72nd International Astronautical Congress (IAC), Dubai, United Arab Emirates, 25-29 October 2021.

IAC-21-4-B6.4

The Path to Crew Autonomy - Situational Awareness in Scheduling and Rescheduling Tasks for Novice Schedulers

Megan C. Shyr[a], Tamsyn E. Edwards[b], Summer L. Brandt*, Jessica J. Marquez*

[a] *Department of Mechanical and Aerospace Engineering, University of California Davis, Davis, CA 95616*, mcshyr@ucdavis.edu
[b] *San Jose State University at NASA Ames Research Center, Moffett Field, CA 94035*, tamsyn.e.edwards@nasa.gov
* *NASA Ames Research Center, Moffett Field, CA 94035*, summer.l.brandt@nasa.gov and jessica.j.marquez@nasa.gov

Abstract

Technical limitations will severely restrict communication between ground-based mission support and onboard crew in future long duration exploration-class missions (LDEM). Thus, mission support tasks like mission scheduling need to be shifted to onboard crew. Currently, crew activities are scheduled over the course of several weeks by ground-based experts with years of experience-based training. These experts display extensive amounts of situational awareness (SA) throughout scheduling by maintaining a mental model of many factors such as constraints (e.g., physical space/layout), abilities and skills of the crew, and crew preferences, allowing them to anticipate and mitigate potential issues. Thus, we propose that SA is a key component in mission scheduling, and support of SA for the onboard crew is essential when mission scheduling tasks are reallocated to non-experts. In this paper, we examine SA in novice schedulers in both scheduling and rescheduling tasks in a spaceflight-like context. Results indicate that there is no significant difference between scheduling and rescheduling tasks with regards to SA in novice schedulers. Additionally, our experiment shows that novice schedulers are less able to develop sufficient SA for constraints that are dependent on more than one activity. We propose that software aids may be useful to support novice schedulers, particularly with these constraints, and may increase SA in scheduling and rescheduling tasks. This work is vital to ensure the successful transfer of mission support tasks to the crew for future LDEM.
Keywords: Situational Awareness, Scheduling, Long-Duration Exploration-Class Missions, Crew Autonomy

Acronyms/Abbreviations

- ISS: International Space Station
- LDEM: Long duration exploration-class missions
- SA: Situational Awareness
- SAGAT: Situation Awareness Global Assessment Technique
- SPAM: Situation Present Assessment Method

1. Introduction

Schedule creation in complex, dynamic environments such as the International Space Station (ISS) can take mission schedulers anywhere from weeks to months to complete [1]. Due to task complexity, scheduling often involves specialists with years of experience-based training [2]. Expert schedulers must ensure strict requirements (such as energy resources) are met and followed [2]. Additionally, they must promptly address conflicts and last-minute changes. Failure to create an effective schedule can compromise crew health and safety and jeopardize mission objectives. In the future, as technical limitations restrict Earth-to-space communication, astronauts will assume the task of scheduling and rescheduling, and must *effectively and efficiently* manage strict timelines without guidance from specialists on the ground.

Interviews with seven expert planners (current or recently retired space mission schedulers) revealed that experienced schedulers maintain a mental model of factors that are not relayed during formal training [3]. These include non-formal constraints such as physical space/layout, abilities and skills of the crew, and crew preferences – knowledge that experts say only comes from experience. Because these constraints are not formally documented, expert schedulers must not only have awareness and knowledge of these constraints, but also integrate them into their mental models when building the schedule, demanding the building and maintenance of situational awareness (SA).

For the purposes of this study, the authors define SA as "the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near futures" [4]. The level of SA that is achieved during a scheduling task indicates a distinction between expert and novice schedulers as novice schedulers would likely only be aware of formally documented constraints and activity priorities [3], preventing them from fully developing SA and integrating all the necessary information into a complete mental model. Thus, SA is a critical component of effective scheduling and is crucial for the successful transfer from experts on the ground to onboard crew [5]. However, limited data exists with regards to SA for novice schedulers in scheduling and rescheduling, especially in a space

mission context and, therefore, addressing this research gap is vital in supporting the feasibility of future LDEM.

Direct measurement of SA has most commonly been evaluated through two methods, Situation Awareness Global Assessment Technique (SAGAT) and Situation Present Assessment Method (SPAM), both with equally predictive performance [6]. These two approaches are predicated on probing participants using questions about the situation/task. They vary in that SAGAT blanks displays throughout query responses while SPAM occurs in real-time with the task displays remaining available to participants as they respond to queries [6]. SAGAT scores SA based on accuracy and classifies these scores into operationally relevant bands. Because participants cannot reference task interfaces while answering queries, accuracy is thought to quantify the situational awareness available through recall. A key feature of SPAM is that SA is predicated on working memory *and* the ability to find the correct answer in a short time frame [7], and therefore, the measure is less reliant on memory which is a criticism of SAGAT [6]. Thus, response time supersedes accuracy when evaluating SA using SPAM. Recent work [3] has developed a SPAM-based framework for measuring SA in scheduling for space-like missions. We aim to use this methodology to expand our understanding of SA in scheduling and rescheduling.

The current paper presents SA data from a ground-based, remote investigation conducted utilizing Playbook, a web-based scheduling platform [8, 9]. Participants were randomly assigned to perform either a series of scheduling tasks or a series of rescheduling tasks, after which, our SPAM-based methodology was used to probe SA. Results presented here provide an initial examination of SA in scheduling and rescheduling for novice schedulers, identifies potential barriers to establishing good SA, and informs the development of countermeasures to enhance SA for novice schedulers.

2. Method

Thirty-one participants took part in the study (18 females, 18-64 years old). All participants held a bachelor's degree or higher and were recruited on a voluntary basis. This study was approved by the NASA Ames Institutional Review Board (HRII 20-07).

2.1 Materials

This study was conducted remotely due to the restrictions necessitated by the public health crisis caused by COVID-19. Participants provided their own hardware which included a video-enabled computer and an iPad. Hardware and software versions were strictly controlled to ensure that participants were using the same technology. Participants interacted with researchers using video conferencing software and a custom experimental platform [10]. Experimental materials were presented to the participants on both their computer browser and iPad. Scheduling and rescheduling tasks were conducted using a self-scheduling software platform, Playbook [8, 9, 11]. Fig. 1 shows the remote experiment setup.

Fig. 2 shows the Playbook user interface. The Timeline displays horizontally, and schedules are made chronologically from left to right (Fig. 2A). Each crew member corresponds to a single row. Activities can either be flexible (marked with a white dot) and can be moved or inflexible (e.g., activities like meals and sleep) and cannot be moved. Activity duration is indicated by block length. The Task List, which contains additional activities not yet scheduled, can be accessed from the hamburger menu, and shows additional information including priority level (low, medium, or high) and any relevant constraint (Fig. 2B). Outside of Playbook, a list of activities is also available to participants on their computer browser. The Scratchpad facilitates transfer of activities from the Task List to the Timeline or vice versa (Fig. 2C). If the placement of an activity creates a violation (e.g., a necessary resource is unavailable), a red outline will appear indicating a violation in the plan (Fig. 2D).

2.2 Task

The study used a 4×2×2 mixed design, with one between-subjects variable (schedule or reschedule) and two within-subject variables. The within-subject variables were type of constraint (4 types) and number of constraints (2 levels). All participants completed nine trials including a baseline trial with no manipulations which was always presented first. A Latin square was used to determine the order of the remaining experimental trials. Each of the remaining eight trials

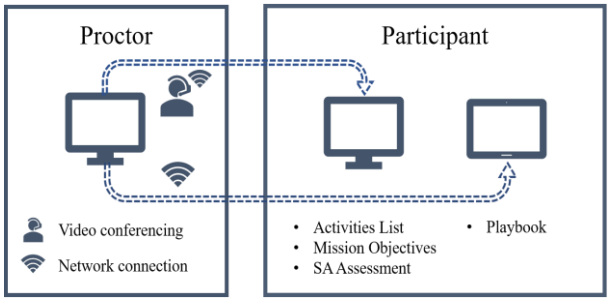

Fig. 1. A diagram of the remote experiment setup. The proctor communicated with the participant through video conferencing and pushed information to either the computer or iPad over a network connection. A comprehensive list of activities and mission objectives were provided on the computer browser as well as the SA assessment following each trial. The iPad was reserved for the scheduling/rescheduling task.

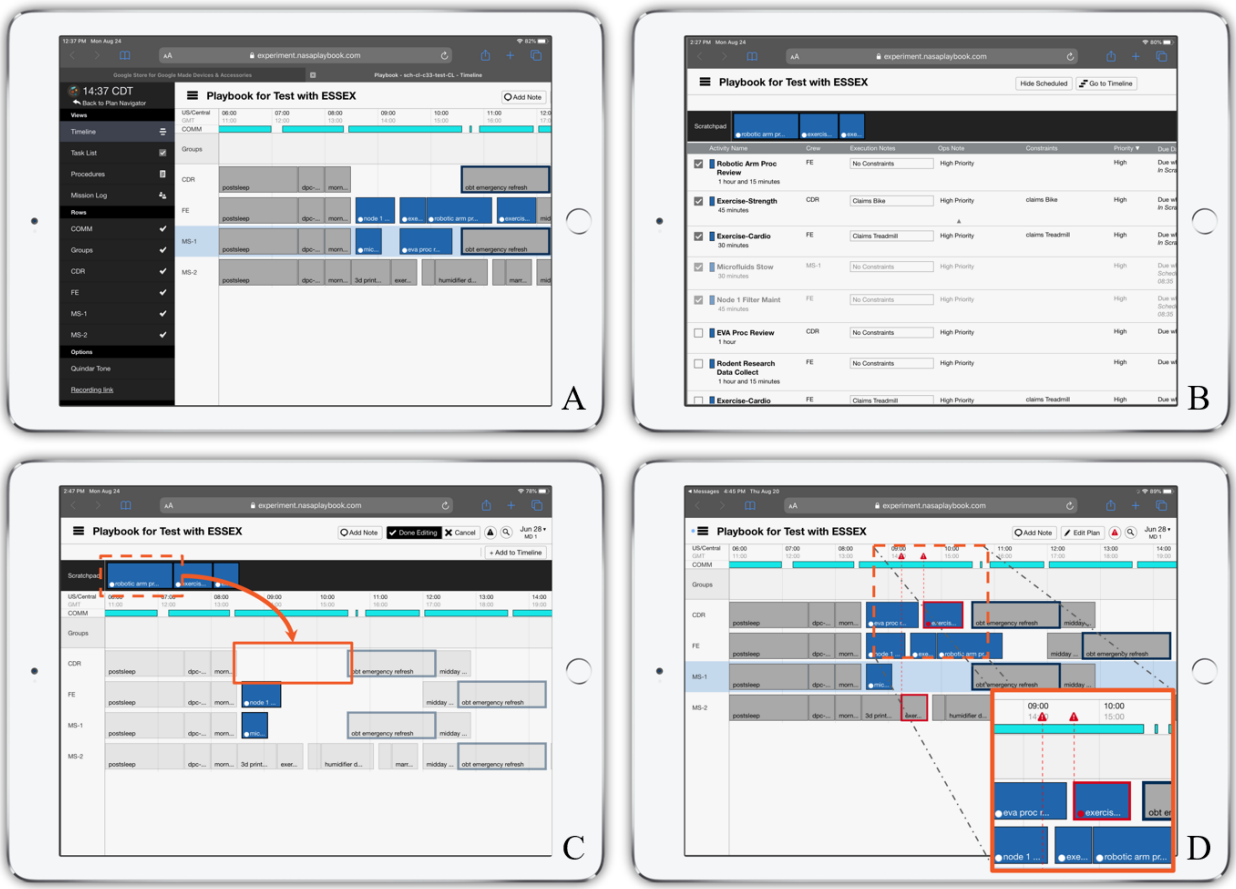

Fig. 2. The three main displays (A, B, C) for our experimental platform are as follows: A) Timeline – displays current schedule with activities arranged chronologically from left to right. Flexible activities are marked with a white dot. The hamburger menu is shown to the left and allows navigation between the displays. B) Task List – lists all activities with relevant information (i.e., priority, constraints). Selecting the check box next to an activity will transfer that activity to the Scratchpad. C) Scratchpad – facilitates transfer of activities from the Task List to the Timeline. D) provides an example of a violation. Violations are indicated by red lines around the affected activities.

consisted of only one type of constraint with either a low number (33% of total activities) or a high number (66% of total activities) of constrained activities. Type of task was assigned randomly with 15 participants completing the scheduling task and 16 participants completing the rescheduling task.

The four levels of the constraint variable were as follows:

- Time Range Constraint (T) limits the time of day an activity can be scheduled (e.g., Activity A must start no earlier than 0900 and end no later than 1030)
- Requires Constraint (R) states that the activity needs to have a particular resource available (e.g., Activity A requires communication availability)
- Claim Constraint (C) describes a specific piece of equipment required for a particular activity (e.g., Activities A and B both claim a treadmill, and therefore cannot be scheduled at the same time).
- Ordering Constraint (O) describes when an activity should be scheduled in relation to another activity (e.g., Activity A must be scheduled before Activity B)

It is worth noting that the first two constraints, T and R, apply to only one activity (e.g., Activity A), while constraints C and O describe a dependency on two activities (e.g., Activity A and Activity B). This distinction will become important in later analyses.

Prior to the start of the experimental trials, participant completed four training trials, after which they completed a competency test. A score of 77% (7 of 9 questions) was required to proceed to the nine experimental trials. During the task, participants were given the hypothetical scenario that they were crew members on a mission to a Deep Space Habitat. Their job was to schedule (or reschedule) a viable timeline for

a single day for themselves and their crew. Participants were instructed to create the timeline as efficiently (quickly) as possible while meeting the following objectives:

- Schedule as many flexible activities as possible.
- Prioritize higher priority activities when unable to schedule/reschedule all activities.
- Clear all violations from the schedule prior to trial completion.

Schedules were made on the iPad using Playbook, and mission objectives, activity information, and the SA assessments were listed/administered on the participants' computer browser (Fig. 1). Lastly, participants were told to work as quickly as they could and not try to create a perfect plan, but rather one that meets the objectives.

2.2.1 Scheduling condition

In the scheduling tasks, participants were presented with a schedule that was empty aside from inflexible (static) activities such as sleep and meals. Inflexible activities occupied 75 hours and 5 minutes of the 96 hours in the plan across four crewmembers. Twenty-four activities were available for participants to schedule with no restrictions in terms of number of movements. The breakdown of the 24 flexible activities were as follows: 8 low priority, 8 medium priority, and 8 high priority. The task was designed so that not all the activities could be scheduled, forcing participants to schedule based on mission objectives.

2.2.2 Rescheduling condition

In the rescheduling tasks, schedules included the same inflexible and flexible activities. The breakdown of the 24 flexible activities were as follows: 8 low priority, 8 medium priority, and 8 high priority. However, 18 of the 24 flexible activities were scheduled in the timeline prior to the start of the trial (8 low priority, 8 medium priority, and 2 high priority). Inflexible activities and pre-scheduled activities occupied 91 hours and 35 minutes of the 96 hours in the plan across four crewmembers. Participants were told that they must reschedule the timeline to include six new high priority activities following mission objectives. Again, there were no restrictions on number of movements, and more activities were provided then were possible to schedule.

2.3 Assessment of Situational Awareness

After participants submitted their completed schedules for each experimental trial, SA was evaluated using a modified SPAM-based methodology [3] consisting of three true-or-false questions administered at the conclusion of each trial. This technique was used to prevent any interruptions during the scheduling/rescheduling task itself, similar to [5]. Participants were asked to answer as quickly and accurately as possible using working memory but could refer to the plan they just created as needed. Trial performance metrics were logged at the conclusion of each trial before SA assessment began. Thus, our SPAM methodology was not intrusive to the task itself and did not bias participants to answer queries during periods of low workload (as could be the case in real-time administration of queries) – two common limitations of SPAM [6].

3. Results

Results include analyses with a response time cutoff and without a response time cutoff. A cutoff is typically used for SPAM because its aim is to capture information within working memory or quickly identifiable within the displays [12]. In both cases, response time should be relatively low. After a certain period, it may be inferred that participants are searching for information rather than having the information available in a mental model or knowing where to find it. In our experiment, we used a 40.5-second cutoff which falls in line with the outliers for response time of our data following the 3.29 × standard deviation (SD) outlier identification method [13]. Responses that were greater than 40.5 seconds (s) were removed from response time analyses and recoded as incorrect.

Potential causes for timeouts are outlined by Cunningham et al. [14] as 1) the operator being unable to figure out the answer, 2) the operator's workload preventing them from answering the question, 3) the context of the scenario no longer aligning with the context of the question in dynamic tasks, or 4) the question requiring more time than allowed to answer. Reasons 2) and 3) are not applicable for our experiment as participants completed the scheduling/rescheduling task prior to answering SA questions. Consequently, our participants likely exceeded the timeout for reason 1) or 4). Reason 1) indicated no SA for information in the probe [14] and reason 4) that relevant information could not be located in an acceptable time. As a result, we expected our cutoff to have a minimal effect on accuracy. Following SPAM, we also expected accuracy to be high (near ceiling) with the ability to look back at task interfaces and displays [7]. Thus, response time should be the primary dependent variable.

Overall average accuracy was 72.90% (73.70% and 72.03% for scheduling and rescheduling, respectively), but increased to 77.51% with the removal of the 40.5-second cutoff (77.97% and 77.08% for scheduling and rescheduling, respectively). Thus, no-cutoff data in our analysis could supply additional information about SA despite not typically being a part of SA evaluation methodologies. Our subsequent linear mixed effects (LME) model analyses examined both cutoff and no-

cutoff data. Cutoff results will be reported for completeness and results reported are for cutoff data unless otherwise specified. Additionally, our overall accuracy, while high, was not at ceiling. Hence, we will also include accuracy along with response time in our analyses.

LME model independent variables were type of task, type of constraint, number of constraints, and all combinations of interactions. Trial (learning effect) was a covariate. Bonferroni-corrected post-hoc tests were conducted on both factors and interactions. Dependent variables were average response time and average accuracy (*3.1 Linear Mixed Effects Model (4×2×2): Average across trials*) and response time and accuracy for individual questions (*3.2 Linear Mixed Effects Model (4×2×2): Individual questions*).

3.1 Linear Mixed Effects Model (4×2×2): Average across trials

First, we examined response time and accuracy as an average over the 3 SA questions administered following each trial. The LME model results are presented below, and Table 1 gives a summary of significant results.

3.1.1 Average response time

The LME model analyzing average SA response time for our 4×2×2 experimental design yielded a significant effect for type of constraint ($F = 4.775$, $p = 0.003$), but no significant effect for number of constraint or type of task. Additionally, the effect of learning (trial order) was not significant. Post-hoc analyses indicated a significant difference between O and R ($p = 0.003$).

With no cutoff, there was also a significant effect for type of constraint ($F = 13.534$, $p < 0.001$) and, additionally, a significant effect for trial order ($F = 5.934$, $p = 0.016$; average SA response time decreased over time). Post-hoc analyses indicated a significant difference between O and C ($p = 0.016$), O and T ($p < 0.001$), O and R ($p < 0.001$), and C and R ($p = 0.044$). Fig. 3 presents the mean and 95% confidence interval for average response time across type of constraints with and without the cutoff**.** Overall, response time increased the most for O and C when the cutoff was removed (Fig. 3). This is also evident by the additional significant differences found between O and C/T without the cutoff and indicates that participants spent more time searching for the answers to the SA questions for these two constraints.

3.1.2 Average accuracy

Average accuracy results also yielded a significant effect for type of constraint ($F = 12.602$, $p < 0.001$) and a significant interaction between type of constraint and number of constraints ($F = 5.264$, $p = 0.002$). There was no evidence of a significant effect for number of constraints or type of task. There was no significant learning effect. Post-hoc analyses indicated a significant difference between O and the other three constraints (C, T, and R; $p < 0.001$).

With no cutoff, there is again a significant effect for type of constraint ($F = 5.346$, $p = 0.001$). There was a significant difference between O and C ($p = 0.003$) as well as O and T ($p = 0.005$). Additionally, a significant effect for number of constraints ($F = 5.328$, $p = 0.022$) was identified. Interestingly, the high number of constraints condition yielded a higher accuracy (estimated means ± standard errors of 74.31% ± 2.73 and 80.92% ± 2.74 for the low and high number of constraints, respectively). This trend was seen for cutoff data as well but was not statistically significant. Again, there was no significant learning effect. Fig. 4 presents the mean and 95% confidence interval for average accuracy across type of constraints with and without cutoff**.**

Table 1. Summary of the 4×2×2 linear mixed effects model results for response time and accuracy.

		Response Time				Accuracy			
		Average		Individual Questions		Average		Individual Question	
		Cutoff	*No Cutoff*	*Cutoff*	*No Cutoff*	*Cutoff*	*No Cutoff*	*Cutoff*	*No Cutoff*
Independent Variables	Type of Constraint	$F = 4.775$ $p = 0.003$	$F = 13.534$ $p < 0.001$	$F = 7.954$ $p < 0.001$	$F = 17.555$ $p < 0.001$	$F = 12.60$ $p < 0.001$	$F = 5.346$ $p = 0.001$	$F = 12.602$ $p < 0.001$	$F = 5.080$ $p = 0.002$
	Number of Constraints						$F = 5.328$ $p = 0.022$		$F = 5.107$ $p = 0.024$
Covariates & Interactions	Type:Number					$F = 5.264$ $p = 0.002$			
	Trial		$F = 5.934$ $p = 0.016$	N/A	N/A			N/A	N/A
Post-Hoc	O – C		$p = 0.016$		$p = 0.006$	$p < 0.001$	$p = 0.003$	$p < 0.001$	$p = 0.004$
	O – T		$p < 0.001$	$p = 0.034$	$p < 0.001$	$p < 0.001$	$p = 0.005$	$p < 0.001$	$p = 0.006$
	O – R	$p = 0.003$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$		$p < 0.001$	
	C – T				$p = 0.030$				
	C – R		$p = 0.044$	$p = 0.003$	$p = 0.004$				

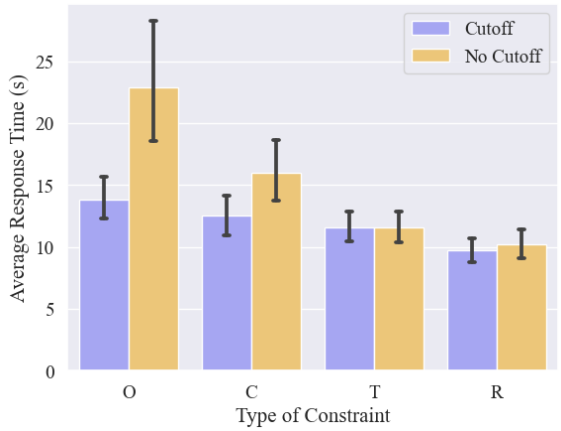

Fig. 3. The average response time (with a 95% confidence interval) by type of constraint for both cutoff and no cutoff.

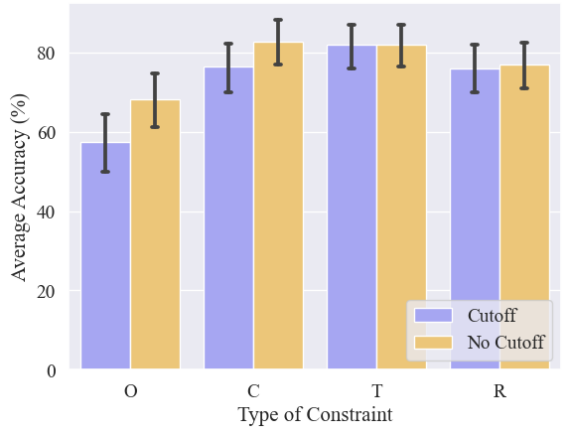

Fig. 4. The average accuracy (with a 95% confidence interval) by type of constraint for both cutoff and no cutoff.

Fig. 4 shows that accuracy improved when cutoff was removed. R and T generally both showed higher accuracy and were less affected by cutoff removal. This suggests that participants were able to answer questions effectively from working memory or quickly find (in < 40.5 s) the answer in Playbook for these two constraints. Response time increased the most for O and C with cutoff removal (i.e., were most effected by the application of the cutoff). In combination with increased accuracy, it seems that participants were able to find the correct answer given more time which motivates future efforts to implement countermeasures in Playbook to make this information more readily available, especially for these two constraints.

3.2 Linear Mixed Effects Model (4×2×2): Individual questions

The next analysis examined the effects of individual question response time and individual question accuracy in our 4×2×2 task design. LME model analyses here followed the LME model analyses outlined above. All SA questions (Table 2) will be referred to as a combination of the constraint (O, C, T, R), number of constraint (low or high), and question number (Q1, Q2, or Q3). For example, question 3 for the T constraint and low number of constraints would be referred to as T-low Q3. Question number/order was arbitrary, but consistent across participants.

3.2.1 Response time

There was a significant effect for type of constraint ($F = 7.954$, $p < 0.001$), but no evidence of a significant effect for number of constraints or type of task. Post-hoc analyses revealed a significant difference between O and T and O and R ($p = 0.034$ and $p < 0.001$, respectively). There was also a significant difference between C and R ($p = 0.003$).

Similarly, analysis without a cutoff yielded a significant effect for type of constraint ($F = 17.555$, $p < 0.001$). A significant difference was found between O and all other constraint types (C, $p = 0.006$; T and R, $p < 0.001$). C was also significantly different from T and R ($p = 0.030$ and $p = 0.004$, respectively). See Fig. 5 for a breakdown of a selection of individual question response times for cutoff and no cutoff. Fig. 1. in Appendix A provides the response times for all individual questions. Again, individual question response time increased without a cutoff most significantly with the O and C constraints. Cutoff and no cutoff trends generally align with our analyses on average response time.

3.2.2 Accuracy

There was a significant effect of type of constraint ($F = 12.602$, $p < 0.001$) for individual question accuracy, but no evidence of a significant effect for number of constraints or type of task. There were significant differences between O and all other constraint types (C, T, and R; all with a $p < 0.001$).

Similarly, with no cutoff, there was an effect for type of constraint ($F = 5.080$, $p = 0.002$) and, additionally, an effect for percent of constraint ($F = 5.107$, $p = 0.024$). O is significantly different from C and T ($p = 0.004$, and $p = 0.006$, respectively). Again, the high number of constraint condition yielded a higher accuracy (estimated means ± standard errors of 74.27% ± 0.03 and 80.95% ± 0.03 for the low and high number of constraints, respectively). Cutoff and no cutoff trends generally align with our analysis on average accuracy. Overall, O had the most questions with < 50% accuracy (3 of 6) and R had the most questions with > 75% accuracy (5 of 6). See Fig. 6 for a breakdown of a selection of individual question accuracies for cutoff and no cutoff. Fig. 2. in Appendix A provides the accuracy for all individual questions.

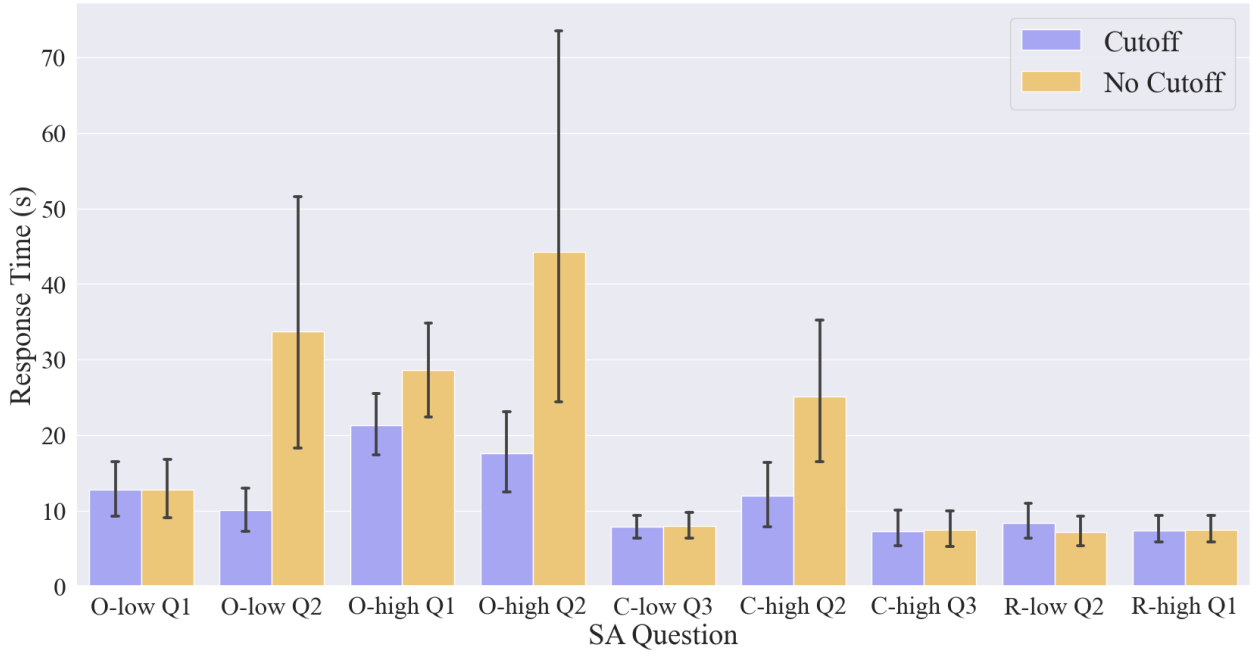

Fig. 5. The response times (with 95% confidence intervals) for a selection of individual SA questions by type of constraint for both cutoff and no cutoff.

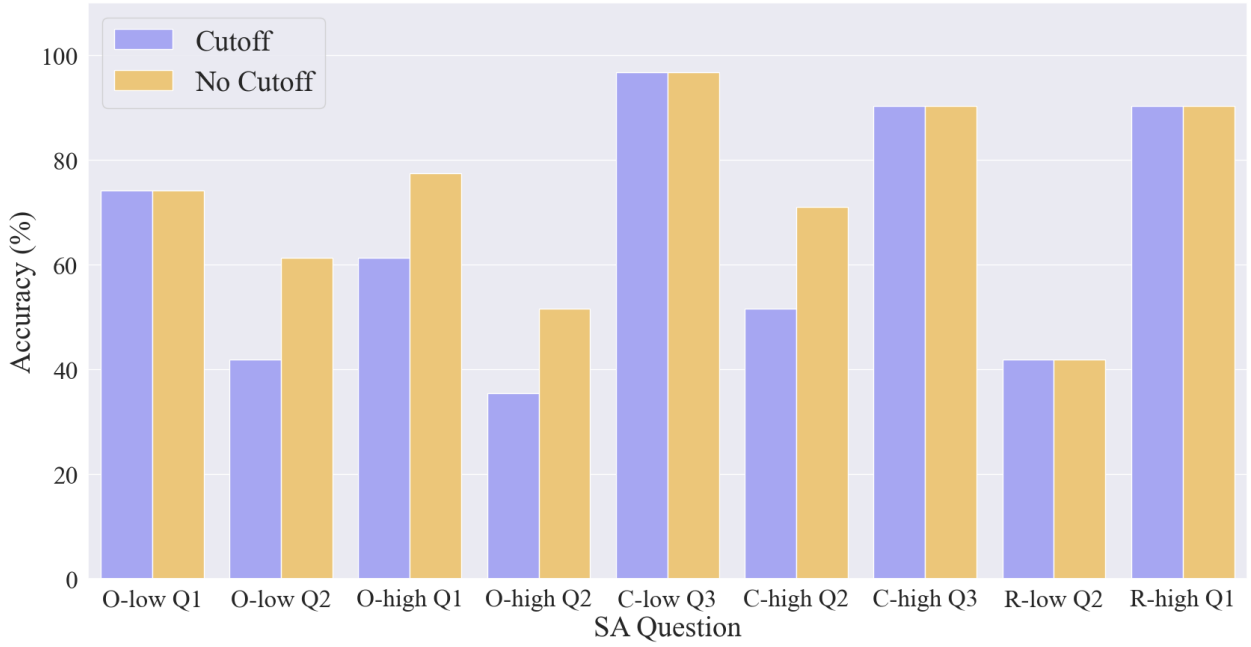

Fig. 6. The accuracies for a selection of individual SA questions by type of constraint for both cutoff and no cutoff. Note that the accuracy for each individual question is the percent of successful responses per question across participants.

Like average accuracy, individual question accuracy with and without the cutoff demonstrated a difference between incorrect responses due to response time (i.e., cutoff) and true incorrect responses. Both O and C accuracy improved with cutoff removal with 5 of 6 and 3 of 6 of the SA questions showing improvement, respectively (Appendix A, Fig. 2). Taking a closer look, O-high Q2 (Table 2) accuracy improved upon cutoff removal (Fig. 6) indicating that the necessary information to answer accurately may not be stored in a participant's mental model but does indicate that the participant could find the answer (and did find the answer in just over 40 s).

Table 2. The questions used to probe situational awareness for each trial according to type of constraint and number of constraints.

Type-Number (of Constraint)	Question Number	Situational Awareness Question
T-low	Q1	Was the earliest required start time for any constrained activity before 08:00?
	Q2	Were 50% or more of the activities in the Task List constrained?
	Q3	Were 2 or more activities constrained to be scheduled before Midday Meal?
R-low	Q1	Were there more activities with constraints in this trial than in the previous trial?
	Q2	At any time did you have 2 or more violations simultaneously in the plan?
	Q3	If COMM was suddenly unavailable after 15:00, would any activities be in violation?
C-low	Q1	If suddenly all activities that claimed a treadmill had to take place on the bike instead, would any activities in your plan be in violation?
	Q2	Did any flexible activities require a bike?
	Q3	Were all crew scheduled to exercise on both the bike and the treadmill?
O-low	Q1	Were all activities that claim a bike scheduled after a Midday Meal?
	Q2	Were any activities constrained to be scheduled before a Midday Meal?
	Q3	If a CDR's Midday Meal was moved an hour later, would that create a violation with any flexible activities?
T-high	Q1	Were 3 or more activities required to be scheduled after a Midday Meal?
	Q2	Was the latest required start time for a flexible activity after 17:00?
	Q3	Were 50% or more of the activities constrained?
R-high	Q1	If COMM was suddenly unavailable before 10:00, would any activities be in violation?
	Q2	Did 8 or more activities have a constraint?
	Q3	At any time did you have 1 or more violation(s) simultaneously in the plan?
C-high	Q1	Did all crew members you scheduled activities for get to exercise on both the bike and the treadmill?
	Q2	If suddenly all activities that claimed a bike had to take place on the treadmill instead, would any activities be in violation?
	Q3	Did any flexible activities require a treadmill?
O-high	Q1	If the first Morning Prep activity was moved an hour earlier, would that create a violation with any flexible activities?
	Q2	Were all activities that claim a bike scheduled after a Midday Meal?
	Q3	Were any activities required to be scheduled after a Midday Meal?

Table 3. Questions that exceeded the 40.5-second cutoff for both scheduling and rescheduling.

Question	Count of cutoff trials
O-low Q2	6
C-high Q2	6
O-high Q1	5
O-high Q2	5
C-high Q1	4
O-low Q3	3
C-low Q2	2
O-high Q3	1
R-low Q1	1
R-low Q3	1
Total	**34**
O	20
C	12
T	0
R	2

3.3 Linear Mixed Effects Model (2×2×2): Average response time and accuracy

The above analyses indicate that across the two task types, constraints involving more than one activity (O and C) seemed to be more difficult for novice schedulers to develop SA as they tended to have longer response times (Fig. 3), exceed the time cutoff (Table 3), and have lower accuracy (Fig. 4) relative to the other constraints (R and T). Therefore, we wanted to investigate if SA was significantly affected when scheduling or rescheduling with O and C versus R and T.

A 2×2×2 LME model analysis with O and C being classed as one constraint group (Two+) and R and T (Two-) being classed as another was conducted. Response time results showed a significant effect for group of constraint ($p < 0.001$) with Two+ having an average response time of 13.01 s ± 0.84 and Two- having an average response time of 10.52 s ± 0.83. Accuracy results showed a significant effect for group of constraint ($F = 14.361$, $p < 0.001$) with Two+ having an average accuracy of 66.93% ± 0.03 and Two- having an average accuracy of 79.05% ± 0.03. Additionally, there was a significant interaction between type of constraint and number of constraints ($F = 8.918$, $p < 0.005$). These results persisted for both response time and accuracy with and without the cutoff and supports our supposition that SA may be more difficult to develop for constraint types dependent on 2 (or more) activities.

4. Discussion

Analysis of both response time and accuracy (average and individual questions) showed type of constraint as a critical factor in SA (Table 1). These results match those found in [3] for response time. Thus, we reiterate that SA is more impacted by type of constraint than number of constraints. One interesting finding was that there was a significant learning effect for average trial response time when the cutoff was removed. Post-hoc analyses indicated that SA response time decreased as trials progressed despite trial order being randomized following a Latin square. Average accuracy did not have a significant learning effect ($F = 0.141$, $p = 0.707$), but the effect of trial order did approach significance when the cutoff was removed ($F = 3.647$, $p = 0.058$) and indicated a decrease in accuracy. However, our experiment seems to indicate that with no significant drop in accuracy across trials it is possible that SA improved in terms of response time over time or, at the very least, participants did not become resigned to the task, become fatigued, and/or start guessing.

Our accuracy analyses (both average and individual questions) hinted that percent constraint played a role as well (evident only in no cutoff data). Interestingly, the high number of constraints (66%) yielded higher accuracy. This could indicate that a higher number of constraints required participants to interact more with the timeline/activities and, thus, built more SA. This should be investigated further in future work.

We also found that average accuracy improved between cutoff (72.90%) and no cutoff (77.51%) data. Encouragingly, the ability of participants to search and successfully find the correct answers indicate that there is potential for Playbook countermeasures to help participants more efficiently locate the information they need. For example, questions O-low Q1 and O-high Q2 (Table 2) are the same (*Were all activities that claim a bike scheduled after a Midday Meal?*), however, the accuracy with the cutoff was 74.19% and 35.48%, respectively. Without the cutoff, the accuracy for O-high improves to 51.69% while O-low remains the same, indicating that it may just take participants longer to track down the answer for a constraint-related question when the trial has a higher number of constraints (i.e., 66% versus 33%). In addition to demonstrating some level of SA, accuracy improvement with cutoff removal also indicates that question comprehension is not a limiting factor.

A deeper dive into the questions themselves provided more insight into how software aids should support novice schedulers locate the information necessary for SA. We began by examining whether there were any questions beyond finding the answer for (i.e., low accuracy and low response time). This trend can be seen for R-low Q2 (Table 2) where extra time

(cutoff removal) did not help improve accuracy. Participants were asked: *"At any time did you have 2 or more violations simultaneously in the plan?"* and results suggest they either know the answer or they guessed (but did not try to search). Markedly, this question was not one that could be easily determined from the Playbook interface. Perhaps a violation count would be useful as a quick reference to track constraints during self-scheduling, especially as task complexity increases. This feature could be extended by allowing schedulers to click on the violation count and see a drop down of the violations themselves. They could then click on a specific violation to jump to the relevant location in the timeline. Doing so would allow schedulers to keep tabs on the number of violations they have at any point during scheduling and prevent them from having to search the timeline for the red lines indicating a violation. Additionally, a count would provide an easy check for confirming they have resolved all violations at the conclusion of the task.

As we indicated in our results, it also appears that participants struggled to develop SA for constraints involving more than one activity. Thus, countermeasures should specifically focus on these constraints. Q2 for both O-low and C-high appear to be challenging questions because results with the cutoff yielded low accuracy and low response time (participants likely guessed) and results with no cutoff yielded low accuracy and high response time (participants could not find the answer). Both questions tasked participants with either being able to recall or determine all activities that were constrained and how, and in the case of C-high Q2, quickly determine how activities may be affected if certain constraint conditions changed. Countermeasures to alleviate the difficulty of these questions could be color-coding activities by constraint type or dependency (i.e., scheduling of activity A depends on scheduling of activity B), including icons to indicate the "claims" (e.g., bike or treadmill) required for the C constraint, or perhaps a method to speed up the search for constraint specifics such as pop-ups when participants hover over activities on the timeline or quick filtering options to pick out activities by constraint details (e.g., activities that claim a bike). These questions contrast with questions that demonstrate SA (high accuracy and low response time) like R-high Q1, C-low Q3, and C-high Q3 as well as questions like O-high Q1 that participants successfully searched for the answer.

5. Conclusion

Future LDEM will put technical limitations on communication between mission support and crew, and as a result, require the transfer of mission support tasks to the crew. Previous work indicates that there is a relationship between situational awareness and effectiveness for scheduling, but SA for novice schedulers in a space mission scheduling task has not been fully explored. Our current work provides initial steps for understanding SA in novice schedulers in both scheduling and rescheduling tasks in a space context. Our results indicated that there is no significant difference between scheduling and rescheduling for the development of situational awareness in novice schedulers. However, results seem to indicate that SA is more difficult to develop for constraints involving multiple activities. Thus, future software aids should be implemented to better facilitate SA, especially with these constraints in mind. Enhancing SA in novice schedulers will enable the shift from expert schedulers to crew and move future research forward on the path to crew autonomy.

Acknowledgements

Research funding was provided by the NASA Human Research Program's Human Factors and Behavior Performance Element (NASA Program Announcement number 80JSC017N0001-BPBA) Human Capabilities Assessment for Autonomous Missions (HCAAM) Virtual NASA Specialized Center of Research (VNSCOR) effort. The authors would like to thank Candice Lee, Casey Miller, and John Karasinski for their assistance in collecting and analyzing this research data.

References

[1] R. C. Dempsey, Ed., *The International Space Station: Operating an Outpost in the New Frontier*. National Aeronautics Space Agency, 2018. Accessed: Sep. 08, 2021. [Online]. Available: http://www.nasa.gov/connect/ebooks/the-international-space-station-operating-an-outpost

[2] J. Barreiro, G. Jones, and S. Schaffer, "Peer-to-peer planning for space mission control," in *2009 IEEE Aerospace conference*, Mar. 2009, pp. 1–9. doi: 10.1109/AERO.2009.4839709.

[3] T. Edwards, S. L. Brandt, and J. J. Marquez, "Towards a Measure of Situation Awareness for Space Mission Schedulers," in *Advances in Neuroergonomics and Cognitive Engineering*, Cham, 2021, pp. 39–45. doi: 10.1007/978-3-030-80285-1_5.

[4] M. R. Endsley, "Design and Evaluation for Situation Awareness Enhancement," *Proc. Hum. Factors Soc. Annu. Meet.*, vol. 32, no. 2, pp. 97–101, Oct. 1988, doi: 10.1177/154193128803200221.

[5] C. Lee, J. Marquez, and T. Edwards, "Crew Autonomy through Self-Scheduling: Scheduling Performance Pilot Study," in *AIAA Scitech 2021*

Forum, American Institute of Aeronautics and Astronautics, 2021. doi: 10.2514/6.2021-1578.

[6] M. R. Endsley, "A Systematic Review and Meta-Analysis of Direct Objective Measures of Situation Awareness: A Comparison of SAGAT and SPAM," *Hum. Factors*, vol. 63, no. 1, pp. 124–150, Feb. 2021, doi: 10.1177/0018720819875376.

[7] F. Durso *et al.*, "Expertise and chess: A pilot study comparing situation awareness methodologies," Jan. 1995.

[8] J. J. Marquez, S. Hillenius, B. Kanefsky, J. Zheng, I. Deliz, and M. Reagan, "Increasing crew autonomy for long duration exploration missions: Self-scheduling," in *2017 IEEE Aerospace Conference*, Mar. 2017, pp. 1–10. doi: 10.1109/AERO.2017.7943838.

[9] J. J. Marquez, G. Pyrzak, S. Hashemi, K. McMillin, and J. Medwid, "Supporting Real-Time Operations and Execution through Timeline and Scheduling Aids," in *43rd International Conference on Environmental Systems*, American Institute of Aeronautics and Astronautics, 2013. doi: 10.2514/6.2013-3519.

[10] B. Kanefsky, J. Zheng, I. Deliz, J. J. Marquez, and S. Hillenius, "Playbook Data Analysis Tool: Collecting Interaction Data from Extremely Remote Users," in *Advances in Usability and User Experience*, Cham, 2018, pp. 303–313. doi: 10.1007/978-3-319-60492-3_29.

[11] J. J. Marquez, S. Hillenius, J. Zheng, I. Deliz, B. Kanefsky, and J. Gale, "Designing for Astronaut-Centric Planning and Scheduling Aids," *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 63, no. 1, pp. 468–469, Nov. 2019, doi: 10.1177/1071181319631386.

[12] F. T. Durso, M. K. Bleckley, and A. R. Dattel, "Does Situation Awareness Add to the Validity of Cognitive Tests?," *Hum. Factors*, vol. 48, no. 4, pp. 721–733, Dec. 2006, doi: 10.1518/001872006779166316.

[13] B. G. Tabachnick and L. A. Fidell, *Using Multivariate Statistics, 4th Edition*, 4th ed. Allyn & Bacon, 2001. Accessed: Sep. 08, 2021. [Online]. Available: https://www.pearson.com/content/one-dot-com/one-dot-com/us/en/higher-education/program.html

[14] J. C. Cunningham *et al.*, "Measuring Situation Awareness with Probe Questions: Reasons for not Answering the Probes," *Procedia Manuf.*, vol. 3, pp. 2982–2989, Jan. 2015, doi: 10.1016/j.promfg.2015.07.840.

Appendix A

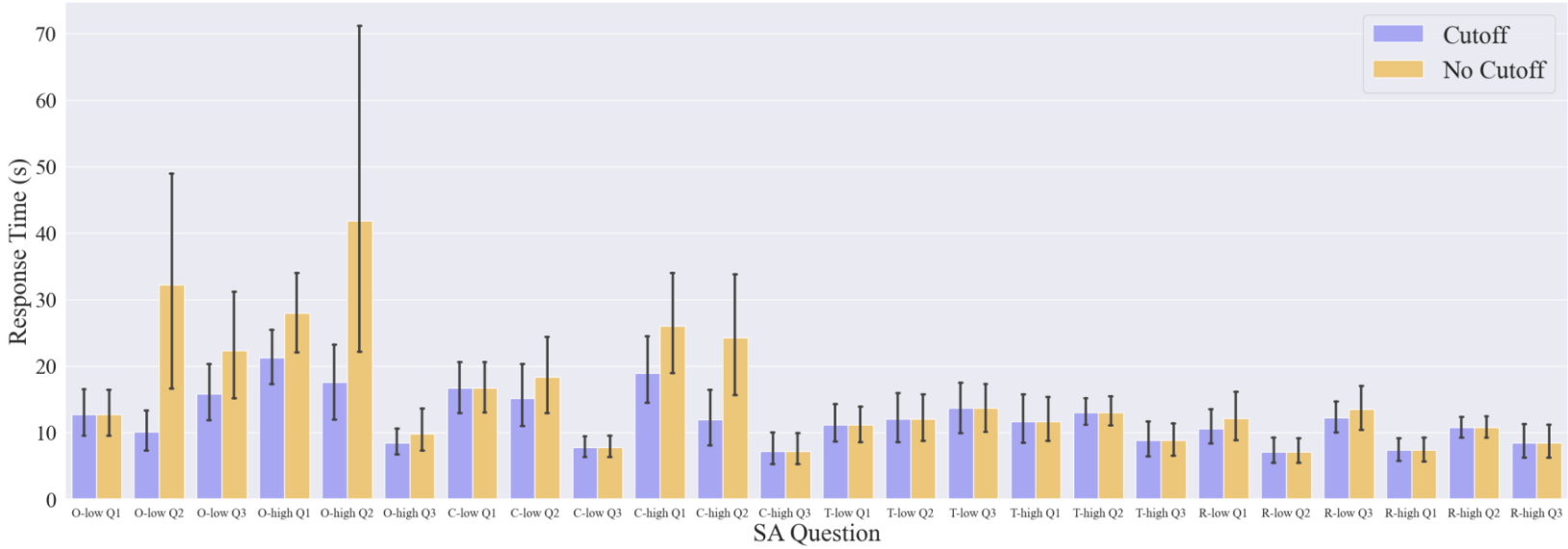

Fig. 1. The response times (with 95% confidence intervals) for all individual SA questions by type of constraint for both cutoff and no cutoff.

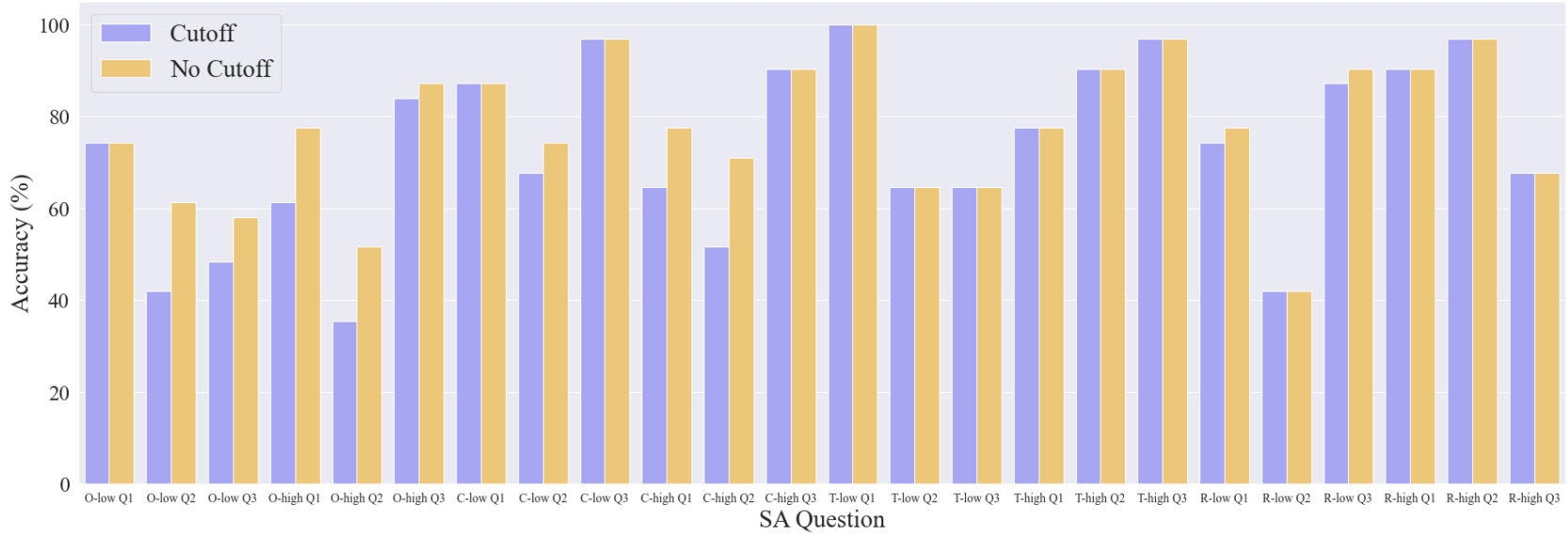

Fig. 2. The accuracies for all individual SA questions by type of constraint for both cutoff and no cutoff. Note that the accuracy for an individual question is the percent of successful responses per question across participants.