On the appropriate participant expertise for display evaluation studies

van Paassen, M.M.; Borst, C.; Mulder, Max

Publication date
2023

Document Version
Accepted author manuscript

Published in
22nd International Symposium on Aviation Psychology

On the Appropriate Participant Expertise for Display Evaluation Studies

Conference Paper · May 2023

CITATIONS
0

READS
26

5 authors, including:

Marinus M. Van Paassen
Delft University of Technology
478 PUBLICATIONS **6,014** CITATIONS
SEE PROFILE

Clark Borst
Delft University of Technology
112 PUBLICATIONS **927** CITATIONS
SEE PROFILE

Max Mulder
Delft University of Technology
546 PUBLICATIONS **6,652** CITATIONS
SEE PROFILE

Gijs de Rooij
Delft University of Technology
8 PUBLICATIONS **4** CITATIONS
SEE PROFILE

ON THE APPROPRIATE PARTICIPANT EXPERTISE FOR DISPLAY EVALUATION STUDIES

M. M. (René) van Paassen, Clark Borst, Max Mulder and Gijs de Rooij
Aerospace Engineering – Delft University of Technology
Delft, The Netherlands
Ferdinand Dijkstra - LVNL Netherlands, Schiphol, The Netherlands
Adam Balint Tisza - EUROCONTROL, Maastricht, The Netherlands

Expert participants may not always be available for evaluation of new displays or support systems, and in some cases, it might be better to use novice participants, particularly when the display or support significantly changes existing work practices. To provide tools and arguments for selecting the expertise level of participants, we propose the use of Rasmussen's decision ladder to analyze where and how a new visualization or a support tool changes the task, and identify steps where a novice participant may learn to perform the task to an acceptable level. A comparison to the support with the current operational interfaces then shows where an expert might have difficulty in stepping away from learned practice. This analysis is applied to the domain of air traffic control, and a selected set of relevant past research with both expert and novice participants is reviewed, revisiting the decision for a participant level in the study.

Introduction

New designs for interfaces, or new support tools, are commonly tested in controlled experiments with human participants. Ideally, the existing version of the interface or support system is tested against the new development in a study that replicates daily operations to such an extent that differences in performance, and expert opinion, indicate whether the new implementation is more efficient and safe. A properly performed evaluation should be indicative of effects in practice, or in other words have external validity (Libby et al., 2002).

Expert participants may not always be available for these interface evaluation studies, and, often out of need, non-expert participants are invited. This may affect the validity of an experiment, primarily by changing the capability of an experiment to correctly assess the effect of manipulations in its experimental conditions, i.e., its internal validity (Libby et al., 2002). But given that there is sometimes no opportunity to use expert participants, and particularly in testing early prototypes, one would prefer to not use scarce opportunities for access to experts, there is often a need to invite and train novices for evaluation.

Expertise by itself is difficult to define; Gobet (2015) defines it in terms of performance with respect to others. Chase and Simon (1973) argue that expert chess players have a vast memory for structures in chess, and can therefore better code and remember chess positions, resulting in improved chunking, indicating how long-term memory and trained perception play a large role in expertise. Given that training for many expert level jobs takes several years, and developing senior expertise in most positions requires at least a decade, it can be argued that any training of a novice or relatively inexperienced participant before experimental evaluation sessions can never approximate true expertise. On the other hand, commonly the number of different conditions presented in an experiment are limited, and only a fraction of an expert's vast store of knowledge will be needed to perform the task. Given due care, it should in many cases be possible to use results obtained with novice participants, properly trained to perform the experiment tasks, to provide a realistic evaluation of a new display, or provide insight into how a (partial) task is approached and performed.

To provide a handle on judging the effect of using participants with an expertise level that differs from the intended end-users, we will apply cognitive task analysis with the "decision ladder" (Rasmussen, 1983; Rasmussen, 1986). The decision ladder model describes processes and knowledge stages in human information processing. It will be used here to assess the potential effect of a difference in expertise level on these different processing stages, and from there on to infer the effect on performance in an experiment. Using the distinction between the processes for different cognitive stages may support a systematic review of the effect of expertise; alternative cognitive models, such as IDA or ACT-R (Anderson et al., 1997; Ritter et al., 2019; Smidts et al., 1997) would offer the means to model and implement these same information stages and knowledge states, but the lack of distinction between these processes does not offer a systematic checklist for evaluating the effect of expertise.

In addition to discussing the decision ladder as a tool to evaluate the effect of using participants with different expertise levels, the paper gives a small overview of some past experiments or evaluations, and reviews these with the new approach.

Cognitive Task Analysis as guidance

Air Traffic Control (ATC) is responsible for the safety and efficiency of air traffic. Commonly, air traffic is monitored and directed through plan views reflecting radar screens, in which fused data from radar systems and on-board navigation systems is used to position aircraft symbols. When providing support with radar vectors, the Air Traffic Controllers (ATCos) provide speed, heading and altitude instructions to the aircraft under their control, both to solve possible conflicts and ensure an efficient and regular traffic flow in the sector. To provide support in these tasks, several interfaces and support systems are in use, and new ones are being developed and evaluated. The continuing shortage of ATCos, the considerable investments needed for selection and training and requirements on safety provide a need for innovation and development of support systems, at the same time there is a lack of available experts to properly test and evaluate these systems.

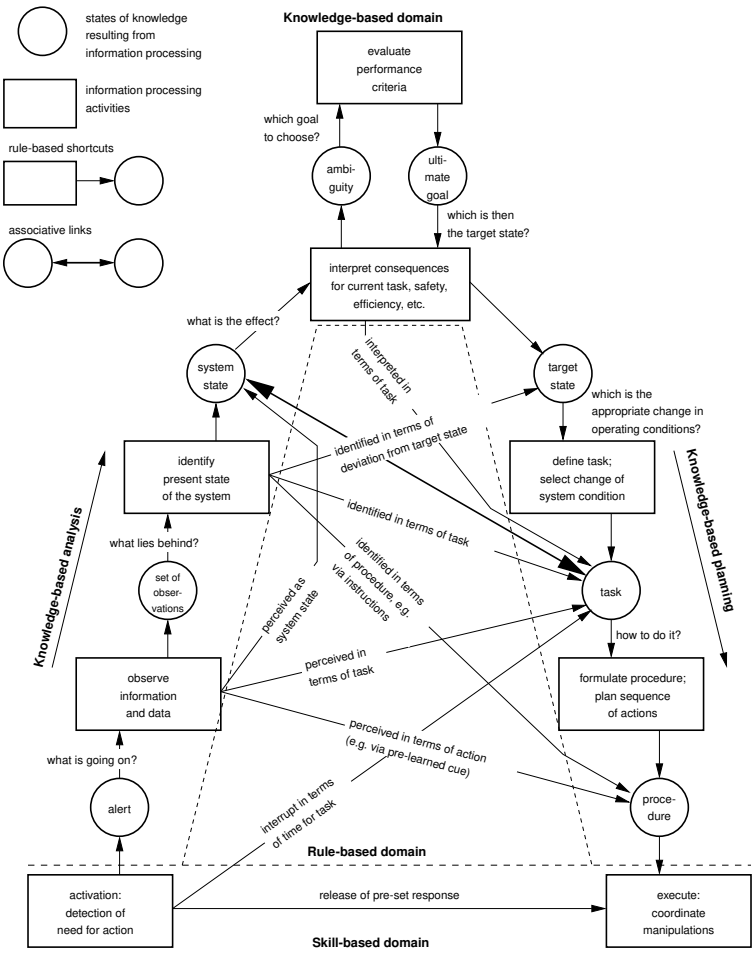

Figure 1: Decision ladder, showing perceptual and cognitive processes, and information stages. After (Rasmussen, 1983)

For the argument in this paper, we assume that a full or partial simulation is set up for the task (ATC tasks in this case), and any interfaces or support systems are tested in representative task scenarios. In the following, we will discuss which factors might affect experiment outcome when participants with an expertise level differing from the target users are invited.

As an example, the research by Somers et al. (2019) used an experiment with both expert ATCo and less experienced participants to generate the data for evaluating different ATC complexity metrics. This project used a simplified, hypothetical sector shape, with a simplified traffic sample and control of the traffic through a mouse-operated interface.

The "decision ladder" model by Rasmussen (Rasmussen, 1986) is used here as a template to consider the effect of participant expertise level on the outcome of experiments. This model is, as explained in (Vicente, 1999), fashioned after models for stages in cognitive processes, augmented with options for rule-based shortcuts that represent the repertoire of standard inferences and responses available to an operator or controller. The decision ladder model distinguishes a number of cognitive processes and their resulting knowledge states; using these distinctions, the potential effect of the level of expertise of experiment participants is discussed.

Consider the graphical representation of the decision ladder in Figure 1. The experience level of a participant might affect these processes in the following ways:

activation Activation provides the initiation of an activity. This starts at a skill-based level, with perception and pattern recognition, in our example with the identification of a need for action, e.g., with an aircraft entering the sector, an aircraft leaving, and thus requiring a hand-over, or the detection that aircraft might become involved in a conflict. Participants with less experience can be expected to have less efficient activation, leading to missing events, or inefficient detection, initializing activities when none are needed. Since currently in ATC the presentation of information is steady and clear, there will likely not be a large effect of experience, in contrast to situations where raw radar images need to be interpreted, or where unlabeled objects would have to be monitored, e.g., in the case of a radar operator responsible for detecting incoming attacks (Klein, 1999).

observation A next step, after a cognitive task has started, is commonly to assemble necessary information. We can assume that if an experiment closely resembles an expert's working environment, the expert can more efficiently gather information, spending less time and effort in this state. On the other hand, if an experimental interface is used, or even a simple re-arrangement of the information has taken place between the working environment and the experiment, an expert's "advantage" quickly disappears.

identification This is the process of understanding and classifying the state of the system to be controlled. If the fidelity of the experimental environment is good enough, the routine and the repertoire or experience of the expert will greatly support this step. Experts will be able to see more nuances, and know more conditions in which the system can be. Applied to the case of ATC, an expert ATCo will likely better understand the flow patterns in the sector, and will be able to group flights, rather than work on a case-by-case basis. This implies that for experiments that require a complete, high-fidelity task, there will be a larger difference between expert and novice participants. On the other hand, interfaces that facilitate the identification of the process (e.g., by presenting the information at a higher level in an integrated fashion), might support novices better, because these will more readily accept and use the new support.

interpretation In this step, the state of the controlled system is checked against the desirable or goal state. Experiment participants with a higher level of expertise most likely have a better definition of their goal state; in a further analysis of Somers' experiment, de Jong and Borst, 2022, found that the expert participants created a regular structure for merging the traffic, where the less experienced participants used more direct-to instructions to the exit waypoint. Thus, the participant's interpretation of the goal state shapes the experimental results. If an experimental interface allows or even promotes a different approach to the work, this might also be the point where experts might raise most objections, most likely ignoring the support and persisting with learned approaches to the work, where novices accept the structure proposed by a support system. This might also indicate a case where more familiarization with the interface, and efforts to explain new support, might help "win over" experts.

evaluation and criteria In the diagram, this is presented as an optional pathway. It is a meta-cognitive phase, in which performance and goals are evaluated, and goals are adjusted if that is deemed necessary. In ATC, the decision to start using an approach stack, or the decision to divert flights when runways need to be closed, e.g., due to weather conditions, fall in this step. Such conditions are seldom investigated in evaluation experiments, and when considered, require the participation of experts, as the performance of non-experts on these tasks is likely to be significantly different.

task definition In this step, the activities needed to achieve a desired state are planned and/or formulated. For experts, these steps may be immediately clear, being associated and known solutions to recognized deviations. Participants with less expertise will require more effort in this stage. Providing support in task definition, for example by offering a menu of resolutions, or a menu of actions, might improve novice performance, but any mismatch between the implementation preferred by experts and offered by the interface might result in rejection of the support.

procedure formulation This will largely rely on experience by the operator. In most experiments, there is a focused, relatively simple task to be performed, and the procedure formulation has limited variation, and can be quickly trained. Again, this is an aspect of the work that might be seen as disruptive to an expert if it is different from what is used in practice.

execution In many cases, the execution step in an experiment differs from the one in actual practice, e.g., entering vectors with a command interface versus radio communication with pilots. Since, contrary to perception skills,

execution skills are in many cases not critical in computer supported work (Vicente, 1999), differences between simulation environments used for experimental evaluation and practice will seldom affect the outcome.

In addition to effects of expertise on different cognitive steps, one can expect differences in rule-based behavior depending on the level of expertise. An expert might have better recognition of routine situations, and know more routine responses to handle these, leading to shunts (shortcuts to knowledge states), and leaps, known associations between recognized knowledge states (Vicente, 1999). When offering practice runs to novices, enough training should be given so that at least common situations in the experiment task can be handled in a rule-based, recognition-driven manner.

From reasoning with the decision ladder model, one can see that the effects of providing additional display support or decision support may in a number of cases be amplified when using novice participants. One should particularly expect larger differences when novices have difficulty with task aspects, and the interface or support system can provide clear distinctions (perception and interpretation stages) or structure (definition and execution stages) needed in the task. In the following, we will review a number of experimental evaluations and try to assess the effect of the expertise level of the participants on the outcomes of the comparison. Table 1 gives a summary of the assessed effects of expertise.

Application to past experiments

The experiments described in Somers et al. (2019) and De Jong and Borst (2022) were intended to provide subjective rating data for comparison against different candidate workload metrics calculated from the traffic state. In the experiments, traffic scenarios in simplified sectors were controlled with a menu-based interface. The modifications to the control input remove the need for radio phraseology, and remove any uncertainty in the execution. This simplifies task planning, and removes any differences due to expertise in execution. The absence of wind in the scenario made them more predictable, and thus reduced the requirements on interpretation of consequences, and removed the need to consider wind in observation. By normalizing the rating data, differences in expertise are largely eliminated. The experiment also consisted of regular traffic, with the exit waypoint shown for each of the aircraft in the sector, reducing planning and structuring effort. Data from both expert and non-expert participants could thus be used; the only difference in behavior was that experts produced a more regular traffic pattern, while non-experts would accept a slightly less regular pattern, with more aircraft on direct headings towards their exit points.

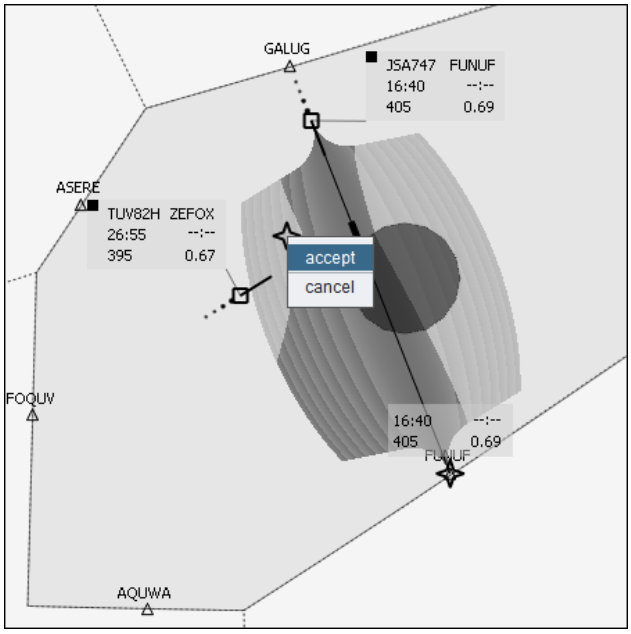

Figure 2: Screenshot of a 4D trajectory manipulation interface (R. E. Klomp et al., 2013).

In a study to investigate strategies of conflict evaluation in ATC, (Rantanen & Nunes, 2005), invited both expert and non-expert participants. The task focused on conflict classification only, with the presentation of only a conflict pair, essentially replacing activation by a fixed cue (the presentation of the next experiment condition) and replacing all execution steps by a press of a key on the keyboard. The effect of expertise on the further development of traffic pattern in the experiment is thus not an issue. The study showed a consistent effect of altitude differences on the required time to analyze conflicts for both the expert and non-expert groups, with the non-experts requiring somewhat more time, and displaying a larger variation when presented with ”difficult” conflict geometry of aircraft at the same flight level.

Table 1: *Summary of consideration for the effect of expertise level on efficiency and behavior at different cognitive processing stages*

Stage	Non-expert	Expert
Detect/alert	Relies on basic features, less efficient, may miss or exhibit spurious false alert	Detect complex patterns, efficient
Observe	More laborious, possibly inefficient, have misses or serendipitous hits	Efficient, look only for needed information, serendipitous misses in low probability scenarios
Identify	More effort needed, coarse identification	Often recognized as pattern, sparse, little superfluous data needed
Interpret	Missing threats, or make mountains out of molehills	Easily evaluate goal state relation
Evaluate criteria	Likely invalid, not enough expertise	Valid only in high-fidelity scenario and simulation, difficult to elicit and interpret in an experiment
Define task	Produces single, simple tasks	May produce more complex tasks
Formulate procedures	Need to provide enough training, may require more time, may limit task formulation	Requires little effort if matching work situation
Execute	Need to provide enough training	Automatic if matching work situation
Shortcuts	Provide enough training to enable rule based shortcuts	Check that learned rules are applicable in the experiment set-up

While Somers' experiment used a largely conventional interface, the research by Klomp, and further developments in that field (R. Klomp et al., 2016; R. E. Klomp et al., 2013; ten Brink et al., 2019), focused on new 4D ATC concepts, see Fig. 2. The controller in these situations effectively produces or modifies four-dimensional planned trajectories, by adjusting a speed, height and lateral profile defined by waypoints and speed and altitude targets. Many facets of a current ATCo's expertise become less relevant; activation is supported by the automation, through highlighting of parts of the trajectory with a future loss of separation, and selection of one of the conflicting aircraft gives an overview of both the currently planned 4D trajectory and the options to modify this trajectory into a conflict-free one, largely supporting observation, identification, interpretation and task definition stages. Most of the work will be new, both to experts in the current system and non-experts. A probable advantage of experts will be the evaluation of the 4D trajectories against aircraft performance limits. Since so much of the task, interface and work instructions is new for these evaluations, initial evaluations can well be performed with non-experts in ATC, and any further evaluations with participants trained in current ATC practice will need ample introduction into the new practice and tools.

An interesting middle ground is found in the evaluation by Mercado Velasco et al. (2021), where a "solution space" display is added to support present-day tactical ATC. This display shows combined speed and heading solutions that are clear from surrounding traffic. Participants at all three expertise levels; novices, intermediate and expert air traffic controllers showed improvement when using the tool. Experts did use it in a different manner, formulating their own solutions, and then using the tool for confirmation, effectively ignoring the support of the tool for most of the cognitive processing stages. Novices and intermediate level participants relied more heavily on the tool, using it for guidance, and they would also select solutions indicated by the tool with smaller separation margins, solutions that the experts, using a more conservative approach to generating solutions, would not consider.

Conclusions

In many types of experimental evaluation or fundamental research, one may be forced to select non-expert participants. The use of the decision ladder model as a template for the cognitive steps in a task, provides a list of cognitive processes, each of which may be influenced by the expertise level of participants in a task. A systematic check should be used to analyze what the effect is of the participants' expertise level in performing each cognitive process, and whether the experiment environment matches or differs from the targeted operational environment at

this level. In addition, the repertoire of trained rule-based behavior for both non-expert participants has to enable routine-based responses to a reasonable degree.

When comparing two or more interface variants or support systems, the participation of non-experts may lead to increased variation in performance, and also to an increased contrast, when participants strongly rely on support given in certain experimental conditions, where expert participants might be able to perform the task without support. Experimental set-ups may also be simplified in comparison to work situations, and care should be taken that the simplification does not unnecessarily constrain the response option of expert participants.

References

Anderson, J. R., Matessa, M., & Lebiere, C. (1997). ACT-r: A theory of higher level cognition and its relation to visual attention. *Human–Computer Interaction*, *12*(4), 439–462.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81.

de Jong, T., & Borst, C. (2022). Determining air traffic controller proficiency: Identifying objective measures using clustering. *IFAC-PapersOnLine*, *55*(29), 7–12.

Gobet, F. (2015). *Understanding expertise: A multi-disciplinary approach*. Palgrave Macmillan.

Klein, G. A. (1999). *Sources of power how people make decisions*. MIT Press.

Klomp, R., Borst, C., van Paassen, M. M., & Mulder, M. (2016). Expertise level, control strategies, and robustness in future air traffic control decision aiding. *IEEE Transactions on Human-Machine Systems*, *46*(2), 255–266.

Klomp, R. E., Borst, C., Mulder, M., Praetorius, G., Mooij, M., & Nieuwenhuisen, D. (2013). Experimental evaluation of a joint cognitive system for 4d trajectory management.

Libby, R., Bloomfield, R., & Nelson, M. W. (2002). Experimental research in financial accounting. *Accounting, Organizations and Society*, *27*(8), 775–810.

Mercado Velasco, G., Borst, C., van Paassen, M., & Mulder, M. (2021). Solution space decision support for reducing controller workload in route merging task. *Journal of Aircraft*, *58*(1), 125–137.

Rantanen, E. M., & Nunes, A. (2005). Hierarchical conflict detection in air traffic control. *The International Journal of Aviation Psychology*, *15*(4), 339–362.

Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man and Cybernetics*, *SMC-13*(3), 257–266.

Rasmussen, J. (1986). *Information processing and human-machine interaction : An approach to cognitive engineering*. North-Holland.

Ritter, F. E., Tehranchi, F., & Oury, J. D. (2019). ACT-r: A cognitive architecture for modeling cognition. *WIREs Cognitive Science*, *10*(3).

Smidts, C., Shen, S., & Mosleh, A. (1997). The IDA cognitive model for the analysis of nuclear power plant operator response under accident conditions. part i: Problem solving and decision making model. *Reliability Engineering & System Safety*, *55*(1), 51–71.

Somers, V., Borst, C., Mulder, M., & van Paassen, M. (2019). Evaluation of a {3d} solution space-based ATC workload metric. *IFAC-PapersOnLine*, *52*(19), 151–156.

ten Brink, D. S. A., Klomp, R. E., Borst, C., van Paassen, M. M., & Mulder, M. (2019). Flow-based air traffic control: Human-machine interface for steering a path-planning algorithm. *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 3186–3191.

Vicente, K. J. (1999). *Cognitive work analysis: Toward safe, productive, and healthy computer-based work*. Lawrence Erlbaum Associates.