Foyle, D. C. and Hooey, B. L. (2003). Proceedings of the 12th International Symposium on Aviation Psychology, 397-402. Dayton, OH: Wright State University

IMPROVING EVALUATION AND SYSTEM DESIGN THROUGH THE USE OF OFF-NOMINAL TESTING: A METHODOLOGY FOR SCENARIO DEVELOPMENT

David C. Foyle
Human Factors Research & Technology Division
NASA Ames Research Center
Moffett Field, California

Becky L. Hooey
Monterey Technologies, Inc.
NASA Ames Research Center
Moffett Field, California

A major challenge facing system designers is to ensure that a system that has been successfully evaluated in the laboratory will be successfully fielded. Off-nominal testing, in which unexpected conditions are evaluated, is proposed as a method to meet this challenge. Off-nominal testing allows for increased understanding of the human-machine system under evaluation, may uncover design issues that can be addressed, and may help determine training issues and procedures. An experimental methodology for conducting off-nominal testing in human-in-the-loop evaluations is presented. The methodology involves developing issues to be tested and off-nominal events addressing those issues, and estimating their disruptiveness on the test participant and the other test trials. Low and moderately disruptive off-nominal events are then incorporated into an experimental design while still allowing for the accurate estimation of dependent measures in the nominal trials. A single final trial can be incorporated for highly disruptive or truly surprising off-nominal events.

Introduction

An important challenge facing system developers is to ensure that a fielded system "scales-up" from laboratory evaluations. Discovering previously unknown interactions, errors, or other negative impacts when the system is fielded may lead to costly setbacks. These setbacks may take the form of unacceptable user demands, poor user/system performance, or user non-acceptance, which, in the extreme, could result in the system being abandoned. One solution to this challenge is to develop evaluation methodologies that can be applied during development prior to the system's fielding. Understanding and rectifying problems prior to fielding will result in a more robust fielded system.

Foyle, Andre, McCann, Wenzel, Begault and Battiste (1996) and Hooey, Foyle and Andre (2002) have described a methodology for the development of human-centered designed systems (see Figure 1). This methodology is based on task and information requirements analyses leading to the development of system requirements. It is iterative in nature, with user testing taking the form of field observation, focus groups, part-task and full-mission simulation and field/flight tests. The nature of the tests in the "iterative evaluation and validation loop" of the human-centered design process, however, has not yet been previously specified.

The present paper proposes an extension of the human-centered design process described above. Specific methodological techniques for the development of scenarios that evaluate off-nominal conditions, while maintaining the integrity of nominal condition evaluations, will be presented. These off-nominal evaluation techniques were used in two experiments: A part-task simulation experiment of display formats for aircraft taxi (Foyle, Hooey, Wilson & Johnson, 2002) and a full-mission simulation experiment of procedures with an advanced taxi display system (Hooey, Foyle & Andre, 2000).

Off-nominal Scenario Testing

A recent white paper prepared for the Federal Aviation Administration by Parasuraman, Hansman and Bussolari (2002) addresses the problems and issues associated with technical transfer of systems from the laboratory to the field within the context of aviation systems for surface operations safety. They argue for early human factors input, not only into the display interface as is more typical, but also into the very system functional requirements. They suggest that failure to do so may lead to "a mismatch between the functionality as specified by the designer, the operating environment (i.e., procedures), and the user's requirements for the system or his or her mental model of system functionality. The result can be inefficient system performance, errors, and possible adverse performance including accidents." (p. 7).

With a similar goal in mind, Leveson (2001a, 2001b) has advocated for the value of "off-nominal" software testing, in which software is evaluated under unexpected conditions. This stands in contrast to testing under "nominal" conditions in which everything goes exactly according to plan and procedure. She notes that software requirements typically avoid specifications of what the software should not do, and that, in fact, some industry standards forbid such negative requirements. Leveson states that this leads to software that specifies nominal system behavior well, but incompletely specifies off-nominal behavior. Furthermore, she argues that this incomplete specification of off-nominal behavior has been a factor in aviation and space-mission accidents.

In contrast to a nominal scenario, in which operations follow standard or formalized procedures, an off-nominal scenario is one in which the unexpected occurs (Leveson, 2001a, 2001b). The unexpected may range from minor deviations that occur frequently in an operational setting to the catastrophic failure of the system or sub-systems. The inclusion of off-nominal scenarios allows for: A full exercise of

the system; a determination of where and why a system fails; the exploration of interactions with other user agents; a deeper understanding of usage; and, the evaluation of procedures and integration issues. Rather than testing only the potential success of the proposed system, the approach advocated here includes tests of "plausible conditions of failure."

Two Philosophies of Scenario Development

Recently, the authors have observed a recurring problem in discussions with systems engineers and designers planning system tests and evaluations. The problem is that in developing test scenarios an inappropriate balance between testing nominal scenarios and off-nominal scenarios is made. Below, we present three actual interactions that the authors have had with system engineers/designers regarding scenario development for human-in-the-loop system testing.

In one case, during the development of a new system, the unstated test emphasis appeared to be primarily aimed at demonstrating the benefits of the system. It was acknowledged that off-nominal scenarios tested interesting and important conditions, but the system engineer feared that the inclusion of off-nominal scenarios would "contaminate" the nominal data. Testing of an off-nominal condition on the last trial for each participant was acknowledged as an option, since nominal data testing would be complete prior to that point. This can be described as:

Philosophy #1: Off-nominal events can only be incorporated into the very last trial. Off-nominal events are so disruptive that there is a need to "protect" nominal data.

In two other separate cases, the systems being developed were user alerting systems. In both of these cases, the test designer's implicit purpose was to demonstrate that the alerting system provided an alert to which the user could respond appropriately. For the first system, the test included alerting conditions on every trial, while for the second system, they occurred on about 90% of the trials. In both cases, the designer/evaluators wanted to evaluate the responses to the alerts, so they felt that most (or all) of the trials needed to incorporate alerting conditions, so as not to waste valuable simulation test time. This can be described as:

Philosophy #2: Tests should not waste time collecting nominal data -- off-nominal data is where the interesting findings lie.

Of course, there is value in both of these philosophies: One does need to ensure that nominal data are valid and not adversely affected by off-nominal testing (Philosophy #1); and, off-nominal data does, indeed, provide interesting and valuable information regarding usage (Philosophy #2). The challenge, however, is in how best to incorporate and balance these two competing goals into the development of scenarios for system evaluation.

Integrating the Two Philosophies

A study design that incorporates both nominal and off-nominal scenarios is highly efficient and has a number of advantages. Such a study, in addition to assessing both normal and non-normal usage, allows the nominal scenarios to act as an experimental control for the off-nominal scenarios -- verifying that the participant was "on-task" prior to the off-nominal event, manipulating the user's expectation of the system, and allowing for comparative performance measurements.

Normal usage assessment. The nominal scenarios allow for system assessment under "normal" conditions, those that will typically be encountered. This can, and should, include a wide-range of possible routine scenarios. Examples from the taxi domain might include runway holds and route amendments. The inclusion of nominal scenarios allows for the assessment of such things as usage patterns, workload, and efficiency gains with the system (e.g., speed, accuracy). A wide variety of nominal scenarios will help ensure robustness and success of a system once fielded.

Non-normal usage assessment. The off-nominal scenarios allow for system assessment when conditions are not normal -- ranging from slightly less-than-perfect operational and environmental conditions to partial or full system failures or inaccuracies. These off-nominal assessments give insight into the users' model of the system and how they interact with it (e.g., the behavior after a system failure may show user complacency or over-reliance). These issues can then be addressed through changes in system design, training and procedures.

On-task performance control. Testing both nominal and off-nominal scenarios together in the same study allows the nominal scenarios to serve as an experimental control, against which the off-nominal data can be assessed. With the inclusion of nominal scenarios, one can specifically compare performance on the nominal scenarios with the nominal portion of the off-nominal scenario (e.g., the time period immediately before the off-nominal event). If there is no performance difference, it allows the ability to assess that the test participant was on-task, performing appropriately at the time of the off-nominal event, and that the observed performance was directly due to the off-nominal event.

User expectancy manipulation. Conducting the evaluation test with both nominal and off-nominal scenarios allows the experimenter to manipulate the user's expectations by adjusting the relative probability of nominal and off-nominal scenarios. By defining approximately 80-90% of the test scenarios as nominal scenarios and the remaining 10-20% of the scenarios as off-nominal, the test participant will

form an expectation that the system will be working normally on each trial. This allows the experimenter to instill normal usage behavior in the participant. In this manner, then, when one of these low-probability off-nominal events occurs, the test participant was more likely to have been engaging in normal usage (i.e., attentional, visual and task allocation), and not acting as a user that was looking for failures and helping to "debug" the system. The type and severity of the off-nominal events would likely affect the relative probabilities necessary to instill normal usage behavior.

Comparative performance measurement. Including both nominal and off-nominal conditions in the evaluation test allows for the assessment of the amount of disruption caused by the off-nominal event, compared to nominal conditions. For example, if the off-nominal event was the presence of a 500 ft patch of fog during aircraft taxi, one might observe a drop in speed from 15 kts to 12 kts (a difference of 3 kts and 20%). This provides a quantitative assessment of how the system will perform in the worst-case scenarios once fielded.

Development of Off-nominal Events

The authors have successfully used these off-nominal evaluation techniques in two experiments: A part-task simulation experiment of cockpit display formats for aircraft taxi (Foyle, Hooey, Wilson & Johnson, 2002) and a full-mission simulation experiment of procedures with an advanced taxi cockpit display system (Hooey, Foyle & Andre, 2000). In developing the off-nominal events for these experiments, the authors followed the following steps, in order:

1. Determined human-system interactions that merit further investigation. There are four categories of human-system interactions to be considered:
a) Unexpected changes in environment or operations
b) Interactions with other human agents in the system
c) Interactions with other equipment or technologies
d) Failures (partial or total) in the system under evaluation

Focus groups were conducted using subject matter experts and future system users to generate a list of domain-specific examples within each category, and rate them for degree of criticality.

2. Identified psychological constructs and created off-nominal events to evaluate human-system interactions. The critical examples were analyzed, and the common underlying psychological constructs associated with them were identified. For instance, several examples shared a common concern related to complacency. Specific off-nominal events were created to assess the underlying psychological constructs, and appropriate dependent measures were determined.

3. Estimated the disruptiveness of the off-nominal events. Each event is assigned a rating of low, medium, or high based on the impact on the participant's task (i.e., potential to alter system usage, visual scan patterns, and procedures) and impact on the following test trials (i.e., potential for negatively effecting system trust and crew interactions).

4. Incorporated these off-nominal events into an experimental design. This is described later in Figure 2, incorporating the disruptiveness estimation.

Off-nominal Event Examples

In order to illustrate this off-nominal event development process, a specific example will be given for each class of off-nominal condition using the two pilot-in-the-loop simulation experiments discussed above. For each example, the human-system interaction class, underlying psychological constructs, the off-nominal event, and the estimated level of disruptiveness to the participants and the following test trials are provided.

Human-system interaction class: Unexpected changes in the environment or operations
Constructs: Situation awareness, display capture
Event: Unexpected taxiway stoplights were presented requiring a near-emergency stop and quick reaction.
Disruptiveness: Moderate. Since the detection required near-emergency braking, this could lead to high physiological arousal. In this particular case, the consequences of a miss were not very high -- if the pilot did not notice the lights, there was no consequence and the pilot was not given feedback.

Human-system interaction class: Interactions with other human agents in the system
Constructs: Complacency, levels of processing
Event: ATC issued an erroneous taxi clearance (by voice, text datalink, or graphical datalink) sending the aircraft to the incorrect concourse. If the pilots detected the clearance error, the controller corrected it; if the pilots did not detect it within 45 sec, the controller amended the clearance.
Disruptiveness: Low. The clearance was always amended at the beginning of each trial. Such amended clearances are typical in actual operation.

Human-system interaction class: Interactions with other equipment or technologies
Constructs: Complacency, trust, situation awareness
Event: Pilots taxied with an Electronic Moving Map (EMM) that depicted traffic on the airport surface. There are known limitations of surface surveillance that would cause traffic to be undetected and thus not depicted on the EMM. Such an undetected aircraft crossed in front of the test-participant requiring a braking response.
Disruptiveness: Moderate. The surprising aircraft crossing in front of the test participant led to emergency braking and higher physiological arousal. However, since the pilots attributed this error to a known limitation in surveillance systems, and not the EMM (the system under evaluation), trust in the

EMM was not affected. Although highly disruptive during the trial, the impact on subsequent trials was estimated to be only moderate.

Human-system interaction class: Failure of the system being tested
Constructs: Crew interaction and display cross-checking
Event: In a partial failure of the system, the head-up display (HUD), available only to the captain, showed a route that was incorrect and different than both the original taxi clearance and the head-down EMM (available to both the Captain and the First Officer).
Disruptiveness: High. Since only the Captain had the erroneous clearance, the possibility existed for an argument, which could affect crew communication and teaming. (In fact, this proved true: One First Officer sharply told the Captain that he needed to communicate more.) Additionally, this system error event could cause the crew to lose trust in the system, thus altering system usage on subsequent trials.

Incorporation into Experimental Design

After defining the issues, developing the off-nominal events, and estimating the amount of disruptiveness, the off-nominal events are incorporated into the test design. Figure 2 shows a generalization of the two designs used in Foyle, Hooey, Wilson and Johnson (2002) and Hooey, Foyle and Andre (2000). As can be seen in Figure 2, there are three general blocks of trials: A set of training blocks, followed by the experimental blocks containing both nominal and off-nominal scenarios, and a final single extreme off-nominal trial.

Training Blocks

Training blocks are almost always used at the beginning of a test to allow for general simulator, controls, and task familiarization. Data, if collected, are generally analyzed separately and for different purposes than the experimental trials. Incorporating off-nominal events into these training blocks depends upon the actual off-nominal event and the issues associated with it. Specifically, if the off-nominal events require familiarization, or if there is a specific required response that needs to be learned, then the off-nominal events would need to be incorporated into the training blocks. For example, the taxiway stop lights described previously were incorporated into the training blocks. This is because taxiway stop lights are not standardized lighting, and without familiarization the pilot may not realize what they are and what the appropriate response should be. Without this training, data collected during the experimental trials may be highly variable and likely invalid. For example, a measure such as "time to brake" could not be used if pilots were not trained that braking was the appropriate response.

A different off-nominal event, such as an incurring aircraft cutting across the pilot's cleared taxi route, would not need to be incorporated into the training blocks, since it is obvious and natural to slow or brake the aircraft to avoid a collision. In fact, one would not want to include this off-nominal event in training because if there were more than two or three such incurring aircraft events during the test day, the test pilot would start to anticipate these events. This cueing would alter the pilots' visual scan behavior and expectations such that the nature of the task would change from normal taxi operations to incursion avoidance. The experimenter would want to familiarize the test pilot with other traffic, however, so that the test pilot would know that other traffic are present in the simulation and what they look like.

During the training blocks, each individual trial may include both nominal conditions and procedures, as well as the off-nominal events that are presented in the training blocks. This doubling-up of nominal and off-nominal conditions during training allows for time savings, which can then be used for the experimental trials.

Experimental Blocks

As discussed above in reference to Philosophy #1, there is some truth that data from the nominal conditions needs to be "protected" in some sense. As shown in Figure 2, this is done by having separate trials for the nominal and off-nominal conditions. In this way, then, the data from the nominal conditions (e.g., speed, accuracy) can be aggregated and have not been affected by the off-nominal events (which may require speed changes or route changes that would greatly impact nominal performance data).

One must be cautious when stating that one trial does not affect another trial. Clearly, off-nominal events estimated as highly disruptive can affect performance on later trials. Off-nominal events that are low or moderately disruptive, by definition, will only minimally affect the test participant or performance on other trials. These low and moderately disruptive off-nominal events are of the type that could normally be expected to occur during actual operations, so that in some respect, these occurrences are already part of the test participant's previous experience. Thus, the off-nominal events included in the experimental blocks are only those which are estimated to be of low and moderate disruptiveness.

In the two studies referred to previously, we paired two off-nominal events within a single trial in the experimental blocks for efficiency. When doing this, placing a low disruptive off-nominal event before the moderately disruptive off-nominal event allows for less interaction between the two off-nominal events (since the more disruptive event occurs last).

Final Trial

The final trial of the experiment allows for the assessment of a highly disruptive off-nominal event. For example, the third off-nominal example given previously (the system failure where the two crew

members had different route clearances) was, indeed, of high disruptiveness. Because of the nature of the system failure, the two crew members had a negative interaction, which would have likely carried over to other trials. Also, because it was a system failure, their trust in the system on future trials would also have been affected. By testing this highly disruptive off-nominal event as the final trial, there were no other trials to be affected.

Wickens and Long (1995) and others have also used this "final trial" technique. In their experiments, they evaluated the detection of an incurring aircraft on the runway on the last trial of the experiment. This was done so that the condition was "truly surprising" and unexpected, since all other experimental trials had a cleared runway. There is similarity between the use of a final trial for a truly surprising event and a highly disruptive off-nominal event, since in both cases the following test trials would be affected. (In this example, the pilot's scan pattern after the runway incursion would change, and future incursions would no longer be "truly surprising".)

The final extreme off-nominal trial provides an opportunity to explore one very disruptive event, and can often yield very interesting observations. However, this technique often suffers from a lack of statistical power as only one data point can be collected from each test participant. As such, data often cannot be compared across experimental test conditions. Nonetheless, subjective and observational findings from these off-nominal events may provide insights into the system under evaluation, or may suggest opportunities for further research, that might otherwise not be observed.

Summary

In this paper, an experimental methodology for conducting off-nominal testing in human-in-the-loop evaluations was developed. Off-nominal testing allows for increased understanding of the human-machine system under evaluation, may uncover design issues that can be addressed, and can allow for the determination of training issues and procedures. The methodology involves developing issues to be tested, off-nominal events addressing those issues, and then estimating their disruptiveness on the test participant and the other test trials. Low and moderately disruptive off-nominal events are then incorporated into an experimental design while still allowing for the accurate estimation of dependent measures in the nominal trials. A single final trial can be incorporated for highly disruptive or truly surprising off-nominal events. The use of off-nominal testing and this proposed methodology will lead to more robust tests and evaluations, which, in turn, will improve the technical transfer success rate of systems and concepts from the laboratory to the field.

Acknowledgments

This work was supported by NASA's VAMS/SEA (Virtual Airspace Modeling and Simulation / Systems Evaluation and Assessment) and AS/AOS/HAIR (Airspace Systems / Airspace Operations Systems / Human Automation Integration Research) programs.

References

Foyle, D. C., Andre, A. D., McCann, R. S., Wenzel, E., Begault, D. and Battiste, V. (1996). Taxiway Navigation and Situation Awareness (T-NASA) System: Problem, design philosophy and description of an integrated display suite for low-visibility airport surface operations. SAE Transactions: Journal of Aerospace, 105, 1411-1418.

Foyle, D. C., Hooey, B. L., Wilson, J. R. and Johnson, W. A. (2002). HUD symbology for surface operations: command guidance vs. situation guidance formats. (Paper 2002-01-3006) Proceedings of the AIAA/SAE World Aviation Congress. SAE International: Warrendale, PA.

Hooey, B. L., Foyle, D. C. and Andre, A. D. (2000). Integration of cockpit displays for surface operations: The final stage of a human-centered design approach. SAE Transactions: Journal of Aerospace, 109, 1053-1065.

Hooey, B. L., Foyle, D. C. and Andre, A. D. (2002). A human-centered methodology for the design, evaluation, and integration of cockpit displays. Proceedings of the NATO RTO SCI and SET Symposium on Enhanced and Synthetic Vision Systems. NATO.

Leveson, N. (2001a). The role of software in recent aerospace accidents. Proceedings of the 19th International System Safety Conference, System Safety Society: Unionville, VA.

Leveson, N. (2001b). Systemic factors in software-related spacecraft accidents. AIAA Space 2001 Conference and Exposition. Paper AIAA 2001-4763. AIAA: Reston, VA.

Parasuraman, R., Hansman, J., and Bussolari, S. (2002). Framework for Evaluation of Human-System-Issues with ASDE-X and Related Surface Safety Systems. (White Paper for AAR-100). Washington, DC: Federal Aviation Administration. (Available at http://www.hf.faa.gov).

Wickens, C. D. & Long, J. (1995). Object vs. space-based models of visual attention: Implication for the design of head-up displays. Journal of Experimental Psychology: Applied, 1, 179-193.

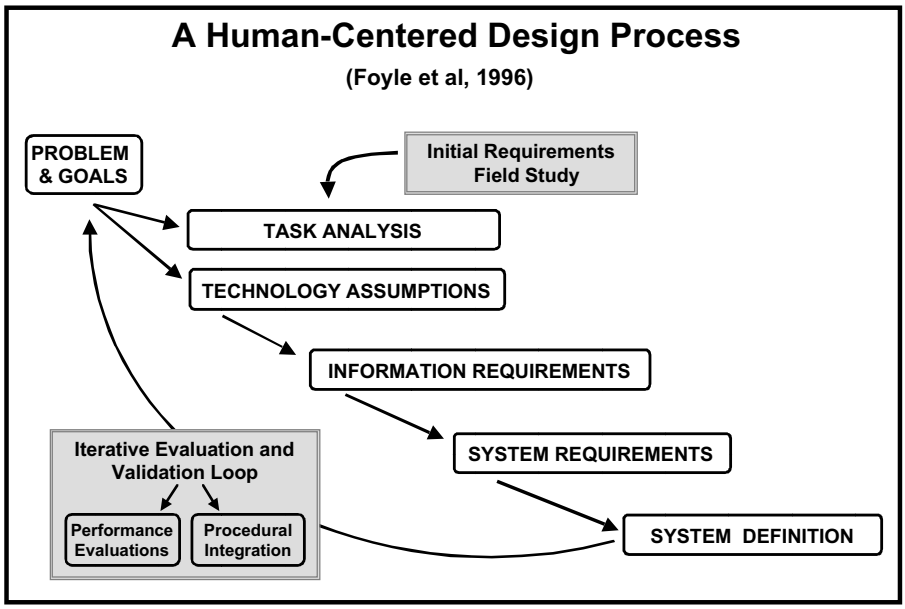

Figure 1. An iterative human-centered design process for system development. A full description of the process and its application to the development of a cockpit display suite for aircraft taxi operations (the Taxiway Navigation and Situation Awareness, T-NASA, system), is described in Foyle et al. (1996) and Hooey, Foyle and Andre (2002).

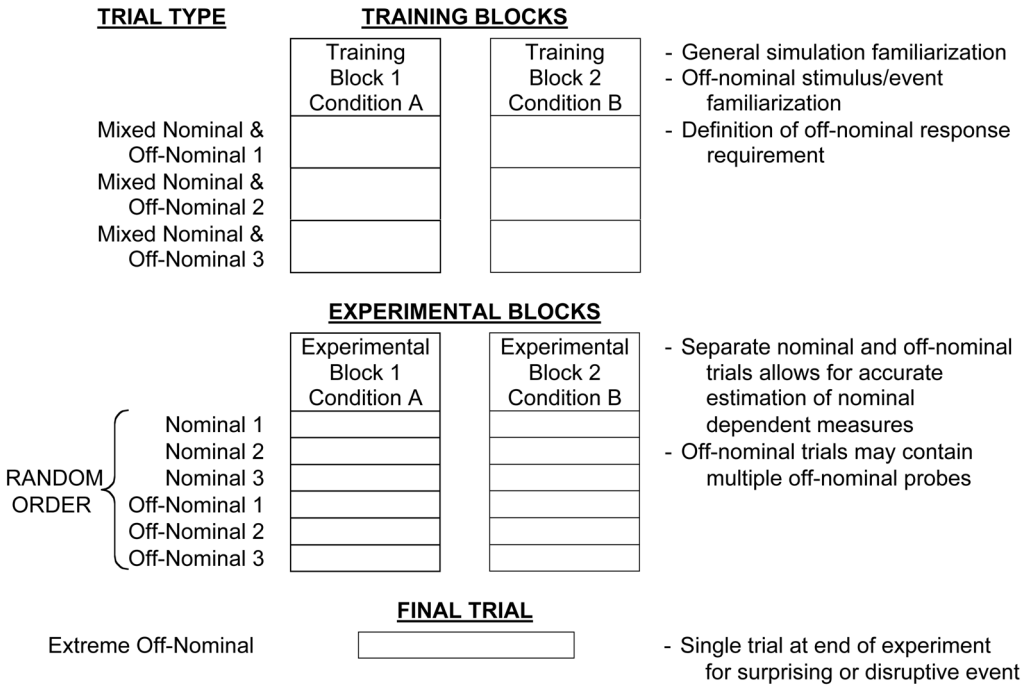

Figure 2. General schematic of nominal and off-nominal testing paradigm showing two test conditions (A and B). The number of trials shown in the training and experimental blocks is for notional purposes only. Counterbalancing of conditions is not shown for simplicity.