Crew Autonomy through Self-Scheduling: Scheduling Performance Pilot Study

Candice N. Lee[1],
San Jose State University/NASA Ames Research Center, San Jose, CA, 95112, USA

Jessica J. Marquez[2]
NASA Ames Research Center, Mountain View, CA, 94035, USA

Tamsyn E. Edwards[3]
San Jose State University/NASA Ames Research Center, Mountain View, CA, 94035, USA

Within the domain of human spaceflight, crew scheduling for International Space Station (ISS) remains a human-driven planning task. Large teams of flight controllers (called Ops Planners) spend weeks creating violation-free schedules for all crewmembers. As NASA considers long-duration exploration missions, the necessary shift of scheduling and planning management from Ops Planners to crew members requires significant research and investigation of crew performance to complete these scheduling tasks. This pilot study was conducted to evaluate non-expert human performance for the task of planning and scheduling, focusing on scheduling problems that increased in complexity based on the number of activities to be scheduled and the number of planning constraints. Nine non-expert planners were recruited to complete scheduling tasks using Playbook, a scheduling software. The results of this pilot study show that scheduling performance decreased as scheduling workload (i.e. number of activities and percent of activities with planning constraints) increased. This paper provides evidence towards developing a model for scheduling task difficulty and identifies potential implications for future automated aids for flight crew scheduling.

I. Nomenclature

χ^2	= Chi-Square
%	= Percent
act	= Activities
ANOVA	= Analysis of Variance
BASALT	= Biologic Analog Science Associated with Lava Terrains
con	= Constraints
DV	= Dependent Variable
ESSEX	= Environment for Self-Scheduling Experiment
HERA	= Human Exploration Research Analog
ISS	= International Space Station
IV	= Independent Variable
M	= Mean
MCC	= Mission Control Center
NASA	= National Aeronautics and Space Administration

[1] Research Assistant, Human Systems Integration Division
[2] Human Systems Engineer, Human Systems Integration Division, AIAA Member
[3] Senior Research Associate, Human Systems Integration Division, AIAA Senior Member

NASA TLX = NASA Task Load Index
NEEMO = NASA Extreme Environment Mission Operations
rs = Correlation Coefficient
SD = Standard Deviation

II. Introduction

As NASA considers long-duration exploration missions, it is envisioned that crew will behave more autonomously as compared to low-Earth orbit missions. It is expected that as missions operate further from Earth, communication latency between the spacecraft and Mission Control Center (MCC) will increase, thus shifting mission control tasks to the crew. In this space environment, this shift in tasks requires tools that enable crew members to have some level of autonomy over their own schedules. By providing crew the means to self-schedule, or reschedule their own timeline, they can minimize the idle time as they wait for MCC to respond or react to a delay in activity execution (Marquez et al., 2017).

However, it is essential that the modified schedules meet all the constraints and requirements necessary for critical spaceflight operations such as creating violation-free plans. Operational Planners are ground flight controllers with years of experience creating an astronaut's schedule and usually spend weeks creating a schedule that meets all of the program's requirements, spacecraft's constraints, and crew members availability and ability (Barreiro, Jones and Schaffer, 2009). Astronauts do not have this experience, nor do they have insight to the dozens of constraints and requirements that must be met; therefore, future self-scheduling needs to support naive planners and help them create violation-free timelines without imposing additional workload to an already overly-subscribed astronaut.

Research is sparse in the domain literature regarding crew performance for self-scheduling (as opposed to Operational Planners); specifically, non-expert human performance for the task of planning and scheduling has not been characterized experimentally. Therefore, the current research aims to address this gap. It is envisaged that this objective will be met through a series of experiments between 2019-2023. This extended abstract presents an initial study that was conducted to investigate self-scheduling performance as a function of plan complexity for naive planners.

III. Methodology

A. Design

The controlled pilot experiment reported in this paper consisted of a total of nine participants who completed scheduling tasks using the scheduling software, Playbook (Marquez et al., 2013). Participants were naive to scheduling tasks and the Playbook software. The sample was self-selected, as participants volunteered to take part. The study utilized a 3x3 design. The independent variables selected for this experiment were the number of flight crew activities to be scheduled, and the percentage of activities to be scheduled that had constraints. Participants received an overview of Playbook and four training sessions that focused on the tools to be used to self-schedule using the playbook software. The training sessions totaled approximately 20 minutes. Participants were then tasked to complete nine trials of scheduling planning problems using the Playbook software on an iPad. Several dependent variables were collected. For the sake of clarity, a subset of these metrics will be reported within the following paper: plan efficiency, plan effectiveness, and workload.

B. Aims

The aims of the research were as follows: (1) understand the effect of the number of activities to be scheduled on human performance, (2) understand the effect of one type of temporal constraint, with one temporal constraint per activity, and (3) identify the average duration to schedule one activity.

B. Scheduling in Playbook

Playbook is a web-based scheduling software used to enable crew self-scheduling (i.e. editing or composing part of their own timeline). Playbook has been used in dozens of analog missions such as NEEMO (NASA Extreme Environment Mission Operations), HERA (Human Exploration Research Analog), and BASALT (Biologic Analog Science Associated with Lava Terrains). Playbook allows crew to visualize timelines, track execution of scheduled activities, as well as complete self-scheduling (Marquez et al., 2013). A few examples of self-scheduling tasks done include re-assigning multiple activities to meet operational priorities and making scheduling changes (Marquez et al., 2017). Thus, this study aims to collect self-scheduling performance data using Playbook.

In Playbook (Fig. 1), an *activity* is a task represented as a block; the length of a block is directly related to the duration of an activity. Each activity has a known duration (expected length of time to complete said task). By convention, any one activity has a duration which is a multiple of 5-minute intervals (which is consistent with International Space Station (ISS) planning and scheduling operations). Each activity is colored with one of four colors selected for this pilot study and contains the activity name on the block. The *task list* a list of activities that are available to schedule. Activities that can be scheduled by the user are called *flexible activities.* For this study, the Task List view in Playbook displays the activity color, priority level, and a description of its associated constraint. The *Scratchpad* sits near the top of the Playbook interface (black horizontal bar in Fig. 1) and facilitates the ability to move activities between the Task List and the Timeline. The *timeline* is the final scheduling or problem space. The Timeline view displays time horizontally (from left to right) and includes information such as the mission day, hour of the day, and crew member assignment. Four (4) horizontal rows (or crew bands) fit across the timeline, one row for each crew member assignment. Any crew member can be assigned to any activity at any time. Some activities had *constraints*. A constraint is a rule, or a requirement associated with the activity. The constraint used for this experiment was a temporal constraint expressed by “Must start after 00:00” or “Must start before 00:00” (where 00:00 was determined by the experimenter). If scheduled and the activity’s constraint was not met, a *violation* would be indicated in the interface. For this pilot study, each activity was assigned a scheduling *priority*: high, medium, and low. The number of high/medium/low activities was evenly distributed for each trial.

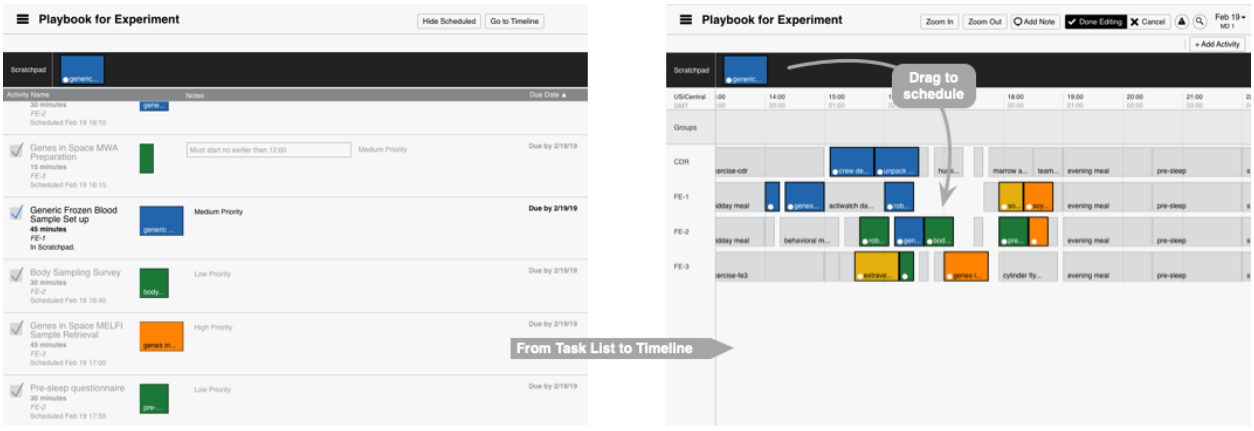

Fig. 1 Scheduling an activity in Timeline from the Task List in Playbook.

In this study, each activity was given a specific duration, in minutes (ranging between 10 minutes - 2.5 hours). The duration for each activity was selected for experimental design purposes, but also reflected typical durations in an operational environment. Each activity block was assigned one of four colors. Colors were allocated

arbitrarily and were intended to create contrast between each activity block. For the purposes of the experiment, there was a set of inflexible, operationally relevant, activities (colored grey) in the timeline, strategically placed on the timeline to facilitate the problem space. The activities' duration and name varied arbitrarily.

The pool of activities available in the Task List corresponded with the number of activities to be scheduled within each condition. Regardless of the number of activities, it was expected that not all the activities could fit in the Timeline. In order to schedule the activity, the user selects the activity in the Task List, which adds it to the Scratchpad. Once in the Scratchpad, the user can navigate to the Timeline and drag the activity from the Scratchpad to the Timeline.

C. Independent variables

Considering the aims of this initial study, two independent variables (IV) were selected: (1) the number of activities and (2) the percentage of total number of activities that had constraints. Within these two IV's, three (3) levels of each were determined, resulting in a 3x3 design. The first IV (number of activities) levels are 12, 24, & 36 activities. Similarly, the second IV (percentage of total number of activities that had constraints) levels are 0%, 33%, and 66% condition.

D. Study Participants

Nine Playbook-naive and non-expert planners were recruited at NASA Ames Research Center for this pilot study. Participant age ranged from 18 to 34 years of age, and education levels ranged from being enrolled in a degree program to having completed a Master's degree.

E. Study Materials

An electronic demographic survey was developed to capture participant data. The survey contained four questions pertinent to the study: age, gender, highest degree achieved and experience with iPad usage. Participants were given an informed consent paper-based form to sign prior to the start of the experiment. Participants were informed of their rights including the right to withdraw at any time, confidentiality of data and anonymity in reporting. Standardized instructions were developed which were read by the researcher to each participant and was presented with a paper copy as well. Lastly, the NASA TLX (Task Load Index) iOS application was used on an iPad to deliver the subjective workload scales.

F. Training

Playbook training slides were created to facilitate participant edification on the key functions of Playbook in order to complete the tasks. The training slides were designed to have the participant view an iPad screenshot screen of Playbook, with callouts describing critical areas of the interface and their functions. The functions and components described include editing the plan, flexible activities, how to move activities, how to add activities from the task list, how to zoom into the interface, and saving/cancelling plan edits.

G. Equipment

The equipment used to facilitate the study was two laptop computers, two iPads, and Playbook software. The software used for data collection included the NASA TLX iOS application for the iPad and Google Forms to administer the surveys and questionnaires. Additionally, a custom platform called ESSEX (Environment for Self-Scheduling Experiment) was developed in order to execute Playbook experimental trials and was used to collect data. ESSEX requires two browsers: one to show questionnaires and instructions to the participant and another to show the different trials (i.e., Playbook plan instances).

In the physical study set up, the participant was seated at a table with one iPad on their right-hand side to use Playbook, and one laptop computer on their left-hand side to view the list of activities and delivery of other instructional materials. Directly placed in front of them was one sheet of printed instructions. The researchers were seated near the participant with one laptop for conducting the experiment and one iPad by their side which had the NASA TLX survey as shown in Fig. 2.

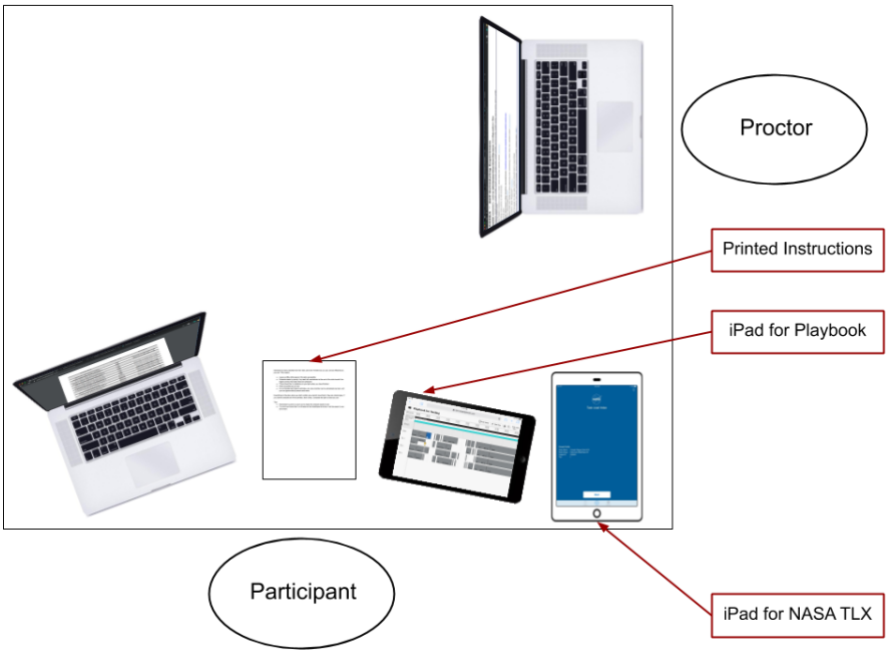

Fig. 2 Experiment configuration with participant and experiment proctor.

H. Training Protocol

Prior to starting the experimental trials, each participant was given four practice self-scheduling problems after viewing the training slides. The individual practice problems were designed to highlight the main functions necessary to complete the main experiment tasks. The purpose of these four specific problems were to allow for practice, reduce possible learning effects, and to clarify any questions or concepts regarding Playbook or scheduling there may have been. The proctor observed the participants complete the practice problems and gave comments and/or corrections when necessary to ensure the correct interpretation of the instructions, correct navigation of Playbook, and the correct solution was achieved.

I. Experiment Instructions

Participants were instructed to schedule as many activities from the Task List into the Timeline as quickly and as efficiently as they could. “Efficiency” was described as:

- Leaving as little white space in the plan as possible;
- Scheduling activity by priority (any activity left unscheduled at the end of the trial should not be a higher priority than activities that are scheduled);
- Resolving all violations from the plan prior to finishing trial (violations are indicated by a red outline around the activity);
- Not stacking or double banding activities; and
- Working as quickly as they could.

The standardized instructions also informed participants that the grey activities on the Timeline were inflexible and could not be rescheduled, and that activity colors and names were not pertinent to the scheduling task. In addition, participants were told that scheduling all of the activities may not be possible and that they should complete the plan as best as they could.

J. Metrics

1. *Workload.* The NASA TLX iOS application (on an iPad) was used to assess participant's workload after each trial. The NASA TLX consists of two parts: a pairwise comparison and weighted scales. These two parts use the same six subjective subscales: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. The pairwise comparisons collected account for differences in a rater's workload definition and sources of workload between tasks (The pairwise comparisons were not collected and only the subscale ratings were reported in this study. The NASA TLX was administered between trials which prompted the participant to rate how much of each subscale of the measure was required to complete the task.
2. *Plan Effectiveness*. Human performance for the task of self-scheduling is measured by how well and how fast the task was completed. Data was collected in order to specify several metrics related to the measure of plan effectiveness, i.e., how "good" was the created plan. Data collected through ESSEX allowed the following metrics: margin, number of activities left unscheduled, and number of violations. Margin, or the amount of white or empty space in the Timeline, is a measure of the sum of duration left in the Timeline after self-scheduling. The number of activities unscheduled would indicate how many activities were able to be scheduled into the Timeline. The number of violations created are measured throughout the trial as well as after the trial had been completed. Participants were instructed to leave no violations in the plan after self-scheduling; a plan left with violations would be considered not valid. Each of these metrics are expanded further in the Results.
3. *Efficiency*. Human performance for the task of self-scheduling is measured by how well and how fast the task was completed. Data was collected in order to specify several metrics related to the measure of efficiency, i.e., how quickly the schedule was created. Data collected through ESSEX allowed the following metrics: time on task and time to resolve violations. Time on task was measured as a key indicator of scheduling efficiency, starting when the participant stated they were ready and stopping when the participant stated they were done. Time to resolve violations was collected starting from the time a violation was created and when that same violation was resolved.

K. Experiment Protocol

The controlled experiment took place at NASA Ames Research Center. Participants were brought individually to the experiment room. Once participants read through and signed the consent form, a general overview and instructions of the experiment was verbally provided by the researcher. Participants were then instructed to complete the demographic survey and complete Playbook training. After the training and practice plans were completed, the proctor verbally stated the instructions for the experimental trials and a total of nine trials were executed in a previously determined randomized order. Every participant completed the experiment in this same predetermined randomized order. Between trials, the participants were verbally reminded of the instructions of the task.

At the start of each trial, subjects were presented with a list of activities (Appendix A) that they were required to schedule on the computer screen on their left hand side. The researcher verbally stated the instructions and launched the experimental plan on the participant's iPad through ESSEX to begin the self-scheduling task. After the participant had informed the researcher of their completion, a multiple choice question was presented. After they had answered the question, the researcher handed the participant the second iPad with the NASA TLX application so that workload measures could be taken. When the participant completed the scales and returned the iPad to the researcher and the next trial would begin. At the end of the experiment, the participant was provided with a debrief that contained the researchers' details.

III. Results

A. Analysis strategy

Results were analyzed using descriptive and inferential statistics. Violations of parametric assumptions were investigated using the Kolmogorov-Smirnov test for normality and, when appropriate, Mauchley's test for sphericity. Effects of the independent variable on dependent variables were investigated using inferential statistics, specifically, repeated-measures, parametric ANOVA, and Friedman's ANOVA when the assumptions of normality were violated. Post hoc tests were conducted for results that were statistically significant. Finally, relationships

between the independent and dependent variables were further investigated using Pearson and Spearman's correlation analysis.

B. Workload

Workload was inferred from the NASA TLX scale. Since pairwise data was not collected, an overall workload score was determined by calculating the averages of all six subscales across each condition to obtain an average workload score (Moroney, Biers, Eggemeier & Mitchell, 1992) (Fig. 3). Combined with a review of descriptive statistics, Fig. 4 shows that the lowest average workload was recorded when participants had 12 tasks to schedule (M= 26.57, SD=16.29) and when no temporal constraints were applied. An increased number of 24 tasks to be scheduled was associated with a greater average workload (M=30.37, SD=18.53). However, the average workload reported when participants had 36 (M=30.92, SD=19.1) tasks to schedule and no temporal constraints was lower than the average workload reported with 24 tasks. An interesting point to note is that the reported workload appears to become more variable as the number of tasks increase, as shown by increasing standard deviations. This may suggest that there were greater individual differences in perceived workload as the number of tasks to be scheduled increases.

Average perceived workload for all task conditions, when 33 percent of tasks had temporal constraints, increased compared to the no temporal constraints condition, suggesting that there may have been an effect of temporal constraints on average perceived workload. However, the same pattern that was observed in the no temporal constraints condition was also seen in the 33 percent constraint condition for all task conditions (12, M=27.78, SD=16.38; 24, M=41.39, SD=15.54; 36 M=33.98, SD=20.10), with 24 tasks associated with the greatest average workload. However, self-reported workload for the condition with 12 tasks with 33 percent constraints (M=27.78, SD=16.38) was only marginally higher than the 12 tasks condition with no temporal constraints (M=26.57, SD=16.29), suggesting that the temporal constraints did not have a large impact on this particular condition. Taken together, these results suggest the possibility of an interaction effect between task number and percentage of tasks with temporal constraints on average reported workload.

An interesting observation from Fig. 3 is that in the condition with 66 percent of tasks with temporal constraints, scheduling 12 (M=30.55, SD=17.23) and 24 (M=42.22, SD=19.76) tasks was reported to be slightly less workload on average than scheduling 12 and 24 tasks in the 33 percent temporal constraints condition, potentially suggesting that the increase in temporal constraints did not have a large effect on reported workload. However, the average workload for the 66 percent constraint condition when scheduling 36 tasks (M=45.55, SD=13.45) was on average, higher than either the 12 or 24 task conditions. This finding again suggests the possibility of an interaction effect between the number and constraints on reported workload.

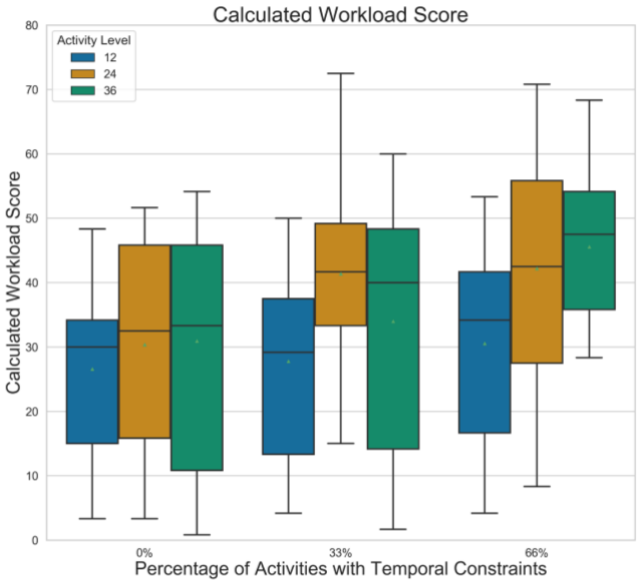

Fig. 3 Workload scores for each condition. A repeated measures ANOVA was conducted on average reported workload[1] for all conditions.

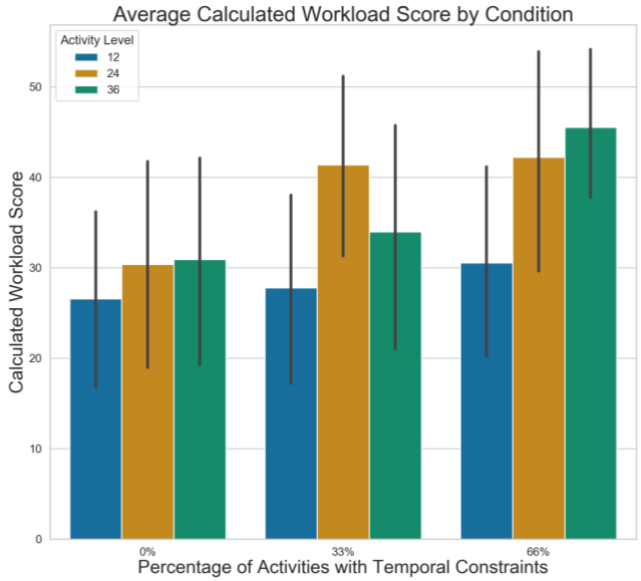

Fig. 4 Average workload scores for each condition.

A significant main effect of the number of activities to be scheduled was found on average reported workload ($F(2,16)=5.97$, $p<0.05$). Pairwise comparisons revealed that, on average, workload was significantly lower in conditions with 12 activities than 36 activities ($p<0.05$). There was no significant difference identified between 12 and 24 activities ($p=0.12$) or 24 and 36 activities ($p=1$). There was also a significant main effect of the number of tasks with constraints on average reported workload ($F(2,16)=17.36$, $p<.001$). Pairwise comparisons revealed that on average, workload was significantly lower in conditions with 0% activity constraints than 66% activity constraints ($p<0.005$). Additionally, workload was significantly lower in conditions with 33% constraints compared to 66%

[1] Analysis for workload's six dimensions were conducted but the results were similar to the unweighted average workload score and thus, not reported in the main text of this report.

constraints (p<0.005)There was no significant difference identified between 0% constraints and 33% constraints on overage reported workload.(p=0.12). No significant interaction between constraints and number of activities was identified for reported workload.

C. Plan Effectiveness

Plan effectiveness was inferred from several independent metrics: Leftover space in the schedule (termed 'margin' throughout this paper) and violations. The results of each metric are considered in this section.

1. Margin

Margin was defined as the amount of available "white space" in the schedule, left over once the participant had completed the trial. "White space" is calculated in time (i.e., sum of the time between scheduled activities). Margin was calculated as a ratio between initial white space in the schedule, and final white space remaining. There were four instances from two participants where participants had "double banded" activities (had overlapping activities) at the end of the trials, resulting in negative margin. This data was therefore removed from the further analysis of margin as the calculated margin would not be comparable.

As seen in Table 1 and Fig. 5, the ratio of margin appears to decrease with the increase of number of activities, which is counterintuitive. Additionally, the "easiest" trial act12-con0 (12 activities with 0% constraint condition) also has the highest ratio margin. These results indicate the ratio of margin is not a good candidate metric for plan effectiveness.

Table 1 Average and standard deviation of ratio of leftover margin per condition

Condition	Mean	SD
act12-con0	0.243711	0.066038
act12-con33	0.218553	0.059104
act12-con66	0.199686	0.034015
act24-con0	0.132075	0.054806
act24-con33	0.139413	0.042197
act24-con66	0.174004	0.044528
act36-con0	0.109164	0.028618
act36-con33	0.112028	0.074015
act36-con66	0.103774	0.032091

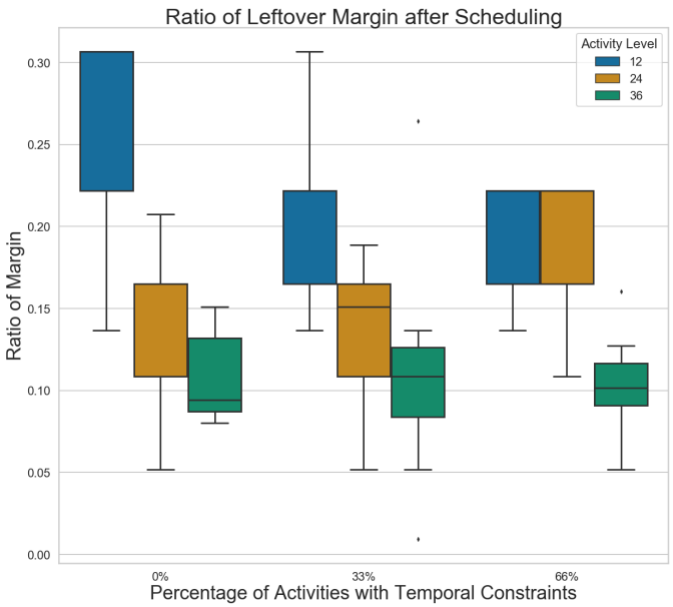

Fig 5. Boxplot of the ratio of margin at the end of scheduling.

2. Violations

Violations are defined as activities that were scheduled but did not meet an activity's constraints[2]. No violations were created during the 0% constraint condition for 12, 24 or 36 tasks, as there were no constraints available to violate.

3. Violations left at end of trial

No participants left any violations at the end of each trial.

4. Total number of violations throughout trial

Violations were made in all conditions with 33% and 66% constrained activities. A general pattern can be discerned that more violations were made with increasing numbers of tasks to be scheduled, both in the 33% (12 activities, M= 0.89, SD= 0.93; 24 activities M=2, SD=2.45; 36 activities M=2.11, SD=1.62) and 66% conditions (12 activities, M= 0.4.33, SD= 2.4; 24 activities M=4, SD=4.5; 36 activities M=9.33, SD= 6.9) (Fig. 6). Most violations were created in conditions of 66% constraints, with increasing violations for 12, 24, and 36 activities, suggesting an interaction effect of constraint and activity numbers on violations made.

A repeated measures ANOVA determined that the percentage of constraints had a significant effect on the number of violations created throughout the trial [$F(1,8) = 12.577$, p=.008]. Mauchly's test indicated that the assumption of sphericity had been violated for the number of activities $\chi^2(2)$, p=.013. Therefore, the Greenhouse-Geisser corrected tests are reported (ε=.585) and the results show that the number of activities had a significant effect on the number of violations created $F(1.169, 9.352)=12.859$, p=.004, while interaction was not significant $F(2, 16)=1.737$, p=.208. The results indicate that the larger the percentage of constraints, the more violations were created. Likewise, the more activities to be scheduled, the more violations were created.

[2] Double banding of activities is also considered a violation but did not present themselves as violations in the plan and were not considered in the analysis as violations.

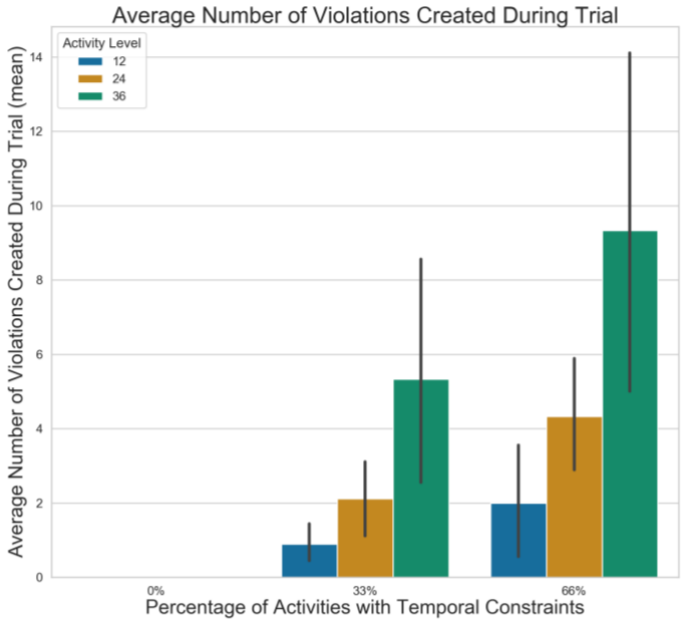

Fig. 6 The number of violations created during each trial. No violations were created in the 0% constraint level as no constraints were available to violate in those conditions.

D. High, Medium, and Low Priority Activities left unscheduled

All High Priority activities were scheduled in all 81 trials (nine participants completing nine trials each). Since all participants left either one or no Medium Priority activities unscheduled, descriptive statistics were conducted on Medium Priority activities and Low Priority activities were analyzed. Participants were able to schedule all Medium activities for the 12 activity with 0%, & 12 activity with 33% constraint conditions as well as the 24 activity with 66% constraint condition (Table 2).

A Friedman's test was conducted on the number of Low Priority activities left unscheduled. Significant differences in the number of Low Priority activities left unscheduled were found between conditions [$\chi 2(8) = 52.35$, $p<0.001$]. A series of Wilcoxon tests were conducted as post-hoc analyses. Four Wilcoxon pair analyses were conducted per condition, and so a Bonferroni correction was applied creating a significance level of 0.01 per condition. Considering comparisons between the number of activities and 0% activities with constraints, significantly more Low Priority activities were left unscheduled in the 24 activity condition compared to the 12 activity condition ($p=0.01$). The difference approached significance between the 12 activity and 36 activity conditions ($p=0.02$) but was not significant between 24 and 36 activity conditions. In the 33% condition, significantly more low priority activities were left unscheduled in the 24 activity condition ($p<0.01$) and 36 activity condition ($p<0.01$) compared to the 12 activity condition, but differences were not significant between 24 and 36 activity conditions. In the 66% of constrained activities condition, there were significantly more low priority activities left unscheduled in the 24 activity condition compared to the 12 activity condition ($p<0.01$), and 36 activities conditions compared to 12 activities ($p<0.01$), but was not significant between 24 and 36 activities.

Comparing constraint conditions, there was no significant difference in low priority unscheduled activities for 12 activities with 0% constraints and 12 activities with 33% constraints. Interestingly, more low priority tasks were left unscheduled in the condition with 12 activities and 0% constrained activities than in the 12 and 66% constrained activities conditions, which approached significance ($p=0.03$). For conditions with 24 activities, there was no significant difference in the number of low priority activities left unscheduled in the 24 activities condition with 0% constraints and 24 activities with 33% constraints. The number of low priority activities left unscheduled in

the 24 activities with 66% constraint condition was more than in the 24 activities with 0% and 33% constraint conditions however, which approached significance (p=0.01). Finally, considering the 36 activity conditions, there were no significant differences between constraint conditions in the number of remaining unscheduled Low Priority activities.

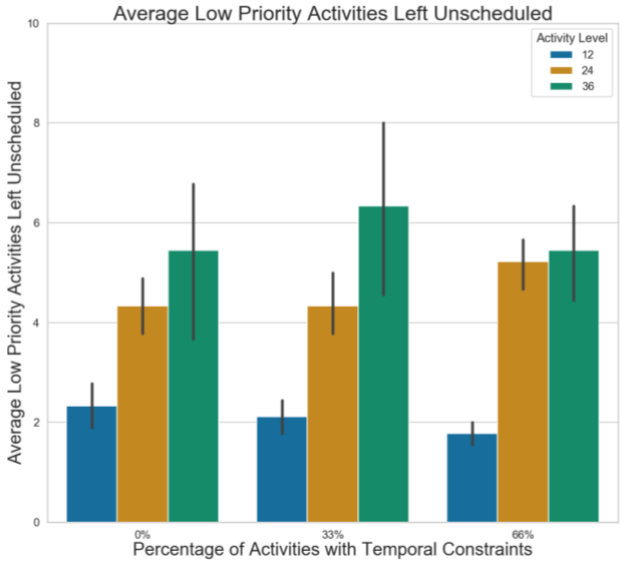

Fig. 7 The average number of low priority activities left unscheduled.

Table 2 Descriptive Statistics for Number of Activities Left Unscheduled by Condition

	Medium Priority Activities			Low Priority Activities		
	Total Count	Mean	SD	Total Count	Mean	SD
act12-con0	0	0	0	21	2.33	.71
act12-con33	0	0	0	19	2.11	.60
act12-con66	1	.11	.33	16	1.78	.44
act24-con0	1	.11	.33	39	4.33	.86
act24-con33	3	.33	.5	39	4.33	1.00
act24-con66	0	0	0	47	5.22	.83
act36-con0	1	.11	.33	49	5.44	2.60
act36-con33	2	.22	.44	57	6.33	2.70
act36-con66	2	.22	.44	49	5.44	1.58

E. Efficiency

Efficiency is a measure of how quickly participants were able to schedule given the scheduling task difficulty. Three metrics were evaluated: time on task, time on task per scheduled activity, and time to resolve violations.

1. Time on Task

Time on task was collected through the experimental software starting when the plan was presented to the participant and until they stated that they had completed planning for each trial. Descriptive statistics were conducted for time on task and showed that the overall average duration for time on task was 392.06 seconds, or 6 minutes and 32 seconds (SD=216.37). As expected, the longest mean duration was for the condition with the most activities and constraints (36 activities, 66% constraints, M=741.56, SD=164.03) as shown in Fig. 8, while the shortest mean duration was for the condition with the least activities and no constraints (12 activities, 0% constraints: M=142.89, SD=15.64) as shown in Fig. 9. Fig. 10 summarizes the time on task per trial as a function of presentation order.

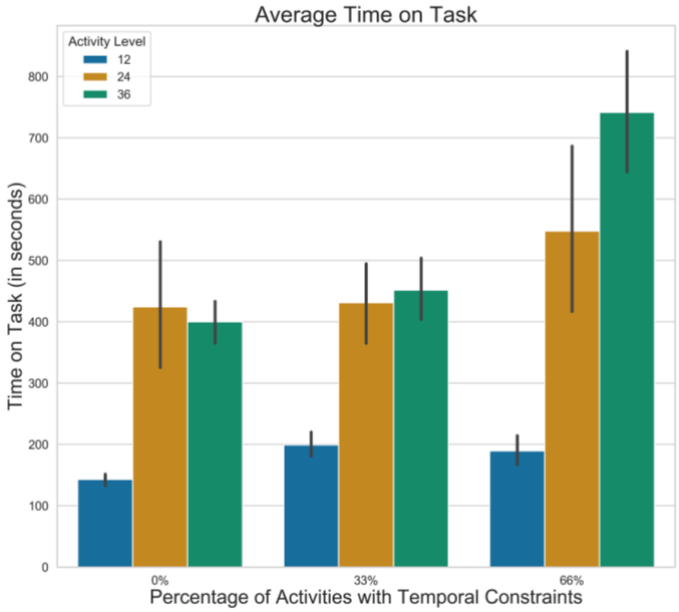

Fig. 8 Average time on task organized by condition

Further, a repeated measures ANOVA was conducted and determined that the number of activities and percentage of constraints had significant effects on increasing time on task [$F(2,16) = 51.978$, $p<.001$], [$F(2,16) = 20.258$, $p<.001$] (Fig. 9). Mauchly's test indicated that the assumption of sphericity had been violated $\chi^2(9)$, $p=.025$ for interaction effects, therefore Greenhouse-Geisser corrected tests are reported ($\varepsilon=.502$). The results show a significant interaction between the number of activities and percentage of constraints [$F(2.008, 16.062)=5.259$, $p=.017$]. Fig. 9 suggests that the condition for 36 activities and 66% constraints caused the interaction as this trial took the longest in comparison to the other trials.

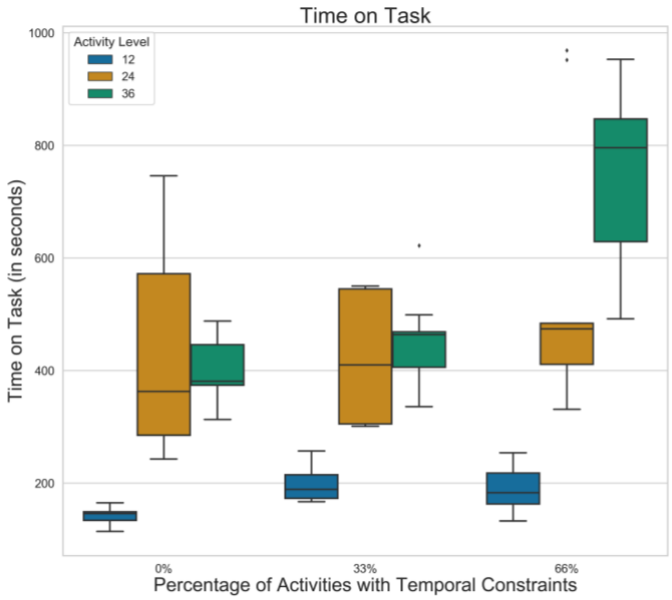

Fig. 9 Time on task organized by condition.

Standard pairwise comparisons (with Bonferroni corrections) were conducted to determine significance between conditions. The condition with 12 activities was significantly shorter than in the 24 activities (p<.001) and 36 activities conditions (p<.001). However, time on task in the 24 activities condition was not significantly shorter than in the 36 activities condition (p=.173) (Table 4). The condition with 66% constraint was significantly longer than in 0% constraints (p=.001) and 33% constraints (p=.001). However, time on task in the 33% condition was not significantly longer than in the 0% constraint condition (p=.170) (Table 4).

This effect may be explained by the trial order (Fig. 10). Participants were tasked to solve one of the highest activity level conditions first and finished with one of the lowest activity level conditions in the experiment which could indicate a performance drop off. Especially coupled with a self-terminating task, it was observed that participants revised their plans much less at the end of the experiment than at the beginning. Further, the lack of significance between the 24 activity condition and the 36 activity condition could indicate a performance cap and thus would not observe any more significant differences if more activities or constraints were added.

Table 4. Pairwise Comparisons

Measure: Activities

(i) Activities	(j) Activities	Mean difference (i-j)	Sig.[b]
1 (12 activities)	2 (24 activities)	-290.593*	.000
	3 (36 activities)	-353.815*	.000
2 (24 activities)	3 (36 activities)	-63.222*	.173

Based on estimated marginal means

* The mean difference is significant at the .05 level.

[b.] Adjustment for multiple comparisons: Bonferroni.

Measure: Constraints

(i) Constraint	(j) Constraint	Mean difference (i-j)	Sig.[b]
1 (0%)	2 (33%)	-38.444	.170
	3 (66%)	-170.630*	.001
2 (33%)	3 (66%)	-132.185*	.001

Based on estimated marginal means

* The mean difference is significant at the .05 level.

[b.] Adjustment for multiple comparisons: Bonferroni.

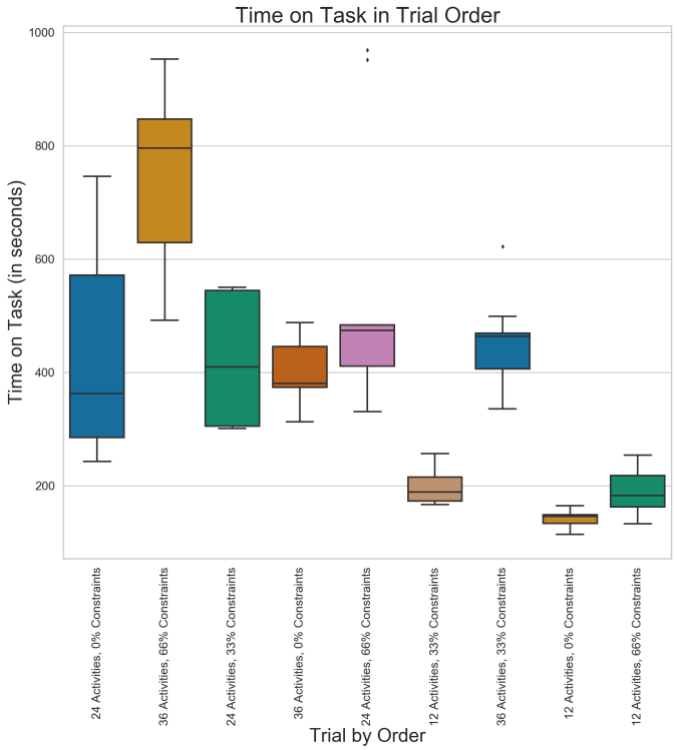

Fig. 10 The average time on task in order of trial.

2. Time on Task per Scheduled Activity

Time on task per scheduled activity was calculated by dividing each trial's time on task by the number of activities that was scheduled (Table 5). When normalizing time on task by the number of actually scheduled activities, we do not see the same trends and variability as with time on task as shown in Fig. 11. This suggests that the time on task is not driven solely by the number of activities.

Table 5. Descriptive Statistics for Time on Task per Scheduled Activity

Condition	Mean	SD
act12-con0	14.84	1.86
act12-con33	20.11	2.48
act12-con66	18.76	3.89
act24-con0	21.89	9.39
act24-con33	22.28	5.57
act24-con66	29.44	13.77
act36-con0	13.20	2.09
act36-con33	15.31	1.95
act36-con66	24.53	5.36

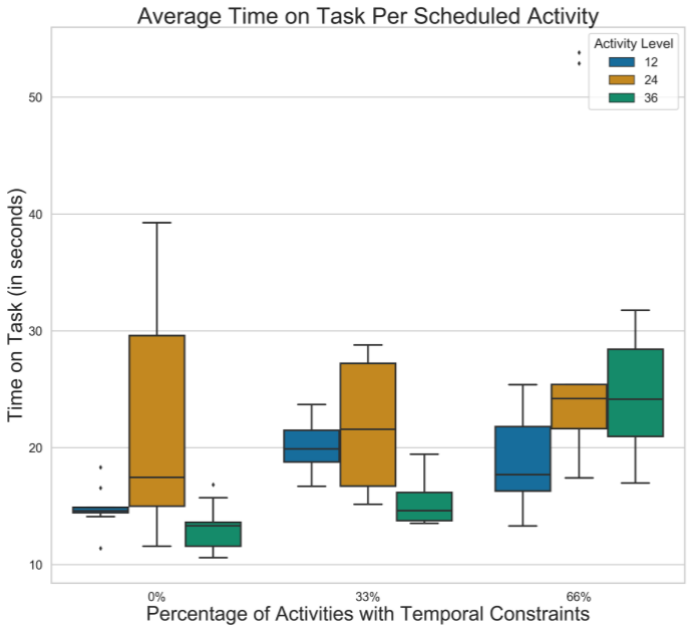

Fig. 11 Average time on task per activity scheduled for each condition.

3. Average time to resolve violations

The average time to resolve violations was calculated measuring the time between creating a violation and solving it. Participants chose to immediately fix the violation created so that calculation was possible. Since there were no violations created in the 0% constraint condition, this average time only pertains to the conditions with constraints present. Figure 12 shows the distribution of this average time across the conditions, which does not appear to greatly vary. The average across all conditions was 8.14 seconds (SD=7.96) (Table 8). We assessed average time to resolve violations because if participants had a lot of violations, we would expect time on task to also increase. However, we did not see a trend in this metric to warrant further evaluation.

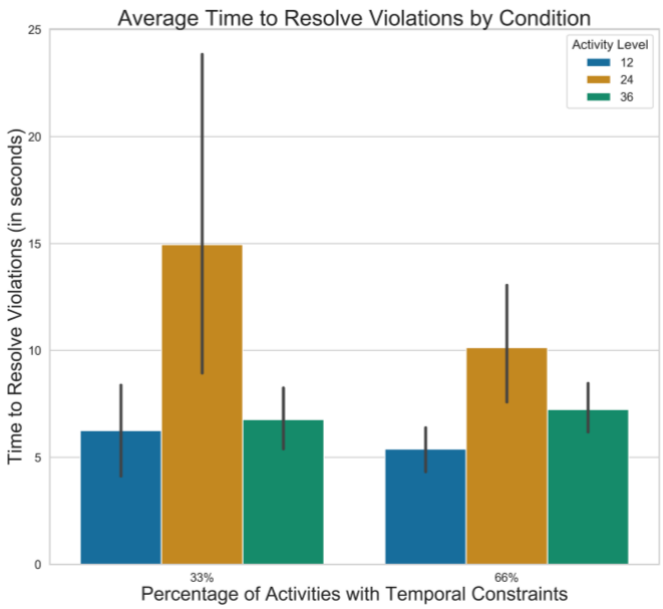

Fig. 12 Average time to resolve violations by condition

Table 8 Average Time (in seconds) to Resolve Violations

Condition	Mean	SD
All conditions	8.14	7.96
act12-con33	6.25	3.24
act12-con66	5.38	2.32
act24-con33	14.94	17.28
act24-con66	10.12	8.86
act36-con33	6.77	5.18
act36-con66	7.23	5.47

F. Correlations

Due to a large number of dependent variables, a subset of measures was selected for correlation analysis. The selection was based on previous findings from inferential statistics and understanding of the variables that were likely to be associated. For brevity, only statistically significant results will be reported below. Correlations were conducted across all conditions for the following dependent variables:

- Time on task
- Total number of violations created during each trial
- Number of total unscheduled activities at the end of each trial
- Average workload

Considering first the conditions with 0% constrained activities, significant, positive relationships were identified between the independent variable of the number of activities to be scheduled and trial duration (*rs* = 0.74,

p<0.001, R2 = 0.55), unscheduled activities (rs = 0.68, p<0.001, R2 = 0.46) and solution quality score (rs = 0.66, p<0.001, R2 = 0.44). In addition, time on task was also significantly positively associated with unscheduled activities (rs = 0.58, p<0.005, R2 = 0.34). Workload was not correlated with anything else in the 0 constraints condition.

A similar pattern of relationships is seen in the 33% constrained activity condition. The independent variable of number of activities to be scheduled was again significantly positively correlated with time to complete trials (rs = 0.72, p<0.001, R2 = 0.52), and number of activities left unscheduled (rs = 0.79, p<0.001, R2 = 0.62). An interesting finding was that the number of activities was not significantly associated with the number of violations as was expected. However, the relationship between these variables approached significance (rs = 0.36, p=0.65). Significant, positive relationships were found however between the number of violations and time to complete trial (rs = 0.48, p<0.05, R2 = 0.23), and remaining unscheduled activities (rs = 0.5, p<0.01, R2 = 0.25).

In the 66% constrained activity condition, a similar pattern of relationships emerged once again, suggesting a robust positive association between the number of activities to be scheduled and time on task (rs = 0.81, p<0.001, R2 = 0.66), and unscheduled activities (rs = 0.78, p<0.001, R2 = 0.61). An interesting finding was that in the 66% constrained condition, the relationship between number of activities to be scheduled and number of violations was highly significant, with more tasks associated with greater violations (rs = 0.58, p<0.005, R2 = 0.34). The finding that this relationship was only significant in the conditions with 66% of activities constrained (as opposed to conditions with 0% and 33% of activities constrained) suggests there may be a mediating or moderating covariate of the % of constrained activities.

Conditions with 12 activities to be scheduled, with either 0%, 33% or 66% of constrained activities, were analyzed. In conditions with 12 activities to be scheduled, a significant, positive, correlation was identified between percentage of constraints and time on task (rs = 0.54, p<0.005, R2 = 0.29). This finding is interesting as combined with the correlation analysis across tasks, this relationship suggests that both task number and constraint percentage is significantly, positively correlated with time on task. A second interesting finding is that constraint percentage was significantly positively correlated with the number of violations (rs = 0.58, p<0.005, R2 = 0.34), even with only 12 tasks to schedule. Taken in conjunction with the previous correlation analysis across tasks, it appears that the percentage of constraints has a stronger association with number of violations. This finding is expected as more activities with constraints would result in a greater chance of creating violations. No other findings were significant.

For conditions with 24 activities to be scheduled, only one relationship between variables was found to be significant. There was a significant positive correlation between the percentage of constrained activities and number of violations (rs = 0.81, p<0.001, R2 = 0.66). Finally, considering conditions with 36 activities, percentage of tasks with constraints was significantly, positively, associated with both time on task (rs = 0.79, p<0.001, R2 = 0.66), and number of violations (rs = 0.81, p<0.001, R2 = 0.66). The final relationship that was significant was a positive correlation between time on task and violations (rs = 0.68, p<0.001, R2 = 0.46).

IV. Discussion

This pilot study was conducted to measure human performance on the task of self-scheduling with naive users. Scheduling problems were manipulated by changing the number of activities to be scheduled and if the activities had a constraint. The within-subjects design resulted in nine total conditions consisting of three levels of the number of activities to be scheduled along with three levels of task constraints expressed in percentages. Tests of differences and relational analyses showed that the number of activities to schedule and the percentage of task constraints in the plan affected human performance, with some varying results. Collectively, the results indicate a consistent trend: the number of activities in a plan has a significant effect on scheduling performance. Increasing the number of activities decreases plan effectiveness and efficiency while increasing workload (i.e., more unscheduled low priority activities, a greater number of violations, longer time on tasks, and higher workload measures). The effect of the percentage of constrained activities was identified for efficiency and workload, with a lesser effect on plan effectiveness (based on the total number of violations created). Additionally, the interactivity of these variables were observed in some measures resulting in a compounding effect.

A. Human Performance for Self-Scheduling

1. Plan Effectiveness

Plan effectiveness was considered as a human performance measure. Specifically, four metrics that related to effectiveness were analyzed: margin, number of activities left unscheduled, total number of violations created, and total number of violations left at the end of the self-scheduling task. Margin was originally proposed as a metric for plan effectiveness; however, the results of the pilot study were counterintuitive as the ratio of margin appeared to decrease with the increase of number of activities. These results would indicate that participants were better able to create schedules with more activities. While this may be true, it does not speak to the participants performance but rather on the scheduling task itself.

The metric of number of activities left unscheduled did not provide sufficient variability across the conditions, except for Low Priority activities. Fortunately, this indicates that the participants were following experimental instructions (as they were asked to schedule higher priority activities first). The trials with only 12 activities had a significantly lower number of unscheduled low priority activities relative to trials with 24 and 36 activities, regardless of number of constraints. This result is not surprising as there were only four low priority activities. Thus, similar to margin, the results relate to the scheduling task as opposed to participant's performance. Surprisingly, despite having a different number of low priority activities, no significant difference was found between 24 and 36 activities with regards to the number of unscheduled low priority activities. As such, alternative metrics of plan effectiveness were considered.

Number of violations were also originally proposed as a metric for plan effectiveness. None of the participants left any violations at the end of the scheduling task, however, if there were violations left, the schedule would be considered not valid. The total number of violations created throughout the scheduling task was also considered a metric for plan effectiveness -- the less violations created would mean the participant had an easier time creating a good plan. Unfortunately, this is not the best metric for this experiment design as one of the condition levels required scheduling of activities with no constraints. Under these conditions, a violation would never be created. However, the results are consistent with the general trends observed, where there is a significant effect due to the number of activities and percentage of constraints. More violations were created when there were more activities and more constraints to resolve. As expected, participants had more opportunities to create violations with more activities and more constraints. Thus, a participant's ability to plan effectively, as measured by the number of violations created, was significantly affected by our two independent variables.

2. Scheduling Efficiency

Scheduling efficiency was considered as a human performance measure. Three metrics were analyzed that related to efficiency: time on task, time on task per scheduled activity, and time to resolve violations. Increasing the number of activities increased time on task for all conditions. Similarly, as the percentage of constraints increased, time on task increased as well. These results do not appear to be consistent with normalized time on task (based on the number of activities scheduled). When considering both independent variables together, there appears to be a greater, compounding effect on increasing time on task. Unfortunately, the compounding effect may have been caused by an order effect as that particular condition (i.e., highest number of activities and constraints) occurred early during the experiment. These results suggest that controlling for percentage of constraints or the number of activities, we should expect a linear relationship with time on task on the other independent variable.

Average time to resolve violations did not show a lot of variability, likely because participants were asked to schedule only one type of constraint. It was observed that the strategy that participants utilized was to immediately resolve violations after they were created, resulting in small variability.

3. Workload

Workload was measured using the NASA Task Load Index. The results suggest that the number of activities and the percentage of constraints had an effect on increasing workload. Participants experienced a higher workload when there were more activities to schedule as well as more constrained activities to schedule. In

conjunction with a similar result for time on task, it suggests that both the number of activities and the percentage of activities with constraints are strong indicators for scheduling task difficulty. Unlike the time on task metric, there were no reported workload interaction effects with the number of activities and percentage of constraints. These results indicate the workload is a valuable measure to include in future experiments and that the unweighted average workload measure is a sufficient measure.

4. Compound Effects

One of the most interesting findings is that our results suggest that there may be a compound effect of a high number of activities and higher percentage of constrained activities no time on task. Another result to note, in terms of performance and workload, was that significant differences could not be detected between trials that had 24 activities and those that had 36 activities. Two possible reasons may be that there were no significant performance differences or that the trial order masked those differences (many of the more difficult or higher activity conditions were presented in the first half of the configuration).

B. Limitations

As an initial study investigating human performance in self-scheduling, several limitations are noted. First, each of the nine participants completed all nine experimental trials in the same order. Although the trial order was randomized in a particular order prior to the study, a limitation in the experimental software hindered the ability to execute unique randomized trials for each participant. Thus, potential order effects may have occurred. Future experiments shall allow each participant to experience a randomized set of trials. Secondly, the small sample size of nine participants limits the generalizability of results and may have resulted in the possibility of a Type II error for the results that did not reach significance.

Finally, motivation may have been a contributor to performance. It was observed that participants spent less time in the latter trials when refining their plans. One participant verbally stated that they simply had no motivation to create an ideal plan and signaled that they were done planning after placing as many activities as they could from the Task List to the Timeline, potentially influencing the data.

C. Future Research

Future research will focus on investigating the impact of different types of constraints on performance in addition to refining the best metrics to measure human performance in self-scheduling. It was found that metrics for determining an objective quantification of plan ‘goodness’, or how efficient or effective a plan was were lacking, and the domain would benefit from future research in this area. In this study, it was discovered that workload was a good measure of performance, but margin and time to resolve violations were not. Future research could take care to initially normalize margin in the experiment design in order to conduct a more complete analysis. Furthermore, it was observed that participants immediately resolved violations after creating them, resulting in small variability in time to resolve violations data. Future research could consider scheduling strategy as a function of scheduling performance, especially with participants of varying levels of scheduling experience (naive vs experts).

Next steps in this research include introducing different types of constraints that Operational planners deal with, in addition to adding more than one single constraint on an activity. It is expected that this addition would create complexity, and this research will align closer to the current context of self-scheduling, as long duration missions will involve multiple constraints and many different types of constraints. Lastly, research in this domain shall be cognizant of investigating scheduling vs. rescheduling. This study provided participants with an initial template with some bounds to that scope of the solution space. Future research should investigate whether it would be more efficient and supportive of human performance to provide a completed template which is then modified, rather than self-scheduling within limited bounds.

V. Conclusions

The future of NASA space exploration depends on providing the tools to allow crew member autonomy. This pilot study successfully demonstrated that there is a large effect due to the number of activities and potentially, percentage of constrained activities on human scheduling performance. Scheduling performance decreased with more activities to plan and at times, the percentage of constrained activities compounded the decrease. These results provide evidence towards developing a model for scheduling task difficulty, which includes the number of activities as well as the number of constrained activities. Further, these findings provide a basis for future studies in the effort to provide scheduling autonomy to crew members. More research in this domain is required, and in addition to investigating the effects of different types of constraints, future research shall identify how the task of self-scheduling can be best provided to crew members, such as the development of automated aids. Flight crew scheduling needs to be revolutionized in order to make future long-duration space missions possible, which can only be achieved with more research in this field.

Appendix

Appendix A.

Appendix A. shows a list of activities for the participant to schedule, shown on the left hand computer screen throughout the trial.

Activity Name	Priority Level	Constraint
XF305 Camcorder Setup	High	
Unpack & Stow Astrobee	High	
MELFI Genes Sample Retrieval	High	
EHS Waste Water Bag Changeout	High	
CRISPR Run Part 4	High	
Extravehicular Unit Maintenance	High	
Robot Micro Conical Tool Stowage	High	
Vascular Echo BP Monitoring	High	
Soyuz 741 Equipment Stowage	Medium	
3-D Printer Maintenance	Medium	
3-D Printing Training	Medium	
FGB Pressurized Vacuum Cleaning	Medium	
Crew Prep for Earth to Return	Medium	
Inspection of Recording Equipment	Medium	
Cyclone Hardware Closeout	Medium	
Robotic Arm Task	Medium	
Robonaut-Orbit	Low	
Blood Sample Collection Set up	Low	
ISS Crew Conference	Low	
Genes in Space MWA Preparation	Low	
Body Sampling Survey	Low	
Pre-sleep Questionnaire	Low	
Medical Survey	Low	
Astrobee Evaluation	Low	

Acknowledgements

This research is sponsored by the Human Research Program, within the Human Factors and Behavioral Performance Element.

References

[1] Bresina, J.L. (2015). Activity Planning for Lunar Orbital Mission. Proceedings of the Twenty-Seventh Annual Conference on Innovative Applications of Artificial Intelligence.

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

[3] Marquez, J.J., G. Pyrzak, S. Hashemi, S. Ahmed, K. McMillin, J. Medwid, D. Chen, and E. Hurtle (2013) "Supporting Real-Time Operations and Execution through Timeline and Scheduling Aids", International Conference of Environmental Systems ICES, Vail, CO. July 2013.

[4] Marquez, J.J., Hillenius, S., Deliz, I., Kanefsky, B., Zheng, J., and Reagan, M. (2017). Increasing Crew Autonomy for Long Duration Exploration Missions: Self-Scheduling. Aerospace Conference, 2017 IEEE, March 2017.