure VFR-presence effect. This point deserves emphasis because the degraded communication constitutes a second manipulated variable embedded within EXP1. The script was applied identically to all nine participants, so it acted as a constant scenario feature rather than a source of between-participant variability. EXP1 nevertheless differed from EXP2 in two respects at once, namely flight-rule composition and communication quality, and the two contributions cannot be statistically separated in the present data. The implications of this confound are examined further in Section 5.

The remaining 75% of traffic comprised IFR flights: five arrivals and three departures at LDZA, plus one IFR overflight (THY34), which was required to descend from FL210 to FL140 within TMA Zagreb. The IFR fleet included four Airbus A320s, two Boeing 737-800s, two Boeing 747-400s, and one De Havilland Canada Dash 8.

EXP2 retained the same total aircraft count as EXP1 (twelve), replacing all VFR and Zulu traffic with IFR equivalents while preserving the original task structure. Each substitution was designed to replicate the specific operational demand of the aircraft it replaced. All other aircraft remained unchanged from EXP1. EXP2 comprised five IFR departures, six IFR arrivals, and one IFR overflight.

EXP3 was identical to EXP2 with the addition of two aircraft to increase the overall workload. EXP3 thus comprised fourteen aircraft: five IFR departures, seven IFR arrivals, and two IFR overflights. All exercises ran to completion (i.e., until the last aircraft was transferred to an adjacent frequency or to Zagreb Tower) rather than to a fixed time limit. Exercise duration therefore varied across participants depending on the expeditiousness of their control actions. Mean exercise durations were comparable across conditions (see Section 4.1), confirming that observed workload differences are not attributable to differences in time-on-task.

3.5.2. Scenario Comparability Matrix

EXP1 and EXP2 were designed to be operationally comparable in terms of overall task structure, including arrival sequencing demands and planned conflict count. Communication load, by contrast, was not matched, since the additional radio-telephony activity associated with VFR operations was an intended part of the experimental manipulation. The resulting comparability matrix is presented in Table 5. The operational equivalence

is grounded in three design principles: (a) each VFR aircraft in EXP1 was replaced by an IFR aircraft in EXP2 that generated a structurally analogous task demand (e.g., a practice approach requiring sequencing, a departure requiring conflict management); (b) the total number of planned conflicts, defined as scripted trajectory pairs that would result in a separation infringement below 5 NM horizontally or 1000 ft vertically absent controller intervention, was held constant at six; and (c) the overall traffic count was matched at twelve aircraft. Note that while the total count is identical, the distribution across flight phases is not perfectly matched (EXP1: 5 IFR + 2 VFR arrivals vs. EXP2: 6 IFR arrivals), and arrival sequencing demand is therefore approximately rather than exactly equivalent. The critical difference in EXP1 is the introduction of flight-rule heterogeneity and associated sequencing demands. As noted in Section 3.1, the fixed scenario order remains a limitation for causal inference.

Table 5. Scenario comparability matrix for experimental exercises (EXP1–EXP3).

Feature	EXP1 (IFR/VFR)	EXP2 (IFR-Only)	EXP3 (IFR + 2)
Total aircraft	12	12	14
IFR aircraft	9	12	14
VFR/Zulu	3	0	0
WTC L/M/H	3/7/2	0/10/2	0/12/2
Arrivals	5 (+2 VFR)	6	7
Departures	3	5	5
Overflights/transits	1 + 1 transit	1	2
Planned conflicts	6	6 (IFR equiv.)	6
FIC coordination	Yes	No	No
Speed diff. > 100 kt	Yes	No	No

Note: WTC categories: C172 = L; A320, B738, DHC8C, AT72 = M; B744 = H.

Because these components covary by construction, the estimand of the primary comparison is the effect of the complete mixed-mode scenario package relative to the IFR-only baseline, rather than the isolated effect of VFR presence or of flight-rule heterogeneity alone. Table 6 decomposes the EXP1–EXP2 contrast into its constituent components and distinguishes those that are an intentional part of the mixed-mode operational construct from those that are uncontrolled confounds.

Table 6. Components of the EXP1–EXP2 contrast and their status with respect to the mixed-mode estimand.

Component (EXP1 Relative to EXP2)	Intentional?	Status
Flight-rule heterogeneity (2 VFR + 1 Zulu vs. none)	Yes	Core of the mixed-mode construct
Aircraft performance mix (three WTC-L C172)	Yes	Inherent to the construct
Speed differential > 100 kt in arrival sequencing	Yes	Inherent to the construct
FIC coordination requirement	Yes	Inherent to the construct
Additional planned scripted R/T activity (88 vs. 69 transmissions)	Yes	Inherent to the construct
Degraded communication on the Zulu flight 9AZAO	Yes	Embedded second manipulation; confound with respect to VFR presence
Sequence position (EXP1 always first)	No	Uncontrolled confound
First exercise without active instructor support	No	Uncontrolled confound

3.6. Workload Measure: NASA-TLX

Subjective workload was assessed using the NASA-TLX [19], one of the most extensively validated subjective workload instruments in the human factors literature. As Hart and Staveland [19] established, the NASA-TLX captures the operator's integrated perception of task demands, an aspect of workload that objective performance measures and task-time analyses cannot access directly. Objective metrics such as communication

duration or conflict detection latency measure outcomes of cognitive processing, whereas the TLX measures the subjective cost of that processing to the operator. In contexts where controllers can maintain adequate performance under very different internal loads (a phenomenon well-documented in ATC research [6]), subjective measurement provides information that objective indicators may miss. The NASA-TLX was therefore selected as the primary workload measure for this study.

The instrument measures perceived workload across six dimensions: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. Participants provided raw ratings (0–100, 5-point increments) and pairwise importance judgments (15 comparisons). The weighted overall score equals the sum of each rating multiplied by its weight (tally/15). The Performance scale is anchored from *Good* (0) to *Poor* (100), so that higher marked values indicate poorer perceived performance. The marked value therefore enters the computation directly, without arithmetic transformation, in accordance with the original protocol [19]. The instrument was administered in paper-based format with 5-point increment scales. The NASA-TLX was administered immediately after each exercise, within approximately one minute of exercise termination, to minimise recall degradation. Participants were familiarised with the instrument and its rating procedure before their first use. Average completion time was under five minutes per administration.

3.7. Statistical Analysis

The primary outcome measure was the overall weighted NASA-TLX score. Because the same participants completed all scenarios, analyses used within-subject paired tests. The NASA-TLX weighted overall score was computed as:

$$TLX_{\text{weighted}} = \sum_{i=1}^6 w_i \cdot r_i \quad (1)$$

where r_i denotes the rating for dimension i as marked on the questionnaire (0–100, 5-point increments) and w_i denotes its corresponding weight derived from pairwise comparisons ($w_i = \text{tally}_i/15$, $\sum w_i = 1$). No arithmetic reversal is applied to the Performance subscale, consistent with the scale anchoring described in the previous subsection.

In addition to subjective workload, three scenario-level task-load metrics were monitored to describe scenario comparability: exercise duration, the planned number of conflicts, and the average number of radio-telephony (R/T) message transmissions per exercise. Because R/T transmissions were generated from a fixed pseudopilot script, this metric is interpreted as a scenario-design indicator of communication load rather than as an independent behavioural outcome.

The primary comparison was EXP1 vs. EXP2 (test of H1). Paired differences were summarised with 95% CIs and Cohen's d_z following standard experimental design principles [40].

The paired-samples t -statistic was computed as:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \quad (2)$$

where $\bar{d}$ is the mean of within-participant differences (EXP1 – EXP2), s_d is the standard deviation of those differences, and n is the number of participants. Within-subject effect size was quantified using Cohen's d_z :

$$d_z = \frac{\bar{d}}{s_d} \quad (3)$$

Effect sizes were interpreted according to conventional benchmarks: small ($d_z = 0.20$), medium ($d_z = 0.50$), and large ($d_z = 0.80$).

Given the three-scenario structure of the design, a one-way repeated-measures analysis of variance (ANOVA) across EXP1, EXP2, and EXP3 was conducted as an omnibus test. Sphericity was assessed with Mauchly's test, with the Greenhouse–Geisser correction to be applied where indicated, and the Friedman test was computed as a non-parametric counterpart. Planned pairwise comparisons followed the omnibus analysis. The contrast EXP1 vs. EXP2 was the primary prespecified comparison, while EXP1 vs. EXP3 and EXP3 vs. EXP2 served as secondary contrasts. In line with the exploratory framing, effect sizes with confidence intervals are emphasised over dichotomous significance decisions. A leave-one-out sensitivity analysis of the primary contrast was additionally performed to assess the influence of individual participants. Normality of paired differences was assessed using the Shapiro–Wilk test. With $n = 9$, this test has low power to detect departures from normality, so a non-significant result does not establish that the distribution is normal. It indicates only that the data are not inconsistent with normality. Paired-samples t -tests were therefore applied together with Wilcoxon signed-rank tests reported as sensitivity analyses. Zero-valued paired differences were excluded from the Wilcoxon test, tied absolute differences were assigned mid-ranks, and exact one-sided p -values were computed from the conditional permutation distribution. For the directional H1, one-sided p -values are reported. The directional form of H1 and H2 was specified during scenario design, before any workload data were collected or analysed, and no interim analyses preceded this choice. Because the study was not prospectively registered, however, primary evidential weight is placed on the two-sided p -values and the 95% confidence intervals of the effect estimates, and all one-sided results are reported as exploratory directional evidence rather than as confirmatory findings. Because effect-size estimates from small samples are upwardly biased, the bias-corrected Hedges' g_z is reported alongside Cohen's d_z ($g_z = d_z \times J$, where $J = 1 - 3/(4(n - 1) - 1) = 0.903$). Subscale comparisons (secondary test of H2) used Holm–Bonferroni correction for multiplicity across six dimensions. Because H2 involves multiple simultaneous tests, it is interpreted as secondary. Subscales that remained significant after Holm–Bonferroni correction are identified as candidates for confirmatory testing in future work.

Given the small sample, two supplementary analyses were conducted to characterise inferential uncertainty beyond frequentist p -values: a Bayesian paired-samples t -test using the default JZS prior (Cauchy scale $r = \sqrt{2}/2$. Rouder et al. [41]), reported in both its two-sided (BF_{10}) and directional one-sided (BF_{+0}) variants, and a 95% bootstrap confidence interval for Cohen's d_z (10,000 resamples, percentile method). Given the exploratory nature of this pilot study, the resulting Bayes Factor is treated not as a formal test of the null hypothesis, but as a supplementary indicator of inferential uncertainty given the small sample size.

4. Results

4.1. Descriptive Statistics

Before conducting the primary analysis, Shapiro–Wilk tests were performed to assess the normality of paired differences. The results did not indicate strong departures from normality for the three pairwise comparisons: EXP1–EXP2 ($W = 0.97$, $p = 0.88$), EXP1–EXP3 ($W = 0.94$, $p = 0.62$), and EXP2–EXP3 ($W = 0.88$, $p = 0.17$).

The structural comparability of the scenarios was characterised using objective task-load metrics (Table 7). Each exercise (EXP1–EXP3) was designed with a fixed count of six planned conflicts and maintained a comparable time-on-task across conditions. Radio-communication demand was characterised using the planned number of scripted trans-

missions in each scenario: 88 in EXP1, 69 in EXP2, and 80 in EXP3, representing 27.5% more planned scripted transmissions in EXP1 than in EXP2. A transmission was defined as one discrete radio utterance by a single speaker; each controller utterance and each pilot or pseudopilot response were counted separately, so that a controller instruction and the corresponding readback constituted two transmissions. All planned radio and coordination communications were included, irrespective of content or purpose, and repeated calls, requests for repetition, and corrective exchanges were counted as additional transmissions when specified as separate utterances in the scenario script. The resulting counts represent the expected communication demand built into each scenario rather than empirically observed transmission totals. Because the pseudopilot followed a fixed script, these counts reflect scenario design rather than participant behaviour and are reported as descriptive context.

Table 7. Objective task-load metrics across experimental scenarios.

Metric	EXP1 (IFR/VFR)	EXP2 (IFR-Only)	EXP3 (IFR + 2)
Total Aircraft	12	12	14
Exercise Duration <i>M</i> (SD), min	39 (1.5)	39 (0.6)	40 (1)
Planned Conflicts	6	6	6
Planned scripted R/T transmissions (<i>n</i>)	88	69	80

Transmission counts are planned scripted totals; each discrete speaker turn was counted separately (an ATCO instruction and the corresponding readback constituted two transmissions). The counts characterise expected scenario communication demand and do not represent observed participant-level communication activity.

Mean overall NASA-TLX workload was 55.33 (SD = 16.21) in EXP1, 46.07 (SD = 16.50) in EXP2, and 46.70 (SD = 18.62) in EXP3. Table 8 shows the average results of the NASA-TLX Questionnaire for exercises EXP1, EXP2, and EXP3. Raw (unweighted) TLX scores, computed as the simple mean of the six dimensions as marked on the questionnaire, yielded a consistent direction but a notably stronger effect: EXP1 ($M = 52.41$, $SD = 16.86$), EXP2 ($M = 41.57$, $SD = 14.80$), EXP3 ($M = 40.56$, $SD = 16.75$). The raw TLX EXP1 vs. EXP2 difference was 10.83 points ($t(8) = 3.06$, one-sided $p = 0.008$, $d_z = 1.02$, Hedges' $g_z = 0.92$), compared with $d_z = 0.67$ for the weighted score. The mean differences under the two scoring methods are similar (10.83 vs. 9.26 points). The difference in effect size arises primarily because the individually varying pairwise-comparison weights add between-participant variability to the weighted difference scores (SD of paired differences: 13.7 weighted vs. 10.61 raw), inflating the denominator of d_z . A secondary contribution is that participants assigned a below-average weight to Frustration in EXP1 (mean weight = 0.148, vs. $1/6 \approx 0.167$ under equal weighting) despite giving it high ratings ($M = 60.56$), so the weighting step also attenuates the subscale showing the largest between-condition difference. Similar considerations have motivated some authors to report raw TLX as the primary outcome in recent NASA-TLX applications [42]. The weighted score was defined a priori as the primary outcome, for consistency with the original Hart and Staveland [19] protocol; the raw score is reported strictly as a sensitivity analysis. Because the two scores answer slightly different measurement questions and the larger raw-score effect partly reflects the lower between-participant variability of the raw difference scores, the discrepancy between them is not interpreted as evidence in favour of the larger estimate. Individual weighted TLX scores for each participant are presented in Table 9.

Table 8. Average NASA-TLX ratings and weights ($n = 9$).

Dimension	EXP1		EXP2		EXP3	
	Rating	Weight	Rating	Weight	Rating	Weight
Mental Demand	62.22	0.259	53.89	0.289	53.33	0.259
Physical Demand	37.78	0.007	28.33	0.030	25.00	0.022
Temporal Demand	52.78	0.193	43.33	0.178	46.11	0.230
Performance	36.11	0.215	35.00	0.237	39.44	0.252
Effort	65.00	0.178	50.00	0.185	51.67	0.156
Frustration	60.56	0.148	38.89	0.081	27.78	0.081
Overall Mean (SD)	55.33 (16.21)		46.07 (16.50)		46.70 (18.62)	

Note: Ratings range from 0–100; the Performance scale is anchored Good (0)–Poor (100) and is reported as marked. Weights are normalised (0–1) based on pairwise comparisons.

Table 9. Individual weighted NASA-TLX scores per participant and scenario.

Participant	EXP1	EXP2	EXP3
1	71.33	54.67	55.67
2	79.67	70.33	56.00
3	45.67	62.33	58.67
4	66.33	64.00	66.33
5	49.00	35.00	30.00
6	39.33	35.00	26.33
7	55.33	30.33	57.67
8	62.33	34.00	57.00
9	29.00	29.00	12.67
M (SD)	55.33 (16.21)	46.07 (16.50)	46.70 (18.62)

4.2. Omnibus Analysis Across Scenarios

The one-way repeated-measures ANOVA across the three research scenarios did not reach statistical significance ($F(2, 16) = 2.56$, $p = 0.109$, partial $\eta^2 = 0.24$, generalised $\eta^2 = 0.06$). Mauchly's test indicated no violation of sphericity ($W = 0.92$, $p = 0.75$), so no correction was required. The non-parametric Friedman test was consistent with this result ($\chi^2(2) = 4.41$, $p = 0.110$, Kendall's $W = 0.245$). The planned pairwise contrasts reported below were prespecified and are presented in accordance with the exploratory character of the study. The non-significant omnibus result reinforces the pilot-study framing of this work. The findings that follow should therefore be read as estimates of the direction and approximate magnitude of the scenario effect rather than as confirmatory evidence.

4.3. Primary Comparison: EXP1 vs. EXP2

The mean within-participant difference (EXP1 – EXP2) was 9.26 points (SD = 13.78), with a moderate effect size (Cohen's $d_z = 0.67$; Hedges' $g_z = 0.61$). The paired-samples t -test yielded a two-sided $p = 0.079$ and a directional one-sided $p = 0.039$ for the prespecified prediction EXP1 > EXP2. The Wilcoxon signed-rank sensitivity analysis was consistent in direction ($p = 0.043$, one-sided). Seven of nine participants exhibited higher workload in the mixed scenario (EXP1) than in the IFR-only baseline, one participant showed the reverse pattern, and one showed no difference. A leave-one-out sensitivity analysis confirmed that the direction of the effect did not depend on any single participant. Across the nine reduced samples, d_z ranged from 0.55 to 1.20, and the one-sided p -value ranged from 0.006 to 0.083. Individual trajectories across the three scenarios are shown in Figure 4.

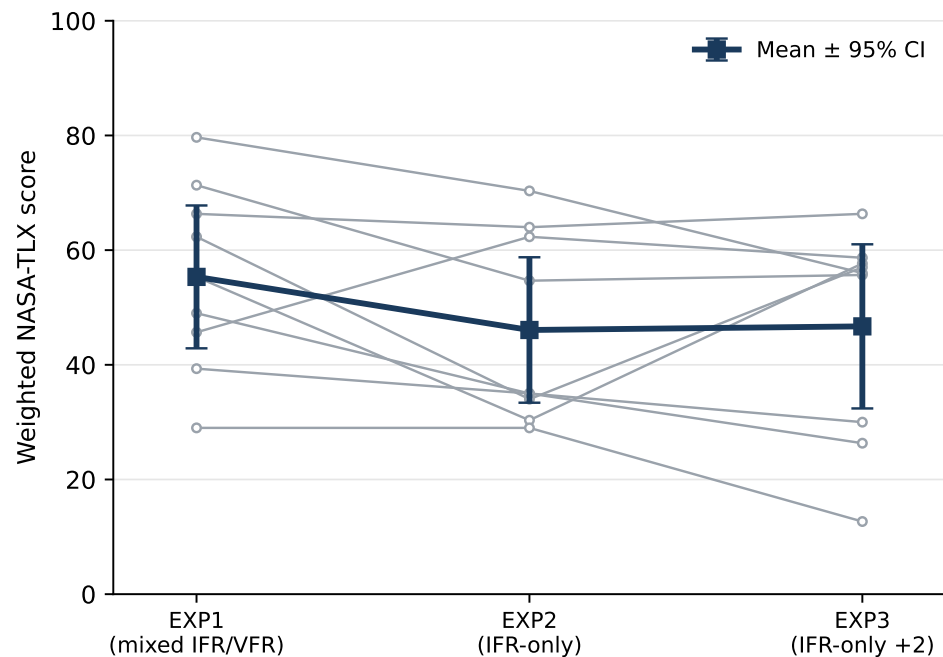


Figure 4. Individual weighted NASA-TLX scores across the three research scenarios ($n = 9$). Grey lines denote individual participants and the dark line denotes the sample mean with 95% confidence intervals.

Two supplementary analyses characterised inferential uncertainty. A Bayesian paired-samples t -test with the default JZS prior (Cauchy scale $r = 0.707$) yielded $BF_{10} = 1.31$ for the two-sided hypothesis and $BF_{+0} = 2.47$ for the prespecified directional hypothesis. On the Jeffreys scale, both values fall within the anecdotal range (1–3). The data therefore provide no more than weak evidence for H1 and do not permit a decisive discrimination between H1 and H0. This inferential uncertainty is consistent with the small sample size. A 95% bootstrap confidence interval for d_z (10,000 resamples, percentile method) ranged from 0.04 to 1.91, spanning negligible to very large effects. If the observed point estimate ($d_z = 0.67$) reflects the true population effect, approximately 20 participants would be required for 80% power with a two-sided test at $\alpha = 0.05$ (24 if the bias-corrected $g_z = 0.61$ is used). However, the wide bootstrap interval means the true required sample size could be considerably larger. This calculation is illustrative rather than a definitive sample-size recommendation, because it conditions on a point estimate obtained from $n = 9$. For planning, a smallest effect of operational interest should be specified in advance; for reference, $d_z = 0.40$, 0.50, and 0.67 would require approximately 52, 34, and 20 participants, respectively (80% power, two-sided $\alpha = 0.05$).

Full pairwise comparisons and individual differences are presented in Table 10 and Figure 5.

Table 10. Paired statistical comparisons: overall NASA-TLX.

Comparison	Mean Difference [95% CI]	Cohen's d_z	p (Two-Sided)
EXP1 – EXP2	9.26 [−1.33, 19.85]	0.67	0.079
EXP1 – EXP3	8.63 [−0.53, 17.79]	0.72	0.062
EXP2 – EXP3	−0.63 [−12.37, 11.11]	−0.04	0.904

Exploratory directional one-sided p -values, prespecified for the contrasts involving EXP1: EXP1 – EXP2, $p = 0.039$; EXP1 – EXP3, $p = 0.031$; no directional hypothesis was specified for EXP2 vs. EXP3. SDs of the paired differences were 13.78, 11.91, and 15.27, with $t(8) = 2.02$, 2.17, and -0.12 , respectively. The 95% CIs are two-sided and should be used to interpret the uncertainty of the effect estimates.

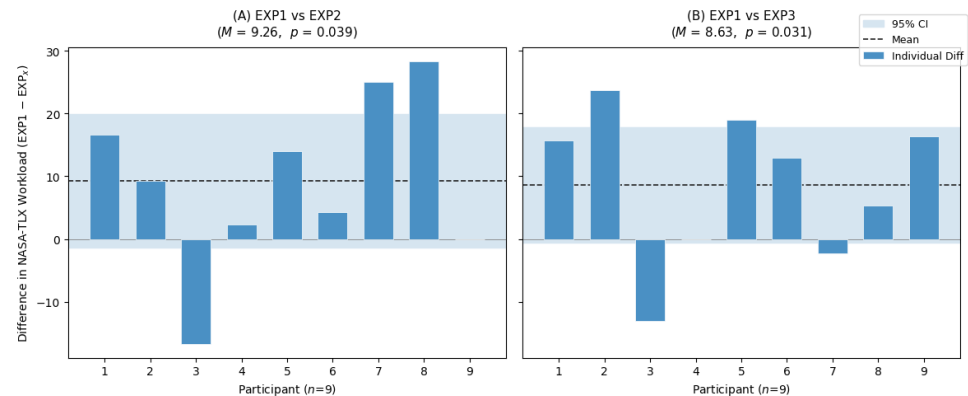


Figure 5. Graph of individual differences in NASA-TLX workload. **(A)** EXP1 vs EXP2 ($M = 9.3$; two-sided $p = 0.079$, directional one-sided $p = 0.039$). **(B)** EXP1 vs EXP3 ($M = 8.6$; two-sided $p = 0.062$, directional one-sided $p = 0.031$). Positive values indicate higher workload in EXP1. Dashed line: mean; shaded area: 95% CI. $n = 9$.

4.4. Secondary Comparisons

The planned contrast between the mixed scenario and the higher-density IFR-only scenario (EXP1 vs. EXP3) showed a pattern similar to the primary comparison. The mean difference was 8.63 points ($t(8) = 2.17$, two-sided $p = 0.062$, one-sided $p = 0.031$, $d_z = 0.72$), and the Wilcoxon signed-rank test was consistent in direction (one-sided $p = 0.031$). The corresponding directional Bayes factor was $BF_{+0} = 2.99$, which again falls within the anecdotal range. EXP2 and EXP3 showed little observed difference in weighted NASA-TLX scores (mean difference = -0.63 , 95% CI $[-12.37, 11.11]$, $d_z = -0.04$, two-sided $p = 0.904$). Because both scenarios occurred later in the fixed sequence, this contrast is reported descriptively and is not used to explain the elevated EXP1 workload. The corresponding Bayes factor ($BF_{10} = 0.32$, i.e., $BF_{01} = 3.1$) constitutes moderate Bayesian evidence for little difference between EXP2 and EXP3 under the specific fixed sequence used in this pilot study.

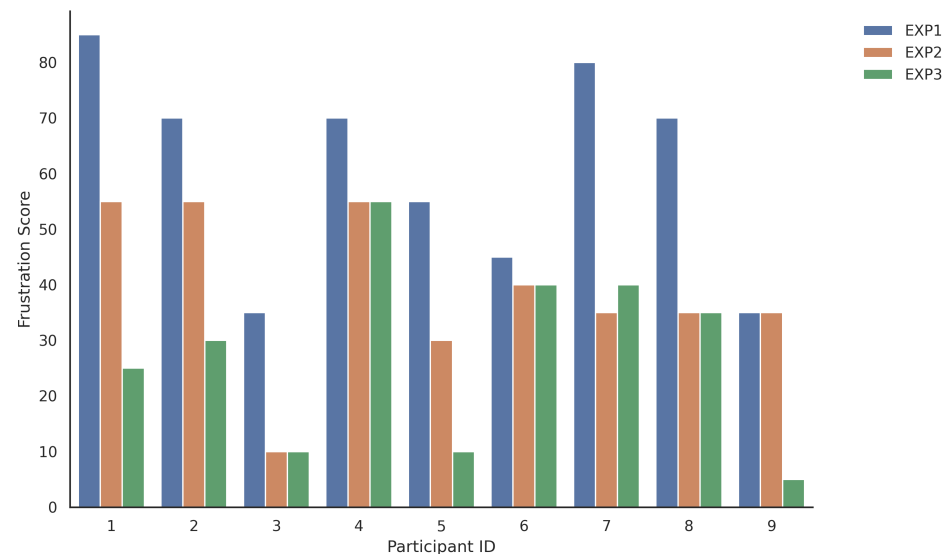
4.5. Subscale Analysis (Secondary)

Dimension-level paired differences between EXP1 and EXP2 are reported in Table 11. After Holm–Bonferroni correction for multiplicity across the six subscales, both Frustration ($M_{diff} = 21.67$, $t(8) = 4.53$, $p = 0.002$, Holm-adjusted $p = 0.012$) and Effort ($M_{diff} = 15.00$, $t(8) = 3.46$, $p = 0.009$, Holm-adjusted $p = 0.043$) remained significant at $\alpha = 0.05$. The remaining four subscales: Mental Demand, Physical Demand, Temporal Demand, and Performance, did not survive correction, although Mental Demand, Physical Demand, and Temporal Demand all showed nominally positive differences in the same direction as the overall TLX result. This pattern is descriptively consistent with H2, which predicted that any workload increase would be most apparent in the Effort and Frustration dimensions, with the caveat that the order confound discussed in Section 3.1 applies to the subscale results as it does to the overall score. The pattern of subscale differences suggests that the higher workload ratings in EXP1 were most clearly expressed as heightened affective load and invested Effort, with Frustration showing the largest observed difference. The Frustration difference of 21.67 points is consistent with the informal post-exercise observations reported in Section 4.6, including controller frustration attributed to non-standard communications, speed differential management, and the perceived unpredictability of requests. Given the sample size ($n = 9$), these dimension-level estimates are unstable; the confidence intervals in Table 11 convey the associated uncertainty, and the subscale findings, including those surviving Holm–Bonferroni correction, should be interpreted as exploratory. Individual frustration ratings across all three scenarios are shown in Figure 6.

Table 11. Subscale paired differences (EXP1–EXP2), $n = 9$.

Subscale	EXP1 M (SD)	EXP2 M (SD)	Diff [95% CI]	d_z [95% CI]	$t(8)$	p	p_{Holm}
Mental Demand	62.22 (17.52)	53.89 (19.49)	8.33 [−5.29, 21.95]	0.47 [−0.23, 1.15]	1.41	0.195	0.529
Physical Demand	37.78 (22.52)	28.33 (19.84)	9.44 [−5.27, 24.15]	0.49 [−0.22, 1.18]	1.48	0.176	0.529
Temporal Demand	52.78 (21.52)	43.33 (20.31)	9.44 [−1.34, 20.22]	0.67 [−0.07, 1.39]	2.02	0.078	0.312
Performance	36.11 (12.19)	35.00 (15.21)	1.11 [−11.08, 13.30]	0.07 [−0.59, 0.72]	0.21	0.838	0.838
Effort	65.00 (23.45)	50.00 (19.84)	15.00 [5.00, 25.00]	1.15 [0.28, 1.99]	3.46	0.009	0.043 **
Frustration	60.56 (18.78)	38.89 (14.74)	21.67 [10.64, 32.70]	1.51 [0.51, 2.47]	4.53	0.002	0.012 **

** Significant after Holm–Bonferroni correction at $\alpha = 0.05$. p values are two-sided paired t -tests; p_{Holm} are step-down Holm-adjusted values. CIs for d_z are noncentral- t intervals derived from the reported t statistics.

**Figure 6.** Individual frustration ratings across all three scenarios. Ratings were higher in EXP1 than in EXP2 for eight of nine participants.

4.6. Informal Post-Exercise Observations

The following observations are reported as informal contextual information rather than as systematically collected qualitative evidence. Participants' spontaneous comments, collected after the final exercise (EXP3), were not obtained through a structured interview protocol and were not subjected to formal qualitative coding; they are reported solely as contextual observations that may inform interpretation of the observed NASA-TLX pattern. Participants referred to frequency congestion due to prolonged non-IFR communications, speed differentials increasing exercise difficulty, dispersed attention due to the non-IFR traffic as a potential conflict, unpredictability from non-standard requests, and frustration from uncertain transmissions. These impressions are consistent with the higher observed Effort and Frustration scores in EXP1, but they do not establish that these scenario features caused the workload difference.

5. Discussion

5.1. Interpretation of Main Findings

The mixed IFR/VFR scenario was associated with higher perceived workload relative to the operationally comparable IFR-only baseline. The largest observed dimension-level differences were in Frustration and Effort. The moderate overall effect size ($d_z = 0.67$) and directional consistency (7/9 participants) suggest a potentially meaningful operational pattern, although the 95% CI [−1.33, 19.85] indicates substantial uncertainty and the need for a larger sample. In operational terms, the primary difference corresponds to an approximately 20% relative increase in weighted NASA-TLX over the IFR-only baseline

(from 46.1 to 55.3 on the 0–100 scale). No validated operational criterion, such as a red-line value, exists for the NASA-TLX in approach control, so the practical significance of a difference of this size cannot be established from subjective ratings alone; relating TLX shifts of this magnitude to objective performance margins is one purpose of the planned confirmatory study.

The constant number of planned conflicts (6) across all exercises indicates that EXP1 was not designed with a higher number of scripted tactical conflict situations. However, the mixed scenario did contain greater scripted communication activity: EXP1 required 27.5% more R/T transmissions than the IFR-only baseline (EXP2) and remained 10% higher than the high-density scenario (EXP3), despite the latter having a larger aircraft count. Because these counts were generated from the pseudopilot script, they should be interpreted as a scenario-design indicator of planned communication demand rather than as an independent behavioural validation of the TLX results.

A plausible mechanism is increased R/T load and attentional switching: integrating slower, less structured VFR flights into IFR streams amplifies sequencing constraints, imposes non-standard communication overhead, and introduces pilot behaviour uncertainty. An important caveat is that only one of the three non-IFR aircraft (9AZAO) was scripted with degraded communication behaviour, while the other two (9ADMB, 9ADDD) used standard phraseology. If the Frustration elevation is primarily driven by the degraded-comms profile of this single aircraft, the present design cannot separate “VFR presence” from “degraded communication quality” as independent contributors. A larger study in which communication quality and flight-rule heterogeneity are manipulated as independent factors would be required to disentangle these explanations. This is reflected in Limitation L4 in Table 12.

Table 12. Summary of limitations, bias directions, and planned mitigations.

#	Threat	Bias Direction	Scope	Mitigation (Planned)
L1	Fixed scenario order (EXP1 → EXP2 → EXP3)	Ambiguous: learning biases against H1; fatigue/first-exercise anxiety biases for H1	Primary	Latin-square counterbalancing
L2	Small trainee sample ($n = 9$, pre-rating)	Widens CIs; limits external validity to trainees	Primary	Confirmatory study with rated approach controllers, powered prospectively for a prespecified smallest effect of operational interest
L3	Change in instructor role (Day 1 → Day 2)	Potentially for H1 (EXP1 is the first exercise under passive support)	Moderate	Uniform passive-observer protocol across all exercises
L4	Single pseudopilot with degraded comms on one Zulu-flight aircraft (9AZAO)	Potentially for H1 (conflates flight-rule heterogeneity with communication quality)	Limited	Separate VFR presence and communication quality as independent variables
L5	No granular objective metrics (separation margins, eye-tracking)	–	Secondary	Full simulator replay-log analysis; eye-tracking integration
L6	Convenience sampling from single institution	Limits generalisability	Secondary	Multi-site recruitment

EXP3 contained two more IFR aircraft than EXP2, but the observed weighted NASA-TLX difference between these two later scenarios was small ($|d_z| = 0.04$). Because both scenarios occurred after EXP1, this descriptive contrast cannot identify the respective roles of aircraft count, traffic heterogeneity, or learning and adaptation; whether structure-based abstractions of the kind described by Histon and Hansman [8] moderate such effects is a question for the counterbalanced confirmatory study.

From an operational perspective, if future counterbalanced studies confirm the association between VFR integration and elevated workload, this would carry implications for CAPAN-based capacity declarations. However, no validated conversion between NASA-TLX scores and CAPAN working-time percentages exists, so the present data cannot quantify the operational magnitude of any such effect. This remains an open question for future research combining subjective and objective workload measures.

5.2. Comparison with Prior Work

5.2.1. NASA-TLX in ATC Research

The NASA-TLX is among the most widely used subjective workload instruments in ATC research [42]. The overall weighted scores observed in the present study (condition means between 46 and 55 on the 0–100 scale) fall in the mid-range of the instrument, indicating moderate workload levels appropriate for an exploratory manipulation with trainee participants.

5.2.2. Traffic Heterogeneity as a Complexity Driver

Kopardekar and Magyarits [5] demonstrated that interactions, not density, drive workload. In the present data, the two later IFR-only scenarios showed little observed difference, while the mixed scenario showed the highest observed mean; because scenario condition and sequence position were confounded, this descriptive pattern cannot test an extension of this principle to flight-rule heterogeneity. Histon and Hansman's [8] model, in which VFR aircraft disrupt traffic regularities and force state-based monitoring, remains a candidate mechanism for the counterbalanced study. Related data-driven findings indicate that interaction and conflict-geometry features carry information about workload beyond aircraft count [13,17]. The present results suggest that flight-rule heterogeneity may be an additional interaction-type feature worth incorporating into such complexity models.

5.2.3. Communication Overhead

Manning et al. [34] identified R/T communication time as a reliable workload proxy. VFR traffic generates non-standard communication: FIC coordination, special clearances, and traffic information required by Class C rules. EXP1 contained more scripted R/T activity and additional non-standard coordination requirements than EXP2; these scenario features provide a plausible communication-overhead mechanism, although the present design cannot isolate their contribution to perceived workload. Participants' informal post-exercise impressions frequently referred to communication demands, speed differentials, and unpredictability; these observations provide contextual support for interpreting the elevated Effort and Frustration ratings, but they were not collected or analysed as formal qualitative data and cannot identify the causal mechanism underlying the scenario difference. CAPAN's inability to account for this supports subjective measurement as a complement to time-based models.

5.2.4. Small-Sample ATC Studies

Within-subject simulator designs with $n < 12$ are not uncommon in exploratory ATC research. Oprins et al. [35] used $n = 10$; Metzger and Parasuraman [7] used $n = 8$. Such designs cannot, however, substitute for adequately powered confirmatory studies. The present work should be interpreted as an exploratory study estimating the magnitude and direction of the effect.

5.3. Limitations and Threats to Validity

Table 12 summarises the principal threats to validity identified in this study, their expected direction of bias relative to H1, and the planned mitigations for the confirmatory

study. Despite these limitations, the within-subject design eliminates between-participant variability, and the observed effect direction is consistent across seven of nine participants and aligns with theoretical predictions regarding traffic heterogeneity [5,8]. The wide 95% CI for the primary comparison ($[-1.33, 19.85]$) and weak, anecdotal-range Bayes factors ($BF_{10} = 1.31$, $BF_{+0} = 2.47$) reinforce the pilot-study framing: the present data are sufficient to estimate the direction and approximate magnitude of the effect, but not to draw firm inferential conclusions. The confirmatory design summarised in Section 5.4.4 is intended to address this.

A further boundary of the present evidence is that it is confined to self-reported workload. The scenario-level metrics (planned conflicts, per-scenario transmission counts) characterise scenario design rather than participant behaviour, and no replay-log analyses of separation margins, losses of separation, issued instructions, transmission durations, readback corrections, or handoff timing were performed. The present data therefore cannot determine whether the higher perceived workload in EXP1 was accompanied by degraded operational performance or reflected greater subjective cost alone; integrating such objective metrics is a core element of the planned confirmatory study. The transmission counts likewise describe the planned communication demand embedded in the scenario scripts; actual transmission frequency, duration, repetitions, and deviations from the script were not quantified at the participant level, so the study cannot determine whether realised communication activity differed across participants or mediated the observed workload ratings.

5.4. Operational Implications and Future Work

The considerations in this subsection are conditional on replication: none is directly tested by the present exploratory data, and each is contingent on confirmation in a counter-balanced study with operational controllers.

5.4.1. VFR Sequencing Procedures

If confirmed in a larger counterbalanced study, the >100 kt speed differential observed in this study would suggest a need for: explicit VFR holding points outside IFR final approach segments; published VFR approach segments avoiding IFR geometries; and controller guidance on minimum sequencing gaps.

5.4.2. Communication Management

Rather than introducing redundant pre-flight briefings or phraseology requirements that are already mandated by national licensing regulations, practical interventions could focus on the ATC–VFR interface itself. These include greater use of the dedicated FIC frequency available in the Zagreb FIR to offload non-tactical VFR coordination from the approach frequency, controller-side decision support for anticipating VFR routing requests, and operational guidance for managing prolonged or non-standard VFR transmissions during peak IFR sequencing phases.

5.4.3. Capacity Planning

If the association between VFR integration and elevated workload is confirmed, CAPAN-based capacity declarations may underestimate mixed-mode workload. ANSPs could consider evaluating VFR-adjusted thresholds during periods of high VFR activity. Near-term tools: IFR–VFR proximity alerting and sequencing aids displaying VFR ground speed.

A further consideration concerns the retrospective analysis of airspace workload. VFR movements in the Zagreb TMA are not systematically recorded in the same manner as IFR flights, a gap also highlighted in recent IFATCA working papers on VFR traffic and airspace

capacity [32], which carries direct implications for any capacity or workload assessment built on historical traffic data. Simulation scenarios reconstructed from such records would, by construction, omit the VFR component, and any resulting workload estimate would systematically underrepresent the true demand placed on the controller. If the present findings are confirmed, this implies that historical-data-driven workload analyses of mixed-mode TMAs are likely to be biased downward, and that capacity declarations derived from them may underestimate actual controller load during periods of substantial VFR activity.

5.4.4. Future Research

Four priorities: (1) Larger confirmatory study with rated controllers and counterbalanced order, with the sample size determined prospectively from a prespecified smallest effect of operational interest (for reference, $d_z = 0.40/0.50/0.67 \rightarrow n \approx 52/34/20$ at 80% power, two-sided $\alpha = 0.05$). (2) Objective measurement: R/T logging, replay analysis, eye-tracking, and voice-based workload markers [29,30]. (3) Replication across different TMAs and operational contexts. In this replication, a factorial design manipulating flight-rule composition and communication quality as independent factors would allow the confounded contribution of the scripted degraded communication in EXP1 to be separated from the effect of VFR presence itself. (4) Extension of the Bayesian analysis reported herein ($BF_{+0} = 2.47$) using informative priors derived from the traffic heterogeneity literature, which may yield more decisive evidence than the default JZS prior used in the present analysis.

6. Conclusions

This exploratory study examined perceived ATCO workload across three within-subject simulator exercises in the Zagreb TMA. In the fixed scenario order employed (EXP1 mixed IFR/VFR, then EXP2 IFR-only, then EXP3 IFR-only with two additional aircraft), the first exercise was associated with higher NASA-TLX scores than the two subsequent IFR-only exercises, with the largest dimensional difference observed in Frustration. Whether this pattern reflects the mixed-mode scenario package, sequence position, the transition from active to passive instructor support, or some combination of these factors cannot be determined from the present data. To the authors' knowledge, no prior study has reported NASA-TLX data comparing mixed IFR/VFR operations against an operationally matched IFR-only baseline in a terminal approach simulation, and the present results estimate effect magnitude and direction for a counterbalanced confirmatory study.

Three key takeaways emerge:

First, the contrast between EXP1 and EXP2 yielded a moderate estimated within-subject effect ($d_z = 0.67$; 95% CI $[-1.33, 19.85]$; two-sided $p = 0.079$), while the contrast between EXP2 and EXP3, which differed only in IFR aircraft count, showed little observed difference ($|d_z| = 0.04$, $p = 0.904$). Because scenario condition and sequence position were fully confounded, these patterns cannot isolate the effects of traffic heterogeneity, density, learning, or the transition to passive instructor support.

Second, the largest observed subscale differences were in Frustration and Effort. Informal post-exercise impressions highlighted communication overhead, speed differentials, and unpredictability as potentially demanding features of EXP1. Because these impressions were not collected through a structured qualitative protocol, they are reported as contextual observations rather than as independent evidence of a causal mechanism. Because the participants were trainees and the scenario order was fixed, these patterns remain exploratory and may not transfer directly to rated controllers.

Third, if confirmed in a counterbalanced study, these findings would suggest that CAPAN-based models may underestimate workload in mixed-mode terminal operations,

particularly during periods of high VFR activity. The 9.26-point TLX difference, while requiring confirmation with a larger sample and counterbalanced design, highlights the possibility that current capacity declarations do not account for VFR-related cognitive overhead. Because VFR movements are not systematically recorded in the same way as IFR traffic, retrospective workload analyses based on historical data are likely to underrepresent actual controller demand in mixed-mode TMAs. Incorporating VFR-adjusted capacity thresholds and supporting controllers with proximity alerting and sequencing tools may mitigate operational risks.

As a within-subject exploratory study with fixed scenario order and a small sample of trainees, these findings motivate a larger confirmatory investigation with certified controllers, counterbalanced designs, and objective workload proxies (R/T analysis, eye-tracking). If the observed effect size ($d_z = 0.67$) reflects the true population effect, an illustrative calculation indicates that approximately 20 participants would provide 80% power for the planned replication (two-sided $\alpha = 0.05$; 24 with the bias-corrected $g_z = 0.61$). However, the wide bootstrap interval around this estimate means the true required sample could be considerably larger.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/aerospace13070653/s1>, Supplementary Protocol S1: Preserved scenario and pseudopilot specifications; Analysis Code S1: Individual scores and dimension-level summary statistics.

Author Contributions: Conceptualisation, A.Ž. and B.A.-F.; methodology, A.Ž., B.A.-F. and K.K.N.; software, A.Ž.; validation, K.K.N., B.A.-F. and P.A.; formal analysis, A.Ž., B.A.-F. and K.K.N.; investigation, A.Ž.; resources, A.Ž. and B.A.-F.; data curation, A.Ž. and K.K.N.; writing—original draft preparation, A.Ž., K.K.N. and B.A.-F.; writing—review and editing, K.K.N., B.A.-F. and P.A.; visualisation, K.K.N.; supervision, B.A.-F. and K.K.N.; project administration, A.Ž. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Union under the National Recovery and Resilience Plan (NPOO), within a project “Digital Center for Education and Innovation in Air Traffic Management DECENT-ATM”, funded through Mechanism 581—Recovery and Resilience Facility (NextGenerationEU).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki. The Ethics Committee of the Faculty of Transport and Traffic Sciences, University of Zagreb, reviewed the study protocol and waived the requirement for full ethical approval due to the non-invasive nature of the simulator exercises and the use of anonymised subjective workload ratings.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The participant-level weighted NASA-TLX scores are reported in Table 9 and the scenario flight plans in Appendix A; the analysis code and the scripting protocol (Supplementary Protocol S1) accompany this article as Supplementary Material. The dimension-level ratings and pairwise weights are available from the corresponding author on reasonable request.

Acknowledgments: The authors would like to thank the instructor Skanija Lukar from Croatia Control Ltd. for her professional guidance and expertise provided during the preparatory phases of the research exercises, and the instructor Aleksandar Jazić for supervising the simulator exercises and mentoring the participants.

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

Appendix A

This appendix contains detailed flight plans for all scenarios.

Appendix A.1. Preparatory Exercise Flight Plans

Table A1. Flight plans for preparatory exercises PRP1 and PRP2.

No.	Callsign	Type	Flight	Start	DEP	DEST	FL	SSR	Route
PRP1									
1	9ADDM	C172	VFR	00:00:15	LDZL	LDZL	A030	0001	W1 PIS ZAG
2	9AOFK	C172	VFR	00:03:00	LDZL	LDVA	A065	0002	N2 N1
3	AUA497	A320	IFR	00:08:00	LOWW	LDZA	F170	6656	PETOV PIS
4	BAW3467	B738	IFR	00:03:40	EGLL	LDZA	F140	4343	ARGOM PIS
5	CTN22P	DHC8C	IFR	00:01:30	LDZD	LDZA	F130	6535	KULEN KOTOR PIS
6	EWG1210	B737	IFR	00:17:00	EDDN	LDZA	F140	5464	ARGOM PIS
7	EZY231	A320	IFR	00:14:00	LDDU	LDZA	F160	3636	TEBLI PIS
8	KLM767C	B744	IFR	00:00:00	LDZA	EHAM	F360	6321	SK PODET KFT
9	QTR754Q	B744	IFR	00:16:00	OEKK	LDZA	F160	5575	VBA PIS
10	RYR303	B738	IFR	00:00:00	LDZA	HECA	F380	5563	SK NIVES SONIK
11	WZZ33G	B738	IFR	00:02:00	LHBP	LDZA	F220	5465	KOPRY PIS
PRP2									
1	EWG33N	A320	IFR	00:04:30	EDDF	LDZA	F140	4544	ARGOM PIS
2	CTN2110	A320	IFR	00:01:40	LDDU	LDZA	F130	3321	RUDIK PIS
3	9AWGH	C172	Z	00:15:00	LDVA	LDOS	A060	0002	RASIN VBA CE
4	THY343	B738	IFR	00:00:00	LDZA	LTBA	F360	5656	SK TEBLI DOBOT
5	CTN498	A320	IFR	00:00:00	LDZA	LIRF	F330	6555	SK KOTOR KULEN NIKOL
6	BAW21NQ	A333	IFR	00:00:00	LDZA	EGLL	F380	1113	SK PODET KFT
7	9ABBQ	C172	VFR	00:00:15	LDZL	LDZL	A060	0003	N2 A1
8	9AOFK	C172	VFR	00:03:00	LDZL	LDRI	A065	0001	W1 TNJ RJK
9	WZZ500	B738	IFR	00:01:00	LHBP	LDZA	F220	5655	KOPRY PIS
10	QTR669	A380	IFR	00:12:15	LLBG	LDZA	F160	4676	VBA PIS
11	AUA707	DHC8C	IFR	00:02:00	LOWW	LDZA	F170	6231	PETOV PIS
12	IBE89	A320	IFR	00:22:00	LSZH	LDZA	F130	3121	KULEN KOTOR PIS
13	RYR74	B738	IFR	00:18:00	LWSK	LDZA	F130	3673	RUDIK PIS

Note: DEP = departure aerodrome; DEST = destination; FL = flight level; SSR = transponder code.

Appendix A.2. Research Exercise Flight Plans

Table A2. Flight plan for EXP1 (mixed IFR/VFR scenario).

No.	Callsign	Type	Flight	Start	DEP	DEST	FL	SSR	Route
1	9ADDD	C172	VFR	00:03:30	LDZL	LDZL	A070	0004	N2 A1 N2 A1
2	9ADMB	C172	VFR	00:00:05	LDZL	LDZL	A030	0001	W1 PIS ZAG
3	9AZAO	C172	Z	00:05:00	LDVA	LDOS	A060	0002	RASIN VBA CE
4	AUA668	DHC8C	IFR	00:00:00	LDZA	LOWW	F210	4333	SK OBUTI GRZ
5	BAW36R	A320	IFR	00:04:00	LDZD	LDZA	F130	4321	KULEN KOTOR PIS
6	CTN101	A320	IFR	00:00:00	LDZA	LIRF	F310	6567	SK KOTOR GISER
7	EWG89	B738	IFR	00:06:30	EHAM	LDZA	F140	6751	ARGOM PIS
8	KLM977	B744	IFR	00:12:00	EDDF	LDZA	F170	3233	PETOV PIS
9	LOT321	B738	IFR	00:12:00	LHBP	LDZA	F220	5454	KOPRY PIS
10	RYR55C	B744	IFR	00:10:00	LLBG	LDZA	F170	3214	TEBLI PIS
11	RYR7879	A320	IFR	00:00:00	LDZA	LDSP	F260	4443	SK NIVES UNIPA
12	THY34	A320	IFR	00:18:00	LTBA	LJLJ	F210	5554	VBA ROLBA DOL

Note: VFR aircraft (9ADDD, 9ADMB) and Zulu flight (9AZAO) highlighted. All other traffic IFR.

Table A3. Flight plans for EXP2 and EXP3 (IFR-only scenarios).

No.	Callsign	Type	Flight	Start	DEP	DEST	FL	SSR	Route
EXP2 (IFR-only baseline)									
1	AUA668	DHC8C	IFR	00:00:00	LDZA	LOWW	F210	4333	SK OBUTI GRZ
2	BAW36R	A320	IFR	00:02:00	LDZD	LDZA	F130	4321	KULEN KOTOR PIS
3	CTN101	A320	IFR	00:00:00	LDZA	LIRF	F310	6567	SK KOTOR GISER
4	CTN335	A320	IFR	00:03:15	LWSK	LDZA	F160	3335	TEBLI PIS
5	DLH663	A320	IFR	00:00:00	LDZA	EDDF	F340	6636	SK PODET KFT
6	EWG89	B738	IFR	00:04:30	EHAM	LDZA	F140	6751	ARGOM PIS
7	KLM977	B744	IFR	00:10:00	EDDF	LDZA	F170	3233	PETOV PIS
8	LOT321	B738	IFR	00:11:00	LHBP	LDZA	F220	5454	KOPRY PIS
9	RYR55C	B744	IFR	00:12:00	LLBG	LDZA	F170	3214	TEBLI PIS
10	RYR7879	A320	IFR	00:00:00	LDZA	LDSP	F260	4443	SK NIVES UNIPA
11	THY34	A320	IFR	00:16:00	LTBA	LJLJ	F210	5554	VBA ROLBA DOL
12	WZZ554	B738	IFR	00:00:00	LDZA	LHBP	F320	5544	SK KOPRY SVR
EXP3 (IFR + 2)									
1	AUA668	DHC8C	IFR	00:00:00	LDZA	LOWW	F210	4333	SK OBUTI GRZ
2	BAW36R	A320	IFR	00:02:00	LDZD	LDZA	F130	4321	KULEN KOTOR PIS
3	CTN101	A320	IFR	00:00:00	LDZA	LIRF	F310	6567	SK KOTOR GISER
4	CTN335	A320	IFR	00:03:15	LWSK	LDZA	F160	3335	TEBLI PIS
5	DLH663	A320	IFR	00:00:00	LDZA	EDDF	F340	6636	SK PODET KFT
6	EWG89	B738	IFR	00:04:30	EHAM	LDZA	F140	6751	ARGOM PIS
7	KLM977	B744	IFR	00:10:00	EDDF	LDZA	F170	3233	PETOV PIS
8	LOT321	B738	IFR	00:11:00	LHBP	LDZA	F220	5454	KOPRY PIS
9	RYR55C	B744	IFR	00:12:00	LLBG	LDZA	F170	3214	TEBLI PIS
10	RYR7879	A320	IFR	00:00:00	LDZA	LDSP	F260	4443	SK NIVES UNIPA
11	THY34	A320	IFR	00:16:00	LTBA	LJLJ	F210	5554	VBA ROLBA DOL
12	WZZ554	B738	IFR	00:00:00	LDZA	LHBP	F320	5544	SK KOPRY SVR
13	RAC0003	AT72	IFR	00:14:00	LHBP	LQBK	F150	1112	KOPRY TEBLI KOMAR
14	EWG66P	A320	IFR	00:18:00	EHAM	LDZA	F170	1115	PETOV PIS

Note: EXP3 adds two aircraft (RAC0003, EWG66P) to the EXP2 baseline. Aircraft start times in EXP2/EXP3 were adjusted relative to EXP1 by 1–2 min to preserve the temporal distribution of planned conflicts after VFR removal.

References

1. Eurocontrol. *EUROCONTROL Seven-Year Forecast 2025–2031. Autumn 2025 Update*; Technical Report; Eurocontrol: Brussels, Belgium, 2025.
2. Hilburn, B. *Cognitive Complexity in Air Traffic Control: A Literature Review*; EEC Note No. 04/04; EUROCONTROL Experimental Centre: Brétigny-sur-Orge, France, 2004.
3. Eurocontrol. *CAPAN Theoretical Foundation and User Manual*; Technical Report; Eurocontrol: Brussels, Belgium, 2004.
4. Majumdar, A.; Ochieng, W.Y. Factors affecting air traffic controller workload: Multivariate analysis based on simulation modeling of controller workload. *Transp. Res. Rec.* **2002**, *1788*, 58–69. [\[CrossRef\]](#)
5. Kopardekar, P.; Magyarits, S. Measurement and prediction of dynamic density. In *Proceedings of the 5th USA/Europe Air Traffic Management R&D Seminar*, Budapest, Hungary, 23–27 June 2003.
6. Loft, S.; Sanderson, P.; Neal, A.; Mooij, M. Modeling and predicting mental workload in en route air traffic control: Critical review and broader implications. *Hum. Factors* **2007**, *49*, 376–399. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Metzger, U.; Parasuraman, R. Automation in future air traffic management: Effects of decision aid reliability on controller performance and mental workload. *Hum. Factors* **2005**, *47*, 35–49. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Histon, J.M.; Hansman, R.J. *Mitigating Complexity in Air Traffic Control: The Role of Structure-Based Abstractions*; MIT International Center for Air Transportation Report; MIT: Cambridge, MA, USA, 2008.
9. Pérez Moreno, F.; Gómez Comendador, V.F.; Delgado-Aguilera Jurado, R.; Zamarreño Suárez, M.; Antulov-Fantulin, B.; Arnaldo Valdés, R.M. How Has the Concept of Air Traffic Complexity Evolved? Review and Analysis of the State of the Art of Air Traffic Complexity. *Appl. Sci.* **2024**, *14*, 3604. [\[CrossRef\]](#)
10. Sultana, T.; Rifat, M.R.I.; Mahmud, M.F.; Hossain, M.D.; Biswas, M.E.; Lee, G.-W.; Huh, E.-N. Workload prediction for air traffic controllers: A comprehensive review of challenges, techniques, and future prospects. *Int. J. Data Sci. Anal.* **2026**, *22*, 148. [\[CrossRef\]](#)

11. Isufaj, R.; Koca, T.; Piera, M.A. Spatiotemporal Graph Indicators for Air Traffic Complexity Analysis. *Aerospace* **2021**, *8*, 364. [\[CrossRef\]](#)
12. Zhang, L.; Yang, H.; Wu, X. Air Traffic Complexity Evaluation with Hierarchical Graph Representation Learning. *Aerospace* **2023**, *10*, 352. [\[CrossRef\]](#)
13. Pang, Y.; Hu, J.; Lieber, C.S.; Cooke, N.J.; Liu, Y. Air traffic controller workload level prediction using conformalized dynamical graph learning. *Adv. Eng. Inform.* **2023**, *57*, 102113. [\[CrossRef\]](#)
14. Henderson, E.; Gould, D.; Everson, R.; De Ath, G.; Pepper, N. Air Traffic Controller Task Demand via Graph Neural Networks: An Interpretable Approach to Airspace Complexity. In Proceedings of the AIAA AVIATION Forum, Las Vegas, NV, USA, 21–25 July 2025.
15. Xie, H.; Zhang, M.; Ge, J.; Dong, X.; Chen, H. Learning Air Traffic as Images: A Deep Convolutional Neural Network for Airspace Operation Complexity Evaluation. *Complexity* **2021**, *2021*, 6457246. [\[CrossRef\]](#)
16. Chen, H.; Zhang, L.; Yuan, L.; Zhu, W.; Liu, L. Air Traffic Complexity Assessment Based on Ordered Deep Metric. *Aerospace* **2022**, *9*, 758. [\[CrossRef\]](#)
17. Pérez Moreno, F.; Gómez Comendador, V.F.; Delgado-Aguilera Jurado, R.; Zamarreño Suárez, M.; Janisch, D.; Arnaldo Valdés, R.M. Determination of Air Traffic Complexity Most Influential Parameters Based on Machine Learning Models. *Symmetry* **2022**, *14*, 2629. [\[CrossRef\]](#)
18. Pérez Moreno, F.; Ibáñez Rodríguez, F.; Gómez Comendador, V.F.; Delgado-Aguilera Jurado, R.; Zamarreño Suárez, M.; Arnaldo Valdés, R.M. Prediction of air traffic complexity through a dynamic complexity indicator and machine learning models. *J. Air Transp. Manag.* **2024**, *119*, 102632. [\[CrossRef\]](#)
19. Hart, S.G.; Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Human Mental Workload*; Hancock, P.A., Meshkati, N., Eds.; North-Holland: Amsterdam, The Netherlands, 1988; pp. 139–183. [\[CrossRef\]](#)
20. Dekker, S.; Hollnagel, E. Human factors and folk models. *Cogn. Technol. Work* **2004**, *6*, 79–86. [\[CrossRef\]](#)
21. Kirwan, B.; Donohoe, S.; Lamoureux, T.; MacKendrick, H.; Shorrock, S. The application of human factors to the ATM system design and integration process. In Proceedings of the 4th USA/Europe Air Traffic Management R&D Seminar, Santa Fe, NM, USA, 3–7 December 2001.
22. Mogford, R.H.; Guttman, J.A.; Morrow, S.L.; Kopardekar, P. *The Complexity Construct in Air Traffic Control: A Review and Synthesis of the Literature*; DOT/FAA/CT-TN95/22; FAA: Atlantic City, NJ, USA, 1995.
23. Seamster, T.L.; Redding, R.E.; Cannon, J.R.; Ryder, J.M.; Purcell, J.A. Cognitive Task Analysis of Expertise in Air Traffic Control. *Int. J. Aviat. Psychol.* **1993**, *3*, 257–283. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Pawlak, W.S.; Brinton, C.R.; Crouch, K.; Lancaster, K.M. A framework for the evaluation of air traffic control complexity. In Proceedings of the AIAA Guidance, Navigation, and Control Conference, San Diego, CA, USA, 29–31 July 1996.
25. Gutiérrez Teuler, G.; Arnaldo Valdés, R.M.; Gómez Comendador, V.F.; López de Frutos, P.M.; Rodríguez Rodríguez, R. Study of the Impact of Traffic Flows on the ATC Actions. *Aerospace* **2022**, *9*, 467. [\[CrossRef\]](#)
26. Delgado-Aguilera Jurado, R.; Gómez Comendador, V.F.; Zamarreño Suárez, M.; Pérez Moreno, F.; Verdonk Gallego, C.E.; Arnaldo Valdés, R.M. Assessment of Potential Conflict Detection by the ATCo. *Aerospace* **2022**, *9*, 522. [\[CrossRef\]](#)
27. Samardžić, K.; Andraši, P.; Radišić, T.; Novak, D. Probabilistic Air Traffic Complexity Analysis Considering Prediction Uncertainties in Traffic Scenarios. *Aerospace* **2024**, *11*, 798. [\[CrossRef\]](#)
28. Muñoz-de-Escalona, E.; Leva, M.C.; Cañas, J.J. Mental Workload as a Predictor of ATCO's Performance: Lessons Learnt from ATM Task-Related Experiments. *Aerospace* **2024**, *11*, 691. [\[CrossRef\]](#)
29. Muñoz-de-Escalona, E.; Leva, M.C.; Gianini, G.; López de Frutos, P.; Jadronova, M. Air traffic controllers communication analysis as a proxy of task demand and mental workload: Using voice recording markers for safety critical task. *Saf. Sci.* **2025**, *191*, 106928. [\[CrossRef\]](#)
30. Yang, J.; Yang, H.; Wu, Z.; Wu, X. Cognitive Load Assessment of Air Traffic Controller Based on SCNN-TransE Network Using Speech Data. *Aerospace* **2023**, *10*, 584. [\[CrossRef\]](#)
31. Netjasov, F. Framework for airspace planning and design based on conflict risk assessment. Part 1: Conflict risk assessment model for airspace strategic planning. *Transp. Res. Part C Emerg. Technol.* **2012**, *24*, 190–212. [\[CrossRef\]](#)
32. IFATCA. *VFR Flights and Airspace Capacity Model Analyses*; Working Paper No. 99, 63rd Annual Conference; International Federation of Air Traffic Controllers' Associations: Singapore, 2024.
33. IFATCA. *VFR Traffic Workload and Capacity Study*; Working Paper No. 160, 63rd Annual Conference; International Federation of Air Traffic Controllers' Associations: Singapore, 2024.
34. Manning, C.A.; Mills, S.H.; Fox, C.M.; Pfeleiderer, E.M.; Mogilka, H.J. *Using Air Traffic Control Taskload Measures and Communication Events to Predict Subjective Workload*; DOT/FAA/AM-02/4; FAA Office of Aerospace Medicine: Washington, DC, USA, 2002.
35. Oprins, E.; Burggraaff, E.; van Weerdenburg, H. Design of a competence-based assessment system for air traffic control training. *Int. J. Aviat. Psychol.* **2006**, *16*, 297–320. [\[CrossRef\]](#) [\[PubMed\]](#)

36. Department of Aeronautics (Zavod za Aeronautiku), Faculty of Transport and Traffic Sciences, University of Zagreb. Laboratory for Air Traffic Control. Available online: <https://www.fpz.unizg.hr/zan/research/laboratory-for-air-traffic-control/> (accessed on 10 July 2026).
37. European Commission. Commission Regulation (EU) 2015/340 of 20 February 2015 laying down technical requirements and administrative procedures relating to air traffic controllers' licences and certificates. *Off. J. Eur. Union* **2015**, *63*, 1–122. [[CrossRef](#)]
38. ICAO. *Annex 11 to the Convention on International Civil Aviation: Air Traffic Services*, 14th ed.; International Civil Aviation Organization: Montreal, QC, Canada, 2016.
39. Croatia Control Ltd. *Aeronautical Information Publication (AIP) Croatia*; Croatia Control Ltd.: Zagreb, Croatia, 2025.
40. Montgomery, D.C. *Design and Analysis of Experiments*, 9th ed.; John Wiley & Sons: Hoboken, NJ, USA, 2017.
41. Rouder, J.N.; Speckman, P.L.; Sun, D.; Morey, R.D.; Iverson, G. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.* **2009**, *16*, 225–237. [[CrossRef](#)] [[PubMed](#)]
42. Hart, S.G. NASA-Task Load Index (NASA-TLX); 20 Years Later. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **2006**, *50*, 904–908. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.