aerospace

MDPI

Article

A Novel Approach Using Non-Experts and Transformation Models to Predict the Performance of Experts in A/B Tests

Phillip Stranger [1,2], Peter Judmaier [3], Gernot Rottermanner [3], Carl-Herbert Rokitansky [4], Istvan-Szilard Szilagyi [5], Volker Settgast [2] and Torsten Ullrich [1,2,*]

1 Institute of Computer Graphics and Knowledge Visualization, Graz University of Technology, 8010 Graz, Austria; phillip.stranger@outlook.com
2 Fraunhofer Austria Research GmbH, 8010 Graz, Austria; volker.settgast@fraunhofer.at
3 Fachhochschule St. Pölten Forschungs GmbH, 3100 St. Pölten, Austria; peter.judmaier@fhstp.ac.at (P.J.); gernot.rottermanner@fhstp.ac.at (G.R.)
4 4D Aerospace Research and Simulation GmbH, 5020 Salzburg, Austria; roki@4d-aerospace.com
5 Division of Medical Psychology, Psychosomatics and Psychotherapeutic Medicine, Department of Psychiatry, Psychosomatics and Psychotherapeutic Medicine, Medical University of Graz, 8036 Graz, Austria; istvan.szilagyi@medunigraz.at
* Correspondence: torsten.ullrich@fraunhofer.at

Abstract: The European Union is committed to modernising and improving air traffic management systems to promote environmentally friendly air transport. However, the safety-critical nature of ATM systems requires rigorous user testing, which is hampered by the scarcity and high cost of air traffic controllers. In this article, we address this problem with a novel approach that involves non-experts in the evaluation of expert software in an A/B test setup. Using a transformation model that incorporates auxiliary information from a newly developed psychological questionnaire, we predict the performance of air traffic controllers with high accuracy based on the performance of students. The transformation model uses multiple linear regression and auxiliary information corrections. This study demonstrates the feasibility of using non-experts to test expert software, overcoming testing challenges and supporting user-centred design principles.

Keywords: user evaluation; user study; air traffic management; statistics; transformation models

check for updates

Citation: Stranger, P.; Judmaier, P.; Rottermanner, G.; Rokitansky, C.-H.; Szilagyi, I.-S.; Settgast, V.; Ullrich, T. A Novel Approach Using Non-Experts and Transformation Models to Predict the Performance of Experts in A/B Tests. *Aerospace* **2024**, *11*, 574. https://doi.org/10.3390/aerospace11070574

Academic Editor: Jacek Skorupski

Received: 1 May 2024
Revised: 28 June 2024
Accepted: 10 July 2024
Published: 12 July 2024

1. Motivation

In today's push towards climate neutrality, the aviation industry is at a crossroads of innovation. The European Union has set itself the goal of "modernising and improving air traffic management technologies, procedures and systems" [1] to make air travel more efficient and environmentally friendly [2]. However, this progress must also ensure the highest safety standards from the very beginning. This requirement makes extensive testing in the software development process essential. At the heart of this testing landscape is the involvement of air traffic controllers (ATCs) themselves, whose expertise ensures that the software meets operational realities and end-user needs. However, this critical need for extensive user testing presents a major problem: the scarcity and high cost of readily available ATCs is a significant barrier to achieving the required test volume. The process of software prototyping, from design prototypes to functional prototypes to pilot systems, requires an ever-increasing number of tests. However, these numbers often exceed the availability of ATCs, in terms of both financial feasibility and organisational logistics. A lack of ATCs for user testing, whether due to organisational constraints or financial factors, limits the scope of testing and consequently reduces the depth of user feedback. This reduction in user feedback not only increases the deviation from user-centred development but also increases the risk of overlooking critical user perspectives in the software development lifecycle. To counter this risk, an attractive solution is to broaden the testing pool by including individuals from outside the air traffic management (ATM) domain. The advantages of this approach are obvious: the pool of test subjects can be expanded and is not limited by the availability of air traffic controllers; moreover, any lack of representativeness in terms of age, gender, etc., can be compensated for more easily if the sample pool is larger. Unfortunately, the most important disadvantage is also obvious: it is no longer the target group that is being tested.

The main goal of the study is to take advantage of the benefits of an extended user group without accepting or at least minimizing its disadvantages. This article describes an approach that makes it possible to partially replace experts with non-experts in A/B testing and to exploit the advantages (see Figure 1) without having to accept the disadvantages. Specifically, this article answers the following research questions:

1. Is it possible to perform a meaningful user test without the relevant user group?
2. How large is the error caused by using the wrong user group and how can it be minimized?
3. If the relevant user group is omitted (i.e., no ground truth is available), can the error still be quantified?

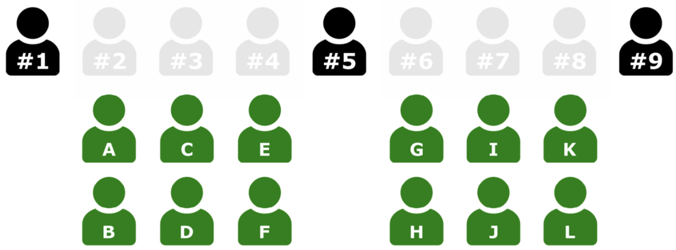


Figure 1. The new approach presented here replaces some expert tests (no. 2, 3, 4, and 6, 7, 8; shown in grey) with non-expert tests (shown in green). Although the wrong target group is used, the results can be converted to the results of the expert tests (indicated by #) through statistical transformations and corrections. If some tests are replaced in this way, and if non-experts are cheaper and more readily available, this approach can both reduce costs and increase the number of tests.

2. Related Work

The EuroControl white paper on human factors highlights that current ATM systems are primarily designed from a functional perspective and focus on presenting a specific set of data to users. However, as Perott et al. note, the presentation of these data often follows a technical rather than a user-centred perspective [3]. As a result, EuroControl advocates a shift towards the user-centred design of ATM systems.

2.1. User-Centred Design

The user-centred design process is a highly iterative approach aimed at rapid prototyping and evaluation to ultimately develop a system that meets user requirements [4]. Research by König et al. demonstrates the suitability of this approach for ATC interface design, as they applied the process to create a planning tool tailored to ATC [5]. Evaluation plays a central role in user-centred design processes [6–9] and represents one of the four phases of the design process [4]. Rubin and Chisnell stress the importance of focusing on users and tasks at an early stage, especially in iterative testing [7]. Similarly, the EuroControl white paper on human factors emphasises the importance of prototyping and evaluation within the iterative design process [3].

2.2. Usability Testing

The evaluation phase of the user-centred design process requires usability evaluation methods to assess the current system. Usability testing involves using real users to test a specific system [7,10,11], with the main objective, as defined by Dumas and Redish, being to improve the usability of the product [12]. Dillon suggests that conducting tests on an application with a group of users performing specific, pre-defined tasks is widely regarded as the most accurate and reliable method for assessing the usability of the application [13]. In addition, Dumas and Redish point out the broad applicability of usability testing in different domains and product types, with test procedures being tailored to the particular context [12]. A comprehensive review by Sagar and Saha highlights usability testing as a prominently used usability evaluation method and covers usability standards, evaluation methods, metrics, and application domains [14].

In practice, usability testing typically involves users performing pre-defined task scenarios, followed by questionnaires or surveys to gather users' opinions or relevant information [15]. For example, in the Bos et al. study, air traffic controllers tested a prototype of an electronic flight strip system. Here, ATCs tested the prototype in two traffic samples, and after each run with the prototype, they completed a questionnaire to evaluate the prototype [16]. In addition, Bos et al. mention that for evaluation purposes, debriefing sessions were held and analyses of simulator logs and video recordings were conducted. Similar methods were used by Huber et al. [17], where ATCs tested prototypes and provided feedback via questionnaires to evaluate interface and interaction concepts.

2.3. A/B Testing

While usability evaluation methods such as usability testing are used to assess a specific system, quantifying the effects of design adjustments requires data-driven methods, of which A/B testing is one of the most common [18]. A/B testing is a method used to evaluate user experience by conducting controlled experiments in which users are randomly exposed to different variants of a service or product [19,20]. Although A/B testing typically involves two variants, it should be noted that any number of variants can be tested, and with a well-designed experiment, the best-performing variant can be identified. As described by Quin et al., A/B testing tests hypotheses in live software systems, with the end users being the participants in the experiment [21]. The hypotheses in this context represent variants of the software system being tested, and the metrics resulting from the A/B test can be used to identify the more user-friendly variant.

A/B testing is widely used in various domains, especially in web, search engine and e-commerce applications. In the web sector, it is mainly social media platforms and news publishers that use A/B testing methods [21]. For example, Hagar and Diakopoulos [22] conducted an interview study examining how newsrooms use A/B testing to select optimal headlines and increase traffic to articles. Other examples include the Wikipedia Foundation, which uses A/B testing to optimise a wide range of aspects [23–25].

2.4. Sampling and Error Correction

The results of statistical testing methods are highly dependent on the quality of the underlying data and the sampling technique used. Errors in the data or inadequate sampling procedures can lead to inaccuracies in the test results, requiring the application of statistical correction methods.

Sampling error is a major source of error in statistical testing methods. As defined by Milanzi et al., sampling error is "generally defined as the difference between the actual value of the population characteristic and an estimate obtained from a sample. This estimate is generally not equal to the true value of the characteristic because of sampling variability [. . .] and bias" [26]. To reduce sampling error, advanced sampling techniques such as stratified sampling are often used. Stratified sampling involves dividing a population into smaller, homogeneous groups called strata. These strata are organised on the basis of characteristics or attributes shared by members of the population [27]. This division helps to prevent the inclusion of extreme samples that may skew the results [28]. In test design, each stratum of a stratified random sample is usually modelled separately to ensure accurate representation. For example, in surveys, strata can be defined based on demographic characteristics such as age, and the sample size for each stratum is determined independently of the survey according to the corresponding age group of the population. An alternative approach to stratification has been proposed by Liberty et al. They use machine learning and regression analysis to address the problem of stratification design [29].

Another effective strategy for reducing sampling error is the use of auxiliary information. Bethlehem notes that auxiliary information can improve both the sampling design and the estimation procedure itself [27]. Bethlehem goes on to provide a comprehensive overview of survey methods, including sampling design, estimators, and the use of auxiliary information to reduce error and bias. Early studies by Raiffa and Schlaifer [30] and Ericson [31] explored the use of auxiliary information in stratified sample surveys. More sophisticated approaches include the use of auxiliary information for two-stage sampling [32] and for determining an optimal compromise allocation of sampling units in multivariate stratified surveys [33]. Building on these foundations, Khan et al. [34], Varshney et al. [35] and Gupta et al. [36] extended the use of auxiliary information to obtain integer optimal solutions. In addition, Deville and Särndal [37] proposed calibration estimators in survey sampling, using auxiliary information to improve the estimation of population statistics. In subsequent work, Singh et al. [38] proposed a calibration approach for improved variance estimators in survey sampling, while Kim et al. [39] proposed various ratio estimators in the calibration approach and Wu and Sitter [40] used auxiliary information in a model calibration approach.

3. A/B Test Setup

The new approach is applied to a test configuration that corresponds to the classic A/B test with experts. In order to control as many factors as possible in the new approach, an A/B test that has already been successfully performed, documented and published in a previous project will be repeated: a comparison of an ATC software (4D-NAVSIM, version 2023; VAST, version 4.14 based on Unity 2019) user interface in 2D and in 3D [41,42]. The setup consists of a prototype, the result of previous efforts [41,43–45], coupled with an existing air traffic simulator [46], which enables realistic air traffic control simulations.

The test involved 28 participants, including eight ATCs (one female, seven male) and twenty students (seven female, thirteen male) with experience in 3D video games. The ATCs work at an international, Austrian Airport, while the students were enrolled in media technology or computer science programmes at the University of Applied Sciences St. Pölten and Graz University of Technology respectively.

3.1. Test Setup and Protocol

The test setup and protocol closely follow those of the previous "Virtual Airspace and Tower (VAST)" project [42]. The tests were conducted in dedicated environments, with ATCs being tested in Salzburg and students being tested at their respective universities. To facilitate a smooth experimental scenario, the test setup consisted of a PC with a powerful GPU, a 4K monitor for the prototype, and standard peripherals. In addition, the air traffic simulator (ATS) ran on separate hardware, and interaction with the traffic simulator was facilitated by voice control via a headset with a microphone.

After a general introduction to the test setting, participants completed a newly developed psychological questionnaire, which was later used as auxiliary information for statistical correction. In a training phase, participants were then free to explore the prototype. Subsequently, as in Rottermanner et al. [42], two test scenarios—Task 1 (2D) and Task 2 (3D)—were performed for 20 min each, with participants using voice control to manage air traffic. The objectives mirrored those of Rottermanner et al. [42], focusing on efficient and safe aircraft landing with a test scenario based on data from Frankfurt airport. As all ATC participants work at an Austrian airport, Frankfurt Airport ensures that all participants are confronted with an unknown air traffic control scenario and environment.

The tasks also remained unchanged; i.e., in Task 1, the 2D task, participants were restricted to an aerial (bird's eye) view of air traffic, while in Task 2, the 3D task, participants were allowed to adjust the viewing angle within a specified range, excluding the aerial option. As in Rottermanner et al. [42], the NASA Task Load Index (NASA TLX) [47] to assess workload and the Situational Awareness for SHAPE questionnaire (SASHA_Q) [8,48] to assess situational awareness were completed by the participants after each task.

3.2. Flight Data

Similar to VAST, the test used real-time flight data from Frankfurt Airport to ensure that participants were exposed to a complex and realistic air traffic control scenario. The data, recorded over one day, included departing and arriving air traffic and were used at four different start times for different scenarios. One scenario was used for training, two were used for the test tasks and one was used as a backup, with all scenarios falling within the 12 pm (noon) to 2 pm time window. This approach prevented participants from anticipating flight behaviour in subsequent tasks [42].

3.3. Performance Measures

During each task, several performance measures were tracked, including the number of aircraft taken over, the time to take over, the number of landings, the deviations from simulation-based optimised routes and landing times, the altitude and distance of unlanded aircraft, the conflicts and the instructions given. These measures were combined to create task-related key performance indicators (KPIs) for each participant. As the simulated ATS traffic was taken as the optimal case, the subjects' performance measures were related to the simulated performance of the ATS. Table 1 lists all key performance indicators.

Table 1. These key performance indicators were used to assess the performance of participants within the test scenarios and were further integrated into the transformation model to establish a mapping between ATCs and students.

	KPI	Description
#1	Taken over (#)	Number of planes taken over by the test subject
#2	Taken over (%)	Percentage of optimal number of taken-over planes
#3	Time until takeover total (mm:ss)	Duration from the radio message from the aircraft to acceptance by the test subject summed across all planes
#4	Time until takeover/plane (mm:ss)	Duration from the radio message from the aircraft to acceptance by the test subject per plane
#5	Landings 1 (#)	Number of planes landed by the test subject
#6	Landings 2 (#)	Number of non-landed planes already in position to land with distance to the runway $<$ 10 km and height $<$ 1000 ft
#7	Landings 3 (#)	Number of non-landed planes already in position to land with distance to the runway $<$ 10 km and height $<$ 5000 ft
#8	Calculated Landings (#)	Number of planes landed by the test subject plus planes close to landing (Landings 2 and Landings 3); calculated via Landings 1 $+\frac{1}{2}$ Landings 2 $+\frac{1}{4}$ Landings 3
#9	Optimum Landings (%)	Percentage of optimum of landed planes
#10	Calculated Optimum Landings (%)	Percentage of optimum of calculated landings
#11	Time deviation to landing total (mm:ss)	Total deviation from the simulated landing times of the ATS
#12	Time deviation to landing/plane (mm:ss)	Deviation per plane from the simulated landing time of the ATS
#13	Distance deviation to landing total (km)	Total deviation from the simulated routes of the ATS
#14	Distance deviation to landing/plane (km)	Deviation per plane from the simulated route of the ATS
#15	Height not landed total (ft)	Total height of the non-landed planes
#16	Height not landed/plane (ft)	Average height per plane of the non-landed planes
#17	Distance not landed total (km)	Total distance of the non-landed planes to the runway
#18	Distance not landed/plane (km)	Average distance per plane of the non-landed planes to the runway
#19	Distance not landed/plane (%)	Average distance per plane of the non-landed planes to the runway in relation to the ATS simulation
#20	Conflicts (#)	Number of losses of separation
#21	Instructions/plane (#)	Number of instructions given by the test subject per plane
#22	Instructions total (#)	Total number of instructions given by the test subject
#23	NASA TLX Average ([0, 100])	Average of NASA TLX results
#24	NASA TLX Average (%)	Percentage of optimal NASA TLX score
#25	SASHA_Q Average ([1, 5])	Average of SASHA_Q results
#26	SASHA_Q Average (%)	Percentage of optimal SASHA_Q score

4. Statistical Error Correction

The basic idea of the new approach is to deliberately introduce a systematic statistical error into the study and then correct it. Under normal circumstances, it is not a good idea to conduct a user test with the wrong target group. However, if the target group is difficult to reach, it may make sense—not for statistical reasons, but for economic, organisational or other reasons—to deliberately introduce this error and then correct it.

The essence of this study is to involve non-domain individuals in the process of testing expert software. To achieve this, the non-domain individuals need to be mapped into the domain of the domain experts. By using auxiliary information, the approach aims to minimise the introduced error of testing expert software with non-domain individuals.

The approach can be easily illustrated for better understanding. Figure 2 provides a visual representation of the main idea of the approach. Basically, the approach aims to construct a model that facilitates the transfer of test results from non-domain experts to domain experts by using auxiliary information. In Figure 2, domain experts are denoted as ATC_i and non-domain individuals are denoted as S_j. Both non-domain individuals and experts are assessed using a single task (Task 1) and a psychological questionnaire that serves as auxiliary information. A linear model is then developed to establish the relationship between the Task 1 results and the auxiliary information of each domain expert (ATC_i) on the one hand and the Task 1 results and the auxiliary information of all non-domain individuals on the other hand. This model consists of a weight vector for each expert; each vector contains the weights to optimally represent an individual expert by non-experts in terms of a linear regression model. Consequently, the model can be applied to the Task 2 performances of the non-experts to predict the Task 2 KPIs of the experts.

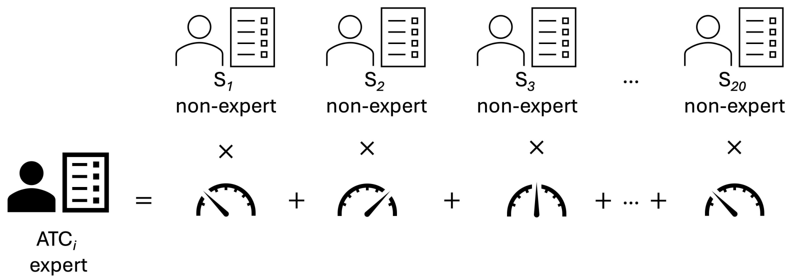


Figure 2. The main part of the transformation model is a mathematical representation of each expert (resp. the expert's KPIs) by a weighted sum of non-experts (resp. their KPIs).

In an actual application scenario, the tests would now be completed (and the controller testing effort saved for Task 2), but in order to not only statistically prove but also clearly demonstrate the accuracy of the predictions, the controller test results are also recorded in Task 2 and compared with the model predictions.

In summary, Task 1 scores are used in conjunction with auxiliary information to create a linear mapping model from non-domain individuals to the expert. The Task 2 scores of the non-domain individuals, together with their auxiliary information, are then used to predict the Task 2 scores of each domain expert.

4.1. Auxiliary Information

In this new approach, auxiliary information is used to counteract the introduced systematic error in the mapping of non-experts to experts. A psychological questionnaire is used as the auxiliary information. In order to create the most suitable psychological questionnaire for providing auxiliary information in the novel mapping process, a series of workshops were conducted with psychologists to define the requirements of the auxiliary information.

A number of characteristics were considered essential to the mapping process for the auxiliary information questionnaire:

1. For procedural reasons, the questionnaire should not provide free text fields for responses but should only allow responses on a numerical scale or be directly mappable to such a scale. Furthermore, as the questionnaire was to be included in a user test, it was imperative that the test could be completed within a limited time (in this case 45 min).
2. The test had to cover a wide range of ATM or ATM-related topics without being too specific, as it was intended to be auxiliary information. If the test was too specific (e.g., a question that all ATCs answered in the same way), the information value of the question would be low; if all non-experts also answered in the same way, the information value would be non-existent. From a statistical point of view, the answers to the questions should ideally have a normal distribution for both the experts and the non-experts. The additional information is not used to select study participants who match the requirement profile of air traffic controllers as closely as possible; participants with a negative correlation to the requirement profile (laypersons who, in extreme cases, do the opposite of professionals) also provide valuable information.
3. Psychological interpretation of the psychological test results was not required for the purposes of this study; i.e., it did not have to be a validated psychological test. The aim is not to create personality or character profiles, and although the tests are conducted anonymously, the questionnaire should not contain any questions that could be ethically or legally problematic.
4. Aspects already covered by the KPIs, in particular the workload and situational awareness questionnaires used, should not be included in this psychological test.

Following several sessions with multiple psychologists, a consensus was reached. The final questionnaire emphasizes various aspects crucial for successful performance in the ATC profession. These encompass personality traits, such as decisiveness, responsibility and teamwork skills, as well as stress management and processing, concentration, cognitive abilities, intelligence and work ethic [49,50]. The questionnaire consisted of 75 questions. Each question was tailored to focus on specific aspects. Questions focusing on the personality traits aspect are based on the Big Five model [51], which includes the five dimensions: surgency, agreeableness, conscientiousness, emotional stability and intellect. For example, questions #1 “I tend to be spontaneous.”, #25 “I have a passion for collecting.” and #32 “I love rituals.” are taken from the psychological questionnaires in the categories of personality traits (#1), stress management and processing (#25) and work ethic (#32). In addition to cognitive and perceptual skills, there are questions designed to assess concentration. The entire questionnaire can be found in Appendix A. It also comprehends two tests (see Appendices A.2 and A.3). As each test is weighted in the same way as each of the 75 questions, the two tests play a minor role. Since the influence of the tests (as well as the individual questions) is an open research question, we opted for more questions and fewer tests due to the time constraints of the complete A/B test setup.

4.2. Transformation Model

As illustrated in Figure 2, each participant is represented by task scores combined with auxiliary information; Specifically, the data for each participant consisted of 26 KPIs, 6 NASA TLX scores, and 8 SASHA_Q scores. The auxiliary information included 77 scores, of which 75 scores were from the psychological questionnaire and 2 scores were from the additional psychological tests focused on assessing concentration, cognitive and perceptual abilities. Combining the task results and the auxiliary information resulted in 117 values per participant. An overview of how the samples are split into the respective components is given in Table 2.

Due to the different ranges of the KPIs and questionnaire responses, normalisation was required. All 117 samples were normalised to the interval between zero and one using the equation

$$x_{norm} = \frac{x - \min}{\max - \min}. \tag{1}$$

Table 2. Each test participant and the corresponding test results consist of 117 values. This table shows how they are allocated to the different components of the test.

Component	Number of Values
KPIs	26
NASA TLX questionnaire	6
SASHA_Q questionnaire	8
Psychological questionnaire (auxiliary information)	77

For continuous variables, such as the KPIs, min and max refer to the minimum and maximum across all tasks and subjects for the specific variable. For discrete variables, such as the questions of the psychological questionnaire, the NASA TLX or the SASHA_Q questionnaire, min and max refer to the minimum and maximum allowed values for the questionnaire. In addition, continuous variables were padded by 10% of their respective min-max range.

The model itself is based on multiple linear regression (MLR) that is carried out with $p = 19$ independent variables; one independent variable per student, with one student removed due to incomplete test results. If Y_i is the score vector of the ATC i (with 117 dimensions as listed in Table 2) and X_j is the score vector of the non-expert student j, then the MLR model consists of the weights $\beta_{i,j}$ and the errors ε_i according to the equation

$$Y_i = \sum_{j=1}^{19} \beta_{i,j} \cdot X_j + \varepsilon_i \tag{2}$$

In general, Equation (2) cannot be solved because it is overdetermined. This is exactly the purpose of auxiliary information. Instead of an exact solution, which is not desirable for numerical reasons and not expected for modelling reasons, a least squares approximation is used. Normal equations and Cholesky decomposition give least squares estimates for the student weights $\hat{\beta}_{i_j}$ and the offsets $\hat{\varepsilon}_i$, ($i = 1, \ldots, 8$ and $j = 1, \ldots, 19$).

The predictions are now calculated by multiplying the results of Task 2 of the non-expert students by the previously calculated weights and adding them together to predict the results of each individual expert.

As the predictions are calculated on normalised data and are therefore in normalised form, denormalisation must be applied. Denormalisation is the reverse process of normalisation and is achieved with the following equation:

$$x_{denorm} = x_{norm} \cdot (\max - \min) + \min, \tag{3}$$

where min and max are the same minima and maxima used in the normalisation process.

The quality of fit of the standard MLR models is assessed by the coefficient of determination R^2. This coefficient, introduced by Wright [52], generally indicates how well the regression model explains the data. More specifically, R^2 can be interpreted as the proportion of variance in the data that is explained by the regression model. Thus, an R^2 value of 0.75 would indicate that 75% of the variance in the data can be explained by the regression model.

The entire transformation model can be evaluated using the quality of fit using the coefficient of determination; for predictions based on such a model, confidence intervals are provided by Olive [53]: the $100(1-\delta)\%$ confidence interval for a prediction $\hat{y}_i$ is calculated via

$$\hat{y}_i \pm t_{n-p-1,1-\frac{\delta}{2}} \sigma^2 \sqrt{1 + x_i^{\mathrm{T}} (X^{\mathrm{T}} X)^{-1} x_i} \tag{4}$$

using the t-distribution, the estimated variance σ^2 of the errors ε_i, and the input values x_j.

5. Results

To illustrate and demonstrate the new approach, we repeated an A/B test of an earlier user study involving air traffic controllers.

5.1. "Virtual Airspace and Tower"

In the specific example of repeating the user interface A/B test from the previous "Virtual Airspace and Tower (VAST)" project [42], the application of the new method is as follows: Task 1 and the psychological test (auxiliary information) were completed by both the expert ATCs and the non-expert students. After the values were normalised, the model parameters were determined using the normal equation and the Cholesky decomposition. Table 3 shows the model parameters. This table also includes statistics such as the minimum (min), maximum (max), mean, standard deviation (std.-dev.), and variance of the weights ($\hat{\varepsilon}$, $\hat{\beta}_1$, $\hat{\beta}_2$, ..., $\hat{\beta}_{19}$) for each model.

Table 3. The least squares estimates $\hat{\varepsilon}$, $\hat{\beta}_1$, ..., $\hat{\beta}_{19}$ represent the multiple linear regression (MLR) models to represent the results of experts by the results of non-experts.

Model	ATC 1	ATC 2	ATC 3	ATC 4	ATC 5	ATC 6	ATC 7	ATC 8
$\hat{\varepsilon}$	0.1354	0.1043	0.1536	0.2074	0.3021	0.2366	0.1571	0.1873
$\hat{\beta}_1$	−0.2409	−0.0143	−0.1072	−0.1475	−0.0673	0.0317	−0.0244	-0.0591
$\hat{\beta}_2$	0.2460	0.1327	0.3101	−0.1876	−0.1702	0.1251	0.1803	−0.0831
$\hat{\beta}_3$	0.3458	−0.0361	0.0436	−0.0904	−0.0125	0.0149	0.1451	0.0362
$\hat{\beta}_4$	0.0775	−0.1603	0.0008	0.2113	−0.0302	−0.2141	−0.0226	0.0626
$\hat{\beta}_5$	−0.0026	0.0228	−0.0076	−0.0375	−0.0083	0.0658	0.0059	0.1014
$\hat{\beta}_6$	0.1586	0.0884	0.1352	0.1280	0.0482	0.2448	0.0550	0.0855
$\hat{\beta}_7$	0.0254	0.0143	−0.1737	−0.2004	−0.1000	0.0210	−0.1031	0.0233
$\hat{\beta}_8$	−0.0843	−0.0230	−0.1020	0.2365	−0.1196	−0.0099	−0.1451	0.1066
$\hat{\beta}_9$	−0.1471	−0.0164	0.0134	0.0755	0.1821	−0.0203	−0.0790	−0.0756
$\hat{\beta}_{10}$	−0.3673	−0.2102	−0.2407	−0.2007	−0.4335	−0.3134	−0.3361	−0.2876
$\hat{\beta}_{11}$	−0.0477	−0.0983	0.0192	−0.1371	−0.0023	−0.0976	−0.1815	−0.1058
$\hat{\beta}_{12}$	0.1427	0.0039	0.1100	0.1477	−0.0613	0.2395	0.1955	0.2445
$\hat{\beta}_{13}$	0.0050	0.1125	−0.0561	−0.0608	0.2664	−0.0121	−0.0247	−0.0120
$\hat{\beta}_{14}$	0.2883	0.3799	0.2121	0.0107	0.3367	0.1448	0.4769	0.0188
$\hat{\beta}_{15}$	0.1091	−0.0761	0.0741	−0.0498	0.1446	−0.1154	−0.1304	0.0282
$\hat{\beta}_{16}$	−0.1307	0.0401	0.0447	0.2868	−0.0672	0.0309	−0.0749	0.1060
$\hat{\beta}_{17}$	0.2900	0.3036	0.3321	0.4234	0.2471	0.4114	0.4032	0.3903
$\hat{\beta}_{18}$	−0.0467	0.0720	−0.1378	0.2000	0.2054	−0.0058	0.1979	0.0681
$\hat{\beta}_{19}$	0.1927	0.2922	0.2627	0.1321	0.0921	0.0819	0.1585	0.0749
min	−0.3673	−0.2102	−0.2407	−0.2007	−0.4335	−0.3134	−0.3361	−0.2876
max	0.3458	0.3799	0.3321	0.4234	0.3367	0.4114	0.4769	0.3903
mean	0.0428	0.0435	0.0385	0.0389	0.0236	0.0327	0.0366	0.0380
std.-dev.	0.1859	0.1485	0.1564	0.1787	0.1772	0.1596	0.1962	0.1361
variance	0.0345	0.0220	0.0244	0.0319	0.0314	0.0254	0.0385	0.0185

Inspection of the Table 3 reveals a visually uniform distribution of weights in the range $[-0.5,\ 0.5]$ with no gross outliers, although no range has been enforced by any constraints. The minimum weight, $\hat{\beta}_{10} = -0.43356191$, corresponds to ATC 5, while the maximum weight, $\hat{\beta}_{14} = 0.47690763$, belongs to ATC 7. Since the selection of non-experts is not limited to people who are as similar as possible to the experts, negative weights also occur. This may lead to invalid values in the prediction and extrapolation of future test results, but it does not restrict the selection of non-experts in any way: an advantage that may justify a possible extrapolation error that does not necessarily occur. If this is not desired, non-experts with negative coefficients—such as $\hat{\beta}_{10}$—should be removed.

In statistics, the coefficient of determination R^2 is used to determine the quality of fit of a model. Specifically, R^2 is the proportion of variation in the dependent variable that can be predicted by the independent variables. In this way, it provides a measure of how well the observed results are replicated by the model, based on the proportion of total variation in the outcomes explained by the model. Table 4 shows how well each ATC's test results can be described by the model of non-experts.

Pearson's correlation coefficients are calculated between the dependent variable y and the independent variables ($x_1,\ x_2,\ \ldots,\ x_{19}$), denoted as $r_{y,x_1},\ r_{y,x_2},\ \ldots,\ r_{y,x_{19}}$. In addition, the correlations between the independent variables themselves are calculated ($r_{x_1,x_2}, r_{x_1,x_3}, \ldots, r_{x_{18},x_{19}}$). The correlation matrix illustrates these coefficients (see Figure 3), where the first column of the correlation matrix shows the correlations between the dependent variable y and each independent variable, while the remaining columns show the Pearson correlation coefficients between all independent variables.

Table 4. The coefficients of determination R^2 and R^2_{adj} can be interpreted as the proportion of variance in the data that is explained by the regression model. The adjusted R^2_{adj} takes the model size into account; the not-adjusted coefficient of determination R^2 automatically increases when additional variables are added to the model.

Coefficient of Determination	ATC 1	ATC 2	ATC 3	ATC 4	ATC 5	ATC 6	ATC 7	ATC 8
R^2	0.6136	0.6555	0.5458	0.4499	0.5154	0.5124	0.5397	0.5546
R^2_{adj}	0.5379	0.5880	0.4569	0.3421	0.4205	0.4169	0.4495	0.4673

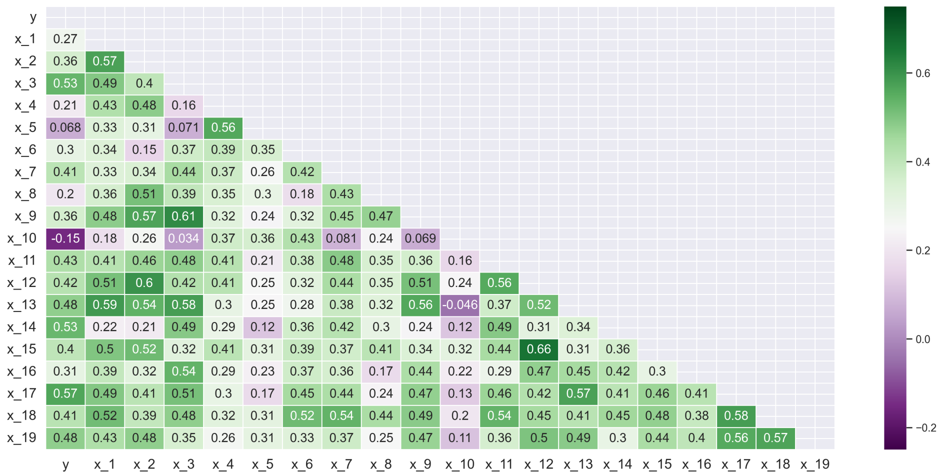


Figure 3. This matrix shows the Pearson's correlation coefficients between dependent and independent variables.

In Figure 3, the highest correlation between the dependent variable y and the independent variables can be seen for x_{17} with $r_{y,x_{17}} = 0.57$. Furthermore, x_3 and x_{14} have correlations with the dependent variable greater than 0.5. Notably, x_{10} is the only independent variable that has a negative correlation with y as $r_{y,x_{10}} = -0.15$. Among the independent variables, the highest correlation coefficient is observed between x_{12} and x_{15} with $r_{x_{12},x_{15}} = 0.66$. Other independent variables with correlation coefficients greater than 0.6 include $r_{x_2,x_{12}} = 0.6$ and $r_{x_3,x_9} = 0.61$; the only negative correlation is observed between x_{10} and x_{13} with a value of $r_{x_{10},x_{13}} = -0.046$. The five smallest correlations in absolute terms are (in decreasing order) $r_{x_3,x_5} = 0.071$, $r_{x_9,x_{10}} = 0.069$, $r_{y,x_5} = 0.068$, $r_{x_{10},x_{13}} = -0.046$, and $r_{x_3,x_{10}} = 0.034$.

The model listed in Table 3 is used to transform the results of Task 2 from the non-expert students to the expert ATCs.

5.2. Transformation Results

The results of the transformation are summarised and listed in Table 5. To illustrate the quality of the transformation, the ATCs also performed Task 2 (observation), and these averaged results are compared with the averaged predictions using the transformation model (prediction) including and excluding the correction using auxiliary information. To facilitate comparison between the KPIs, the relative errors of the normalised values (according to Equation (1)) are also given. As the relative errors depend on the size of the range interval, i.e., the minimum and maximum values of all test results by ATCs and non-ATCs, the listed percentages are sensitive to outliers. Nevertheless, it makes sense to normalise the data in order to be able to compare the error values of the individual categories, which can differ by orders of magnitude.

Table 5. The transformation model uses the non-expert (student) results to predict the expert (ATC) results. Compared to the real test results of the experts in Task 2, the transformation model achieves an accuracy with a relative error of less than 1% in 1 out of 26 KPIs, a relative error between 1% and 5% in 9 out of 26 KPIs, a relative error between 5% and 10% in 5 out of 26 KPIs and a relative error greater than 10% in 11 out of 26 KPIs. The main concept of the transformation model is based on auxiliary information. To illustrate its power, the transformation results based on a linear model without auxiliary information have been included as well.

KPI	Observation	Without Aux. Info.		With Aux. Info.		Improvement
		Prediction	Error	Prediction	Error	
Taken over (#)	9.750	10.597	14.2%	9.565	3.1%	+11.1%
Taken over (%)	0.886	0.963	14.2%	0.870	3.1%	+11.1%
Time until takeover total (mm:ss)	172.625	−275.687	19.9%	521.359	15.5%	+4.4%
Time until takeover/plane (mm:ss)	17.750	−39.838	21.4%	59.750	15.6%	+5.8%
Landings 1 (#)	4.000	5.573	32.5%	3.874	2.6%	+29.8%
Landings 2 (#)	0.500	1.648	95.8%	0.796	24.7%	+71.1%
Landings 3 (#)	1.625	2.114	13.6%	1.232	10.9%	+2.7%
Calculated Landings (#)	4.656	6.830	37.7%	4.483	3.0%	+34.7%
Optimum Landings (%)	0.80	1.115	32.4%	0.775	2.6%	+29.8%
Calculated Optimum Landings (%)	0.776	1.155	34.6%	0.764	1.1%	+33.5%
Time deviation to landing total (mm:ss)	−73.875	−17.069	6.3%	14.316	9.8%	−3.5%
Time deviation to landing/plane (mm:ss)	−10.250	15.990	6.9%	2.711	3.4%	+3.5%
Distance deviation to landing total (km)	4.929	4.545	0.3%	8.494	3.0%	−2.7%
Distance deviation to landing/plane (km)	2.392	0.857	4.0%	2.122	0.7%	+3.3%
Height not landed total (ft)	46,901.500	57,265.587	25.6%	50,987.712	10.1%	+15.5%
Height not landed/plane (ft)	6671.031	9012.775	52.6%	7111.841	9.9%	+42.7%
Distance not landed total (km)	132.568	156.490	10.3%	171.081	16.6%	−6.3%
Distance not landed/plane (km)	18.904	25.167	29.7%	23.018	19.5%	+10.2%
Distance not landed/plane (%)	0.868	0.761	11.9%	0.835	3.7%	+8.2%
Conflicts (#)	0.375	−2.346	28.4%	1.315	9.8%	+18.6%
Instructions/plane (#)	5.924	4.149	28.4%	4.460	23.4%	+5.0%
Instructions total (#)	57.250	46.359	20.6%	42.990	27.0%	−6.4%
NASA TLX Average ([0, 100])	37.396	34.796	2.8%	54.647	18.5%	−15.7%
NASA TLX Average (%)	0.626	0.723	2.8%	0.521	11.2%	−15.7%
mental	56.562	20.909	35.8%	65.717	9.2%	+26.6%
physical	31.250	85.875	54.6%	51.462	20.2%	+34.4%
temporal	37.188	37.900	0.7%	66.266	29.1%	−28.4%
performance	27.500	41.185	13.7%	44.414	16.9%	−3.2%
effort	47.188	29.451	17.7%	71.225	24.0%	−6.3%
frustration	24.688	1.848	22.8%	37.308	12.6%	+10.2%
SASHA Q Average ([0, 5])	3.438	4.012	43.0%	3.507	5.2%	+37.8%
SASHA Q Average (%)	0.688	0.802	43.0%	0.701	5.2%	+37.8%
manageable	4.750	6.664	38.3%	3.531	24.4%	+13.9%
next steps	4.625	6.562	38.8%	3.768	17.2%	+21.6%
heavy focus	2.125	2.174	1.0%	3.052	18.5%	−17.5%
find info	2.500	−1.420	78.4%	1.116	27.7%	+50.8%
valuable info	3.375	5.508	42.5%	3.771	7.9%	+34.6%
attention	3.000	3.940	18.8%	4.179	23.6%	−4.8%
understanding	3.500	3.431	1.3%	4.172	13.4%	−12.0%
awareness	3.625	3.866	4.8%	2.979	12.9%	−8.1%

The transformation model deliberately allows for negative coefficients (see Table 3); if all non-experts with negative weights had been removed (as discussed above), the number of subjects would have been significantly reduced. Only 5 of the 19 non-experts have consistently positive weights. As already mentioned, this increases the likelihood of semantically unreasonable values in the extrapolation/prediction (e.g., a negative prediction when in reality only semi-positive values are meaningful and possible). Nevertheless, the transformation model is convincing. It shows improvements over models without auxiliary information. The improvement column lists the average improvement (reduction in errors) in percentage points of the relative errors through the use of auxiliary information. The use of auxiliary information improves the prediction results by reducing the error by 12% on average.

In the intended application scenario of the transformation model—replacing unavailable or difficult-to-reach air traffic controllers in the test with an alternative target group for cost and/or organisational reasons—the real observations are not known. The proposed interpretation of an A/B test prediction can be based on the confidence intervals (see Equation (4)): In an A/B test setting, the relevant question is whether version A or version

B is better. If the test results (KPIs) of the ATCs in Task 1 t_1 and their prediction for Task 2 t_2 differ, the confidence interval $t_2 \pm conf(\delta)$ can be determined depending on the confidence level δ in such a way that a separation $t_1 \notin t_2 \pm conf(\delta)$ with maximum delta is ensured. This view allows the test question to be answered in terms of how confident you can be that one version (A or B) is better than the other and that the test result is not random. Such a representation is shown in the appendix in Tables A1 and A2.

The results presented in "Design and Evaluation of a Tool to Support Air Traffic Control with 2D and 3D Visualizations" [42] could not be reproduced completely; in this repeated study, the A/B Test showed significant differences between Task 1 (2D) and Task 2 (3D) according to Mann–Whitney-U-tests only for

- Distance not landed/plane % [$U = 59$, p-value = 0.003],
- Distance not landed total (km) [$U = 7$, p-value = 0.007],
- Distance not landed/plane (km) [$U = 8$, p-value = 0.010].

Inspecting the KPIs in the Tables A1 and A2 reveals four KPIs showing high-confidence percentages across all models, namely "Landings 2", "Distance not landed total (km)", "Distance not landed/plane (km)" and "Distance not landed/plane (%)".

Unfortunately, this study suffers from the same problem that it seeks to solve: the statistical tests could not be carried out to the necessary extent with ATC subjects. Despite the severe limitation of having only eight ATC participants, the transfer model was able to show that the essential statements of the A/B test could be generated with the non-expert students.

6. Conclusions

The aim of this new approach was to test the feasibility of involving non-experts in the evaluation process of expert software, focusing specifically on whether a transformation model could be constructed to predict test results for ATCs using students' test results. Using auxiliary information in the form of a newly developed psychological questionnaire, we constructed a novel transformation model from the students' Task 1 results to the Task 1 results of each ATC. We then predicted Task 2 results for each ATC based on the students' Task 2 results.

Using multiple linear regression to create the transformation model, we achieved accurate predictions for the majority of the defined KPIs for Task 2 for the ATCs using the students' Task 2 performance. In other words, the first research question, whether it is possible to perform a meaningful user test without the relevant user group, can be answered in the affirmative. The errors of the averaged predictions were generally small, with the majority of KPIs showing errors of less than 10% and all KPIs showing errors of less than 30%. The examination of the quality of fit revealed coefficients of determination between 45% and 66%. On average, the coefficients of determination resulted in 54.8% of the variance in the dependent variable being accounted for by the independent variables, underlining the predictive power of the approach. This example answers the second research question about the expected errors.

The selection of the questionnaire remains an open question; to the best of our knowledge, we suspect that the questionnaire is only dependent on the field of application (air traffic management). This is an example of constructive error correction for error minimization. However, further research is needed to confirm this hypothesis. Furthermore, the number of auxiliary questions is an open research question. On the one hand, a comparison of models with and without auxiliary information indicates that the prediction improves when some auxiliary information is used. In our example, the prediction improved by 12% on average (see Table 5). On the other hand, the auxiliary information and the KPIs to be predicted are part of the same transformation model. As the number of auxiliary questions increases, the impact of the KPIs on the transformation will diminish, potentially reducing the prediction accuracy. The optimal number of additional questions and tests is unknown and remains an open research question. Furthermore, all questions and tests are used with a uniform weight in the transformation model, despite the possibility that some questions may be more important than others. It is also unclear which questions are the most important ones.

A notable restriction of this study was the limited size of the test pool. Ideally, the proposed approach would be validated with a larger pool involving more ATCs and students. However, this is limited by the availability of ATCs for testing purposes—the very limitation that this approach aims to alleviate. Even if the error cannot be avoided, it can at least be limited by confidence intervals; i.e., you do not have to blindly trust the transformation model. This answers the third research question about error quantification.

In summary, our results highlight the potential of the presented approach to improve the evaluation process of expert software by involving non-experts in the testing phase. By developing and validating a novel transformation approach that incorporates auxiliary information from a newly developed psychological questionnaire, we have demonstrated the ability to predict the performance of ATCs based on students' test scores. The approach allows testing with non-experts, while ATCs are only needed at the beginning to build the transformation model. However, as shown in Figure 1, we recommend involving experts in testing at key milestones and, at the end of the software development process, validating the end result.

The new approach not only avoids the challenge of obtaining a sufficient number of ATCs for testing but also increases the frequency of testing while ensuring that a wider range of perspectives are incorporated into the evaluation process. With this approach, more tests can be performed for the same financial value, resulting in better-tested and more user-centred software, in line with the push for user-centred design by EuroControl [3].

Author Contributions: Conceptualization, I.-S.S. and T.U.; Data curation, P.S., P.J., G.R. and V.S.; Formal analysis, T.U.; Funding acquisition, T.U.; Investigation, P.S., P.J., G.R., C.-H.R., V.S. and T.U.; Methodology, T.U.; Project administration, V.S.; Resources, P.S., P.J., G.R., C.-H.R., I.-S.S. and V.S.; Software, P.S., P.J., G.R., C.-H.R. and V.S.; Supervision, T.U.; Validation, P.S., P.J., G.R., C.-H.R., I.-S.S., V.S. and T.U.; Visualization, P.S., G.R., C.-H.R., V.S. and T.U.; Writing—original draft, P.S., P.J., G.R., C.-H.R., I.-S.S., V.S. and T.U.; Writing—review and editing, P.S., P.J., G.R., C.-H.R., I.-S.S., V.S. and T.U. All authors have read and agreed to the published version of the manuscript.

Funding: This article is supported by TU Graz Open Access Publishing Fund. The presented research results are based on the previous research projects "Virtual Airspace and Tower" (VAST), "Sondierung eines Prototyping & Evaluation Frameworks für (teil-) automatisierte Air Traffic Control Software" (EMMSA), and "Cross Level User Evaluation" (CLUE).

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Acknowledgments: Open Access Funding by the Graz University of Technology.

Conflicts of Interest: Authors Phillip Stranger, Volker Settgast and Torsten Ullrich were employed by the company Fraunhofer Austria Research GmbH, Peter Judmaier and Gernot Rottermanner were employed by the company Fachhochschule St. Pölten Forschungs GmbH, Carl-Herbert Rokitansky was employed by the company 4D Aerospace Research and Simulation GmbH. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A. Auxiliary Information

Appendix A.1. Questionnaire

The following questions were presented to the test participants with the answer options of agreement and disagreement according to the scale:

1. not true at all
2. do not agree
3. rather disagree
4. disagree a little
5. neither

6. somewhat agree
7. rather agree
8. quite true
9. very true
10. completely agree

The original questionnaire was written in the German language; the following questions are a translation that is as close to the original as possible:

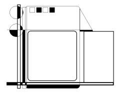
1. I tend to be spontaneous.
2. I enjoy getting to know other people.
3. I enjoy giving presentations to large groups.
4. I prefer to create a cosy seclusion at home rather than going out and socialising.
5. I am an optimist.
6. In my work, I try to plan ahead as much as possible.
7. I face challenges with optimism.
8. I prefer to solve problems at work independently rather than as part of a team.
9. My favourite job is one where I can take on a high level of responsibility.
10. I really enjoy monotonous professional activities.
11. I usually make my decisions impulsively and on instinct.
12. I adapt my work activities immediately according to the situation at hand.
13. I am easily persuaded by others.
14. I like to be the centre of attention.
15. I always work purposefully to achieve my work results.
16. To cope with more difficult tasks, I seek the approval of a colleague to be on the safe side.
17. Mastering an unfamiliar professional task causes me discomfort and anxiety.
18. If necessary, I can assign clear tasks in a work context.
19. I can concentrate on monotonous tasks over a longer period of time.
20. I am depressed after challenging tasks at work.
21. I am able to communicate easily in stressful situations.
22. A stressful job is unimaginable for me.
23. People who have achieved more professionally than I have are enviable.
24. I am very resilient in my job.
25. I have a passion for collecting.
26. It makes me uncomfortable if I don't have a situation under control.
27. I'm good with numbers.
28. I can relax after strenuous activities with exercise.
29. I find some traffic rules nonsensical.
30. I don't follow rules that don't make sense to me in certain life situations.
31. Standardised work processes are important to me.
32. I love rituals.
33. I see stressful situations as a kind of obstacle for me.
34. I see challenging situations as an opportunity.
35. My abilities unfold in situations that trigger stress.
36. Complex work situations should be dealt with as part of a team.
37. I am able to recognise patterns and structures in certain situations or activities where others do not see them.
38. I relax when I do sport.
39. Music is a form of relaxation for me.
40. I have to work to earn a living, but I wouldn't do it if I didn't have to.
41. I enjoy learning something new.
42. I take regular breaks from strenuous activities.
43. I am a creative person.
44. I play at least one musical instrument well.
45. I put other people's needs before my own.

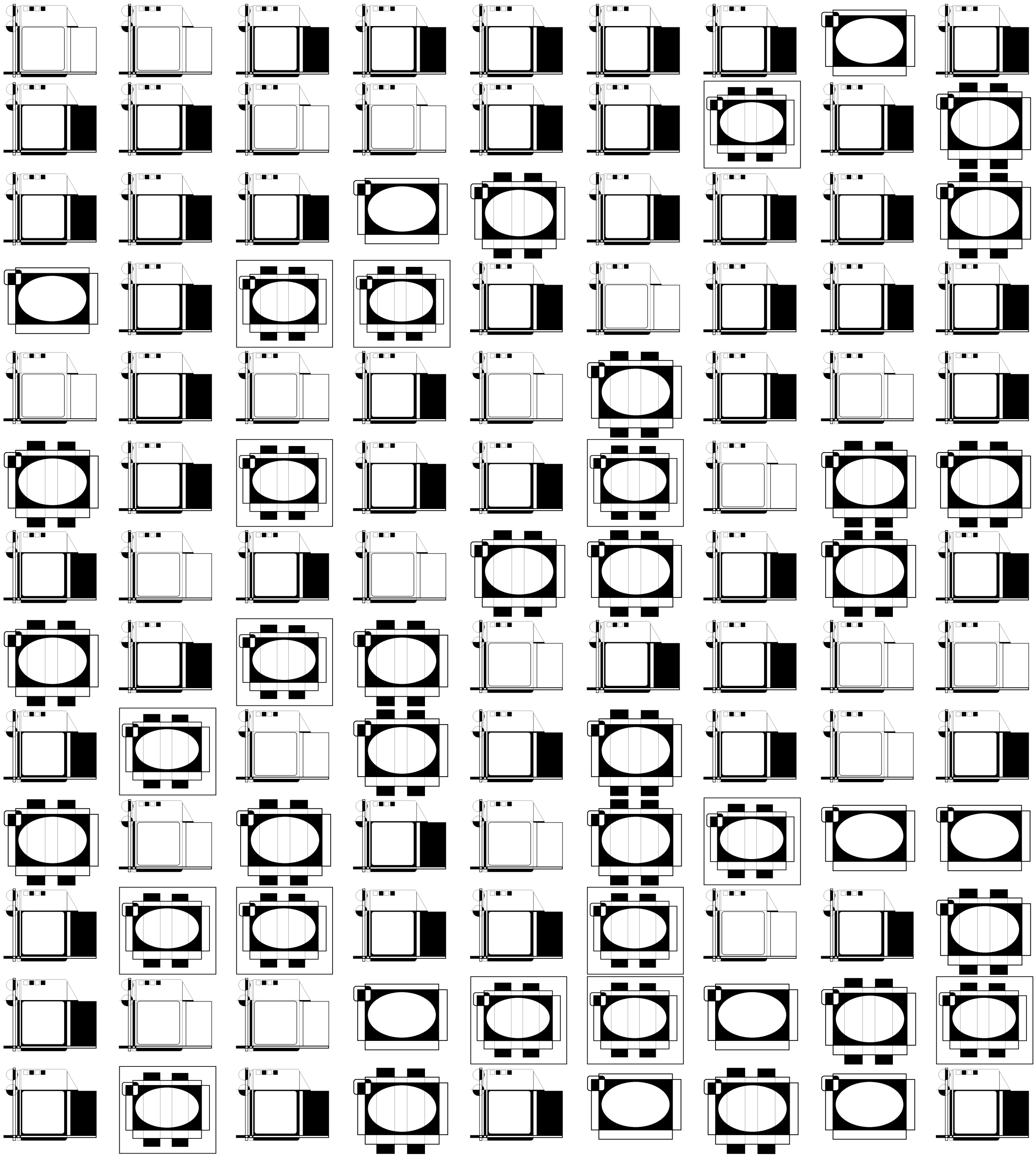
46. I avoid conflicts.
47. I often forget what I wanted to do a few minutes ago.
48. I get angry quickly if something doesn't fulfil my wishes.
49. I'm not allowed to show emotions at work.
50. Sometimes I tend to let my feelings run wild.
51. I have suffered from illnesses for no apparent reason.
52. I tend to carry out tasks quickly, but with mistakes
53. I always stand behind the decisions I make.
54. I have high expectations of myself.
55. It is very important to me that I am always committed.
56. I can change work steps quickly if necessary.
57. I often experience the feeling of losing control in my everyday life.
58. It wouldn't be a problem for me to work a lot of overtime.
59. In difficult situations, I take a solution-orientated approach.
60. I don't want others to realise when I can't do something.
61. I like working alone.
62. I can easily prioritise my work.
63. I can reduce stress by using relaxation techniques.
64. I am able to concentrate on work processes despite a heavy workload.
65. I take my anger out on bystanders.
66. As soon as I get too stressed at work, I take a coffee or smoke break to relax again.
67. I find it very easy to listen.
68. If necessary, I can easily manage a clear division of tasks.
69. I find it very difficult to make a short-term decision under great pressure.
70. Treating colleagues respectfully and appropriately in the workplace is not particularly relevant to me.
71. A job where you have to speak English is out of the question for me.
72. I am able to empathise with the feelings and sensitivities of another person.
73. After a stressful day, I prefer to relax with my family or friends.
74. I can't switch off after a stressful day.
75. I am very good at dealing with criticism.

Appendix A.2. Psychological Test #1

Indicate the frequency of occurrence of the target motif by marking (crossing out) the target motif. You have 20 s to complete this task.

Target motif





Appendix A.3. Psychological Test #2

Please identify and mark (using a highlighter!) all "ä" letters in a maximum of 20 s. Make sure you do not make any mistakes and process as many correct characters as possible.

a ä a a a ä g ä a ä a a ä a a a a a ä a ä a a a ä a ä a a a ä ü ä a a ä a a ä a ä a a

a a ä ä ä a ä a a ä a a ä a a ä ä a a a ä a a a ä a a a a a ä ä ä a ä ä a a ä x a ä a

ä a ä ä a ö a a a a a a a a a a ä ä a a x a a a a g a a a a a a a a a a a ü a a a a a

ü ü ä a a ä a a a a a a a a a a a ä a a a a a a g a a a a a a ü a a a a ü a a a a a a

a ä a a a a a a g ä d a a a a a a a a a a a a a a a a a a a a ä ä ä a ä a ä a a ä a a

a a a ä a a ä a a a a a a a a a a a a a a a a a a a ö a a a a ä a a a a a a a a ä a a

a a a a a a d a a a a a a d a a a ä a a a a d a a a a a a a a a a ä a a ä a a ä a a a

a d a a d a a a ä a g a d a a ä a a a a a d a a ö a a a g a a a a a a a a a a a a a a

a a a ä a a d a a a ö a d d a a a a d a ö a a a a a d a a a a a g ü ö a a a a a a a a

a ä ö a a a ä a a a a d a a a ö a a ö ö ü a a a a a a a a a a a a a a a a d a a a a a a

a a d a a a a ö a ö a a ä d a a a ö a ö a a a d a a a a a a a a a a d a a a ü ö ö a a a

a ä a a d a ä d a ä a a d a a d a a a a a a d a a a ö a a d a a g a a d a a a d a a ö

a a a a ö a a a ö a a ö a a a ö a a a ö a a a a a a a a a a a a a a ö a a a a ö a ö a a

a a d a d a a ö a a a a a g a ö a d a a a a ä ä a ö a ö a a a a a a ö a a a a ö a a a

a a a a a a ö d a a a a a ü a a a a a a ä ä a a d a a a a a a a a a a a a a a a a ü a

a a a a a g a a a a a ä a d a a ä a a a a a a a a a a a a a ö a a g a a a ä a a a a ä

a a g a a g a a a a a ö a a ö a a ö ö a a a g a a a ä ä a ä a ä ä a ö a a a ö a a ö a

a a a a ö a a ö a a a a a ä a a ä a ä a ä a ä a a ä a ä ä ä a g a a a a a a a a a a a

a a a a a ä a a ö a a a a a a ä a a ä a a a a a a ä a a ä ä a ä ö a a a a a a a a a a

a a a ö ä a g a a ä a g a a a a ä ä a a a ä a a g a a a g a a a a a a a a ö ö ö ü ä a

Appendix B. Detailed Transformation Results

Table A1. Thetransformation model is able to calculate a prediction of the result for each KPI and for each ATC. To interpret the result—to decide which version is better in an A/B test—the confidence level is determined that the test result prediction of one version being better than the other is not a coincidence.

KPI	T1 Obs.	ATC 1			ATC 2			ATC 3			ATC 4		
	Mean	**T2 Pred.**	**±**	**Conf. %**	**T2 Pred.**	**±**	**Conf. %**	**T2 Pred.**	**±**	**Conf. %**	**T2 Pred.**	**±**	**Conf. %**
Taken over #	10.125	9.557	0.491	20.0	9.561	0.51	25.0	9.089	1.033	45.0	9.648	0.33	10.0
Taken over %	0.92	0.869	0.045	20.0	0.869	0.046	25.0	0.826	0.094	45.0	0.877	0.03	10.0
Time until takeover total	209.625	432.808	189.579	20.0	411.18	196.767	25.0	408.771	168.424	20.0	738.025	461.484	35.0
Time until takeover/plane	20.75	50.38	28.65	25.0	43.49	18.78	20.0	48.038	25.453	25.0	88.283	64.108	40.0
Landings 1	4.5	3.699	0.755	35.0	3.777	0.721	40.0	3.632	0.776	40.0	4.582	0.141	<1.0
Landings 2	0.125	0.642	0.487	65.0	0.437	0.288	50.0	0.43	0.275	45.0	1.171	1.034	85.0
Landings 3	1.5	1.167	0.325	20.0	0.829	0.635	45.0	1.465	0.071	<1.0	1.983	0.441	20.0
Calculated Landings	4.938	4.24	0.624	25.0	4.148	0.734	35.0	4.135	0.79	35.0	5.551	0.502	15.0
Optimum Landings	0.9	0.74	0.151	35.0	0.755	0.144	40.0	0.726	0.155	40.0	0.916	0.028	<1.0
Calculated Optimum Landings	0.898	0.719	0.162	35.0	0.701	0.177	45.0	0.703	0.19	45.0	0.945	0.03	5.0
Time deviation to landing total	−223.25	110.315	321.609	70.0	45.174	265.277	70.0	109.627	318.146	75.0	−51.637	160.295	30.0
Time deviation to landing/plane	−48.875	54.504	100.055	55.0	6.061	49.343	35.0	51.515	99.209	60.0	−32.404	11.186	5.0
Distance deviation to landing total	−4.259	18.135	20.773	40.0	9.171	12.57	30.0	15.233	18.455	40.0	3.855	6.731	10.0
Distance deviation to landing/plane	−0.841	6.894	7.259	40.0	1.263	1.43	10.0	5.483	5.575	35.0	−1.389	1.174	<1.0
Height not landed total	42,327.0	56,172.602	13,038.33	65.0	50,691.314	7719.965	50.0	51,405.661	8314.942	50.0	45,807.772	2352.338	10.0
Height not landed/plane	6512.208	7489.741	924.834	45.0	7061.755	490.422	30.0	6977.802	436.547	25.0	6781.833	262.61	10.0
Distance not landed total	101.135	194.77	90.121	75.0	180.968	74.336	75.0	186.313	80.065	75.0	120.001	13.133	10.0
Distance not landed/plane	15.501	25.367	9.359	80.0	24.337	8.72	85.0	24.288	8.315	80.0	17.804	1.835	15.0
Distance not landed/plane %	1.09	0.747	0.316	70.0	0.749	0.325	80.0	0.768	0.313	75.0	1.06	0.026	5.0
Conflicts	0.125	1.784	1.494	35.0	−0.032	0.17	<1.0	1.726	1.535	40.0	2.698	2.344	40.0
Instructions/plane	5.057	4.758	0.259	10.0	4.515	0.432	20.0	4.631	0.347	15.0	4.657	0.352	10.0
Instructions total	51.125	45.374	5.561	25.0	43.516	7.566	40.0	42.745	8.149	40.0	43.166	7.545	25.0
NASA TLX Average [0, 100]	45.729	55.132	7.68	20.0	55.324	7.971	25.0	54.879	8.586	25.0	58.59	10.419	20.0
NASA TLX Average %	0.543	0.528	0.019	<1.0	0.481	0.048	15.0	0.489	0.052	15.0	0.559	0.026	<1.0
SASHA Q Average [0, 5]	3.578	3.44	0.122	20.0	3.372	0.18	35.0	3.404	0.165	30.0	3.807	0.208	25.0
SASHA Q Average %	0.716	0.688	0.024	20.0	0.674	0.036	35.0	0.681	0.033	30.0	0.761	0.042	25.0

Table A2. Continuation of Table A1.

KPI	T1 Obs.	ATC 5			ATC 6			ATC 7			ATC 8		
	Mean	**T2 Pred.**	**±**	**Conf. %**	**T2 Pred.**	**±**	**Conf. %**	**T2 Pred.**	**±**	**Conf. %**	**T2 Pred.**	**±**	**Conf. %**
Taken over #	10.125	9.74	0.312	10.0	9.787	0.276	10.0	9.299	0.671	20.0	9.84	0.278	10.0
Taken over %	0.92	0.885	0.028	10.0	0.89	0.025	10.0	0.845	0.061	20.0	0.895	0.025	10.0
Time until takeover total	209.625	602.965	370.135	30.0	384.841	160.485	15.0	676.498	464.453	35.0	515.786	272.442	25.0
Time until takeover/plane	20.75	70.214	44.453	30.0	41.674	19.274	15.0	76.202	47.332	30.0	59.721	32.72	25.0
Landings 1	4.5	3.475	0.968	35.0	4.125	0.356	15.0	3.606	0.875	30.0	4.095	0.359	15.0
Landings 2	0.125	0.791	0.625	65.0	0.989	0.765	80.0	0.811	0.665	65.0	1.097	0.87	85.0
Landings 3	1.5	0.957	0.524	25.0	1.55	0.091	<1.0	0.55	0.921	40.0	1.356	0.092	5.0
Calculated Landings	4.938	3.955	0.968	30.0	4.883	0.139	<1.0	4.07	0.852	25.0	4.884	0.14	<1.0
Optimum Landings	0.9	0.695	0.194	35.0	0.825	0.071	15.0	0.721	0.175	30.0	0.819	0.072	15.0
Calculated Optimum Landings	0.898	0.686	0.208	35.0	0.835	0.051	10.0	0.692	0.188	30.0	0.831	0.051	10.0
Time deviation to landing total	−223.25	−53.14	151.513	30.0	6.59	208.498	45.0	−2.038	219.912	40.0	−50.365	159.027	35.0
Time deviation to landing/plane	−48.875	−23.235	21.192	10.0	0.531	47.592	25.0	−12.772	33.968	15.0	−22.507	18.874	10.0
Distance deviation to landing total	−4.259	3.764	6.362	10.0	8.23	11.355	20.0	7.751	10.198	15.0	1.813	5.666	10.0
Distance deviation to landing/plane	−0.841	1.542	2.223	10.0	2.146	2.961	15.0	1.786	2.367	10.0	−0.748	0.988	<1.0
Height not landed total	42,327.0	48,251.071	5644.19	25.0	51,279.231	8235.991	40.0	54,466.841	11,311.126	45.0	49,827.201	7168.135	35.0
Height not landed/plane	6512.208	6713.634	123.847	5.0	7290.622	674.502	30.0	7565.874	956.702	35.0	7013.468	445.983	20.0
Distance not landed total	101.135	165.186	59.31	45.0	178.581	74.21	60.0	190.649	89.316	60.0	152.18	46.292	40.0
Distance not landed/plane	15.501	22.011	6.221	50.0	24.008	7.667	65.0	25.484	9.227	65.0	20.842	4.904	45.0
Distance not landed/plane %	1.09	0.9	0.176	35.0	0.835	0.232	50.0	0.708	0.349	60.0	0.913	0.156	35.0
Conflicts	0.125	1.622	1.343	25.0	0.268	0.234	<1.0	0.647	0.281	5.0	1.805	1.447	30.0
Instructions/plane	5.057	3.859	1.022	30.0	4.6	0.443	15.0	4.089	0.899	25.0	4.571	0.446	15.0
Instructions total	51.125	39.481	10.169	35.0	45.467	5.014	20.0	39.389	10.828	35.0	44.784	5.048	20.0
NASA TLX Average [0, 100]	45.729	48.628	2.436	5.0	48.433	2.155	5.0	60.58	13.195	25.0	55.611	8.771	20.0
NASA TLX Average %	0.543	0.565	0.025	<1.0	0.606	0.044	10.0	0.404	0.135	25.0	0.536	0.022	<1.0
SASHA Q Average [0, 5]	3.578	3.395	0.156	20.0	3.607	0.034	<1.0	3.257	0.299	35.0	3.775	0.175	25.0
SASHA Q Average %	0.716	0.679	0.031	20.0	0.721	0.007	<1.0	0.651	0.06	35.0	0.755	0.035	25.0

Table A3. The tests NASA TLX and SASHA Q each consist of individual subtests. Their predictions and partial results are listed separately in this table.

Item	T1 Obs.	ATC 1			ATC 2			ATC 3			ATC 4		
	Mean	T2 Pred.	±	Conf. %	T2 Pred.	±	Conf. %	T2 Pred.	±	Conf. %	T2 Pred.	±	Conf. %
NASA TLX Average [0, 100]	45.729	55.132	7.68	20.0	55.324	7.971	25.0	54.879	8.586	25.0	58.59	10.419	20.0
NASA TLX Average %	0.543	0.528	0.019	<1.0	0.481	0.048	15.0	0.489	0.052	15.0	0.559	0.026	<1.0
mental	61.562	54.651	6.39	15.0	80.49	18.898	50.0	61.33	1.881	0.0	72.7	8.668	15.0
physical	27.5	54.237	25.672	50.0	41.614	12.05	30.0	54.56	25.581	55.0	52.677	23.356	35.0
temporal	43.437	59.155	13.463	30.0	69.937	24.429	60.0	68.516	23.575	55.0	79.218	32.096	50.0
performance	48.75	62.723	13.38	30.0	30.149	17.166	45.0	48.466	1.931	0.0	37.793	8.896	15.0
effort	58.75	60.221	2.078	0.0	82.274	23.218	60.0	67.816	7.466	20.0	80.29	20.457	35.0
frustration	34.375	45.881	11.194	25.0	32.282	1.815	5.0	35.715	1.955	0.0	38.059	2.985	5.0
SASHA Q Average [0, 5]	3.578	3.44	0.122	20.0	3.372	0.18	35.0	3.404	0.165	30.0	3.807	0.208	25.0
SASHA Q Average %	0.716	0.688	0.024	20.0	0.674	0.036	35.0	0.681	0.033	30.0	0.761	0.042	25.0
manageable	4.625	3.814	0.693	30.0	3.169	1.399	65.0	3.328	1.213	55.0	3.712	0.777	25.0
next steps	4.5	3.839	0.566	25.0	3.696	0.77	40.0	3.423	1.07	50.0	4.258	0.151	5.0
heavy focus	2.375	4.08	1.634	65.0	2.841	0.455	25.0	3.106	0.699	35.0	2.108	0.147	5.0
find info	3.0	1.457	1.439	60.0	0.678	2.066	85.0	1.817	1.145	55.0	1.764	1.21	40.0
valuable info	3.375	3.832	0.441	20.0	4.025	0.553	30.0	3.398	0.097	0.0	3.872	0.446	15.0
attention	3.625	3.488	0.109	5.0	4.431	0.754	40.0	3.541	0.097	0.0	5.861	2.227	65.0
understanding	3.5	3.834	0.315	15.0	4.295	0.723	40.0	4.177	0.673	35.0	5.195	1.532	50.0
awareness	3.625	2.833	0.786	35.0	2.916	0.649	35.0	2.759	0.808	40.0	3.739	0.147	0.0

Table A4. Continuation of Table A3.

KPI	T1 Obs.	ATC 5			ATC 6			ATC 7			ATC 8		
	Mean	**T2 Pred.**	**±**	**Conf. %**	**T2 Pred.**	**±**	**Conf. %**	**T2 Pred.**	**±**	**Conf. %**	**T2 Pred.**	**±**	**Conf. %**
NASA TLX Average [0, 100]	45.729	48.628	2.436	5.0	48.433	2.155	5.0	60.58	13.195	25.0	55.611	8.771	20.0
NASA TLX Average %	0.543	0.565	0.025	<1.0	0.606	0.044	10.0	0.404	0.135	25.0	0.536	0.022	<1.0
mental	61.562	64.309	2.716	5.0	64.225	2.402	5.0	61.912	2.892	0.0	66.116	2.419	5.0
physical	27.5	33.004	3.043	5.0	56.379	25.778	45.0	61.749	31.025	45.0	57.472	29.319	50.0
temporal	43.437	47.043	2.804	5.0	61.226	15.273	30.0	78.648	32.302	50.0	66.389	20.959	40.0
performance	48.75	45.145	2.787	5.0	33.949	12.544	25.0	50.711	2.967	0.0	46.371	2.482	0.0
effort	58.75	80.333	19.336	35.0	61.934	2.358	5.0	69.523	8.562	15.0	67.407	7.162	15.0
frustration	34.375	36.06	2.821	0.0	23.72	10.092	20.0	48.359	12.146	20.0	38.39	2.513	5.0
SASHA Q Average [0, 5]	3.578	3.395	0.156	20.0	3.607	0.034	<1.0	3.257	0.299	35.0	3.775	0.175	25.0
SASHA Q Average %	0.716	0.679	0.031	20.0	0.721	0.007	<1.0	0.651	0.06	35.0	0.755	0.035	25.0
manageable	4.625	3.419	1.047	35.0	4.02	0.516	20.0	2.741	1.864	55.0	4.046	0.52	20.0
next steps	4.5	3.097	1.367	45.0	4.52	0.126	0.0	2.925	1.455	45.0	4.384	0.127	0.0
heavy focus	2.375	3.179	0.707	25.0	2.587	0.123	5.0	3.62	1.242	40.0	2.896	0.501	20.0
find info	3.0	0.203	2.542	75.0	0.641	2.249	75.0	1.007	1.965	60.0	1.359	1.473	55.0
valuable info	3.375	3.477	0.14	0.0	3.863	0.373	15.0	3.531	0.149	5.0	4.173	0.766	30.0
attention	3.625	3.198	0.421	15.0	4.103	0.373	15.0	4.016	0.298	10.0	4.793	1.044	40.0
understanding	3.5	3.87	0.268	10.0	3.906	0.357	15.0	3.803	0.286	10.0	4.297	0.734	30.0
awareness	3.625	2.838	0.707	25.0	3.419	0.123	5.0	1.991	1.601	50.0	3.333	0.248	10.0

References

1. European Commission. Reducing Emissions from Aviation. Available online: https://climate.ec.europa.eu/eu-action/transport/reducing-emissions-aviation_en (accessed on 8 April 2024).
2. EUROCONTROL. Aviation Outlook 2050: Air Traffic Forecast Shows Aviation Pathway To Net Zero CO_2 Emissions. 2022. Available online: https://www.eurocontrol.int/article/aviation-outlook-2050-air-traffic-forecast-shows-aviation-pathway-net-zero-co2-emissions (accessed on 8 April 2024).
3. Perott, A.; Schader, N.T.; Leonhardt, J.; Licu, T. Human Factors Integration in ATM System Design. *White paper, EUROCONTROL*, 2019.
4. IOS. *Ergonomics of Human-System Interaction—Part 210: Human-Centred Design for Interactive Systems*; International Organization for Standardization: Geneva, Switzerland, 2019.
5. König, C.; Hofmann, T.; Bruder, R. Application of the user-centred design process according ISO 9241-210 in air traffic control. *Work* **2012**, *41*, 167–174. [CrossRef] [PubMed]
6. Norman, D.A. *The Design of Everyday Things*; Basic Books: New York, NY, USA, 2002.
7. Rubin, J.; Chisnell, D. *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*, 2nd ed.; John Wiley & Sons: Hoboken, NJ, USA, 2008.
8. Stanton, N.A.; Salmon, P.M.; Rafferty, L.A.; Walker, G.H.; Baber, C.; Jenkins, D.P. *Human Factors Methods: A Practical Guide for Engineering and Design*, 2nd ed.; CRC Press: London, UK, 2013.
9. Tullis, T.; Albert, W. *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*, 2nd ed.; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2013.
10. Bach, C.; Scapin, D.L. Comparing inspections and user testing for the evaluation of virtual environments. *Int. J. Hum.-Comput. Interact.* **2010**, *26*, 786–824. [CrossRef]
11. Nielsen, J. *Usability Engineering*, 1st ed.; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1993.
12. Dumas, J.S.; Redish, J. *A Practical Guide to Usability Testing*; Intellect Books: Bristol, UK, 1999.
13. Dillon, A. The evaluation of software usability. In *International Encyclopedia of Ergonomics and Human Factors*; Karwowski, W., Ed.; Taylor & Francis: Hoboken, NJ, USA, 2001; pp. 1110–1112.
14. Sagar, K.; Saha, A. A systematic review of software usability studies. *Int. J. Inf. Technol.* **2017**, 1–24. [CrossRef]
15. Bastien, C.J.M. Usability testing: A review of some methodological and technical aspects of the method. *Int. J. Med. Inform.* **2010**, *79*, e18–e23. [CrossRef] [PubMed]
16. Bos, T.; Schuver-van Blanken, M.; Huisman, H. Towards a Paperless Air Traffic Control Tower. In Proceedings of the 2nd International Conference on Human Centered Design, Orlando, FL, USA, 9–14 July 2011; pp. 360–368. [CrossRef]
17. Huber, S.; Gramlich, J.; Pauli, S.; Mundschenk, S.; Haugg, E.; Grundgeiger, T. Toward User Experience in ATC: Exploring Novel Interface Concepts for Air Traffic Control. *Interact. Comput.* **2022**, *34*, 43–59. [CrossRef]
18. King, R.; Churchill, E.F.; Tan, C. *Designing with Data: Improving the User Experience with A/B Testing*, 1st ed.; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2017.
19. Kohavi, R.; Henne, R.M.; Sommerfield, D. Practical Guide to Controlled Experiments on the Web: Listen to Your Customers Not to the Hippo. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '07, San Jose, CA, USA, 12–15 August 2007; pp. 959–967. [CrossRef]
20. Young, S. Improving Library User Experience with A/B Testing: Principles and Process. *Weav. J. Libr. User Exp.* **2014**, *1*. [CrossRef]
21. Quin, F.; Weyns, D.; Galster, M.; Costa Silva, C. A/B testing: A systematic literature review. *J. Syst. Softw.* **2024**, *211*, 112011. [CrossRef]
22. Hagar, N.; Diakopoulos, N. Optimizing Content with A/B Headline Testing: Changing Newsroom Practices. *Media Commun.* **2019**, *7*, 117. [CrossRef]
23. Meta. Fundraising/2013-14 Report—Meta, Discussion about Wikimedia Projects. 2020. Available online: https://meta.wikimedia.org/wiki/Fundraising/2013-14_Report (accessed on 8 April 2024).
24. MediaWiki. Page Previews/2016 A/B Tests—MediaWiki. 2022. Available online: https://www.mediawiki.org/wiki/Page_Previews/2016_A/B_Tests (accessed on 8 April 2024).
25. MediaWiki. Page Previews/2017-18 A/B Tests — MediaWiki. 2020. Available online: https://www.mediawiki.org/wiki/Page_Previews/2017-18_A/B_Tests (accessed on 8 April 2024).
26. Milanzi, E.; Njeru Njagi, E.; Bruckers, L.; Molenberghs, G. Data Representativeness: Issues and Solutions. *EFSA Support. Publ.* **2015**, *12*, 759E. [CrossRef]
27. Bethlehem, J. *Applied Survey Methods: A Statistical Perspective*; John Wiley & Sons: Hoboken, NJ, USA, 2009.
28. Parsons, V.L. Stratified Sampling. In *Wiley StatsRef: Statistics Reference Online*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2017; pp. 1–11. [CrossRef]
29. Liberty, E.; Lang, K.; Shmakov, K. Stratified Sampling Meets Machine Learning. In Proceedings of the 33rd International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; Balcan, M.F., Weinberger, K.Q., Eds.; Volume 48, pp. 2320–2329.
30. Raiffa, H.; Schlaifer, R. *Applied Statistical Decision Theory*; Harvard University: New Haven, CT, USA, 1961.
31. Ericson, W.A. Optimum Stratified Sampling Using Prior Information. *J. Am. Stat. Assoc.* **1965**, *60*, 750–771. [CrossRef]

32. Hidiroglou, M.A.; Särndal, C.E. Use of auxiliary information for two-phase sampling. *Surv. Methodol.* **1998**, *24*, 11–20.
33. Ahsan, M.J.; Khan, S. Optimum allocation in multivariate stratified random sampling with overhead cost. *Metr. Int. J. Theor. Appl. Stat.* **1982**, *29*, 71–78. [CrossRef]
34. Khan, M.G.; Maiti, T.; Ahsan, M.J. An Optimal Multivariate Stratified Sampling Design Using Auxiliary Information: An Integer Solution Using Goal Programming Approach. *J. Off. Stat.* **2010**, *26*.
35. Varshney, R.; Siddiqui, N.; Ahsan, M.J. Estimation of more than one parameters in stratified sampling with fixed budget. *Math. Methods Oper. Res.* **2012**, *75*. [CrossRef]
36. Gupta, N.; Ali, I.; Bari, A. An Optimal Chance Constraint Multivariate Stratified Sampling Design Using Auxiliary Information. *J. Math. Model. Algorithms* **2013**. [CrossRef]
37. Deville, J.C.; Särndal, C.E. Calibration Estimators in Survey Sampling. *J. Am. Stat. Assoc.* **1992**, *87*, 376–382. [CrossRef]
38. Singh, S.; Horn, S.; Chowdhury, S.; Yu, F. Theory & Methods: Calibration of the estimators of variance. *Aust. N. Z. J. Stat.* **2002**, *41*, 199–212. [CrossRef]
39. Kim, J.M.; Sungur, E.; Heo, T.Y. Calibration approach estimators in stratified sampling. *Stat. Probab. Lett.* **2007**, *77*, 99–103. [CrossRef]
40. Wu, C.; Sitter, R. A Model-Calibration Approach to Using Complete Auxiliary Information From Survey Data. *J. Am. Stat. Assoc.* **2001**, *96*, 185–193. [CrossRef]
41. Rottermanner, G.; Settgast, V.; Judmaier, P.; Eschbacher, K.; Rokitansky, C.H. VAST: A High-Fidelity Prototype for Future Air Traffic Control Scenarios. In Proceedings of the 17th European Conference on Computer-Supported Cooperative Work, Salzburg, Austria, 9–12 June 2019; Volume 3; Reports of the European Society for Socially Embedded Technologies. [CrossRef]
42. Rottermanner, G.; de Jesus Oliveira, V.A.; Lechner, P.; Graf, P.; Kreiger, M.; Wagner, M.; Iber, M.; Rokitansky, C.H.; Eschbacher, K.; Grantz, V.; et al. Design and Evaluation of a Tool to Support Air Traffic Control with 2D and 3D Visualizations. In Proceedings of the 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Atlanta, GA, USA, 22–26 March 2020; pp. 885–892. [CrossRef]
43. Rind, A.; Iber, M.; Aigner, W. Bridging the gap between sonification and visualization. In Proceedings of the AVI Workshop on Multimodal Interaction for Data Visualization (MultimodalVis), Castiglione della Pescaia Grosseto, Italy, 29 May–1 June 2018.
44. Rottermanner, G.; Wagner, M.; Kalteis, M.; Iber, M.; Judmaier, P.; Aigner, W.; Settgast, V.; Eggeling, E. Low-fidelity prototyping for the air traffic control domain. *Mensch Comput.* **2018**, 605–614.
45. Rottermanner, G.; Wagner, M.; Settgast, V.; Grantz, V.; Iber, M.; Kriegshaber, U.; Aigner, W.; Judmaier, P.; Eggeling, E. Requirements analysis & concepts for future european air traffic control systems. In Proceedings of the Workshop Vis in Practice-Visualization Solutions in the Wild, IEEE VIS, Phoenix, AZ, USA, 1–6 October 2017.
46. Steinheimer, M.; Gonzaga-López, C.; Kern, C.; Kerschbaum, M.; Strauss, L.; Eschbacher, K.; Mayr, M.; Rokitansky, C.H. Air traffic management and weather: The potential of an integrated approach. In Proceedings of the International Conference on Air Transport (INAIR), Vienna, Austria, 10–11 November 2016; Hromádka, M., Ed.; University Of ŽIlina: Žilina, Slovakia; pp. 120–126.
47. Hart, S.G. *NASA Task Load Index (TLX). Volume 1.0*; Paper and Pencil Package; NASA: Washington, DC, USA, 1986.
48. Jeannot, E.; Kelly, C.; Thompson, D. *The Development of Situation Awareness Measures in ATM Systems. Report Eurocontrol HRS*; Technical Report, HSP-005-REP-01; EUROCONTROL: Brussels, Belgium, 2003.
49. Durso, F.T.; Manning, C.A. Air Traffic Control. *Rev. Hum. Factors Ergon.* **2008**, *4*, 195–244. [CrossRef]
50. Hilburn, B. Cognitive complexity in air traffic control: A literature review. *EEC Note* **2004**, *4*, 1–80.
51. Goldberg, L.R. The Development of Markers For the Big Five Factor Structure. *Psychol. Assess.* **1992**, *4*, 26–42. [CrossRef]
52. Wright, S. Correlation and causation. *J. Agric. Res.* **1921**, *20*, 557–585.
53. Olive, D.J. Multiple Linear Regression. In *Linear Regression*; Springer International Publishing: Cham, Switzerland, 2017; pp. 17–83. [CrossRef]