Identifying Gender Differences in Information Processing Style, Self-efficacy, and Tinkering for Robot Tele-operation

Dilruba Showkat[1] and Cindy Grimm[2]

Abstract— **As robots become more ubiquitous it is important to understand how different groups of people respond to possible ways of interacting with the robot. In this study, we focused on gender differences while users were tele-operating a humanoid robot that was physically co-located with them. We investigated three factors during the human-robot interaction (1) information processing strategy (2) self-efficacy and (3) tinkering or exploratory behavior. Experimental result show that the information on how to use the robot was processed comprehensively by the female participants whereas males processed them selectively *($p < 0.001$)*. Males were more confident when using the robot than females *($p = 0.0002$)*. Males tinkered more with the robot than females *($p = 0.0021$)*. Tinkering might have resulted in greater task success and lower task completion time for males. Similar to existing work on software interface usability, our results show the importance of accounting for gender differences when developing interfaces for interacting with robots.**

I. Introduction

As robot capabilities grow, we will increasingly see them applied to a wider range of applications such as surgery, health care, engineering education, helping elderly senior citizens for self-care, socialization, child care as robotic assistant, and as companions. Unlike traditional software interfaces, robotic interfaces have an in-person, physical context to them — the robot may be in the room with the user, be physically interacting with the environment, or be directly interacting with the user. An open question is: will we see the same gender and age-related differences in robotic interfaces that we see in traditional software ones, or does the physical nature of the interaction mitigate these differences? In this paper we show that — for a simple tele-operation manipulation task where the user and the robot are in the same room — we *do* see the same sorts of interaction differences for gender.

Previously, gender differences have been studied in social psychology, communication, education, creativity, human-computer interaction (HCI), web psychology and much more [1], [4], [6], [20], [29]. Gender differences have also received increased attention in human robot interaction (HRI). Previous gender related studies in HRI have focused on questions such as: how a robot should approach a person sitting on a chair (front, left or right) [10], the influence of a robot's voice (voice's fundamental frequency) [30], the robot's facial features (head dimensions, chin length) [30], the robot's gender (male and female) [36] and how all these play out for male and female interactees. Existing research also confirms that there exists significant gender differences in the robot's perceived social presence, social facilitation, disclosure, and persuasiveness [34], [36]. Additional researchers studied how a robot can persuade different individuals to disclose their private information [31], how different individuals perform in given task (arithmetic) at the presence of a robot [34], and how receptive people were to the robot suggestions [30], [36]. These research findings have focused primarily on the robot's role as a social agent, and how a robot is perceived in different social situations from a gender perspective.

In this paper we take a step back from the social aspect of robotics and focus on interacting with the robot as a physically embodied computer, and examine tele-operation as a user-interface task (see Figure 1). Research has shown that for standard software interfaces the gender lens matters [6], [14], and, moreover, that it is possible to evaluate and redesign an interface to make it more inclusive. Our goal is to determine if the same gender-based differences exist when the user-interaction component has an embodied element (the robot). If so, then the same approaches used when analyzing software interfaces can be employed in order to improve robotic tele-operation interfaces.

Pulling from existing research, we examined three factors that influence how a user will respond to an interface: i) information processing strategy; ii) self-efficacy; and iii) tinkering. These factors have been shown to correlate both with gender and with user's problem solving ability [6], [14]. For the task, we asked users to tele-operate a humanoid robot in order to grasp and manipulate several household objects. For all of the users, this was the first time they had tele-operated this robot. We evaluate the user's task performance and the three factors above through a mix of observations and surveys. We now briefly outline the three factors in more detail; for all factors, these are *aggregate* behaviors.

Information processing style: According to the selectivity hypothesis males and females use different information processing strategies [7]. In general, males tend to attend to a task using discrete segments of configuration, while females are more inclined to pay attention to the configuration as a whole. Similarly, males tend to engage in selective- or heuristic-based processing, making use of single cues that are highly available and most noticeable in the current context in order to make a single inference. In contrast, females tend to gather all of the available cues first, then use the collective as the basis of judgment for information processing [12], [13],

*This work supported in part by NSF grants CNS 1659746 and CNS 1730126

[1]Dilruba Showkat graduated with a Masters in EECS from Oregon State University showkatd@oregonstate.edu

[2]Cindy Grimm, Collaborative Robotics Institute, Oregon State University, OR 97330, USA Cindy.Grimm@oregonstate.edu

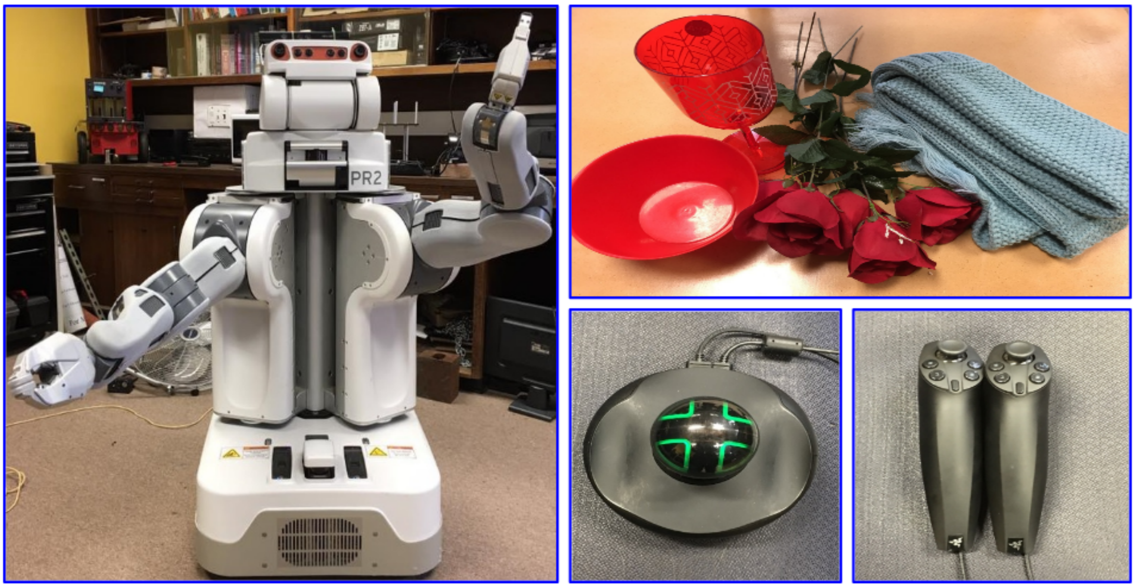

Fig. 1. Personal robot (PR2) used in the study (left). Objects (top-right), reference point (right-bottom-left) and joysticks (right-bottom-right) used in the study.

[18], [26].

Self-efficacy: Social cognitive theory [2] suggests that self-efficacy (an individual's personal judgment about their own capabilities) plays a crucial role in the choices that they make, the amount of effort they put in, and task retention when faced with adversity. Researchers report that males tended to have more self-efficacy, less math anxiety, and a higher performance score compared to females for mathematical problem solving [25], [28].

Tinkering: Cognitive playfulness — or tinkering — has been demonstrated to positively impact test performance in a field study of full time employees [23]. Students also benefit in their scientific understanding as they tinker or play with tools [16], [21]. The education literature shows that males are more likely to tinker than females [38].

In this study, we are taking our first step towards understanding gender differences in robot tele-operation (manipulating a robot from a distance) by investigating these three factors. We already know robots are *perceived* differently by males and females when the robots are presented as social agents. This work studies gender differences in information processing style, self-efficacy and tinkering in the context of *controlling* a robot. Robots may all, eventually, be fully autonomous, but in the meanwhile direct (or assisted) control of the robot plays a very important role in deploying robots in the real world. Improving the accessibility of these interfaces may help to increase the diversity of the predicted burgeoning robotic work-force.

Specifically, our research questions are:

- **RQ1:** Are there any differences in the information processing style across gender when learning to tele-operate a robot?
- **RQ2:** Are there any gender differences in self-efficacy that impact the efficient use of the robot?
- **RQ3:** Are there any gender differences related to tinkering that impact efficient use of the robot?

II. Related Work

In this section we review related work on gender and robots as well as gender and interface development.

A. *Gender and its influence in Human Robot Interaction*

Research has shown that a robot's appearance (gender, voice, facial features) has a significant influence on how recommendations provided by the robot were received, and these influences varied by gender of the person interacting with the robot [31], [27]. A robotic tutor similarly influences task performance (easy and hard math) [34]. The study reported that males tend to think of robots as more human like whereas females think of robots as more machine like. As a result, males reported feeling socially facilitated by the robot while performing arithmetic tasks.

Furthermore, research shows that by changing the robot's persona we can gather different level of information, for example a study reported that males expressed more information to the female robot and females expressed more information to the male robot. In another research, males and females participants reported that they find opposite-sex robot to be more trustworthy and engaging. Researchers also found gender differences in the negative attitude toward robots [27], [31], [34], [36]. All these findings are interesting from the perspective of the latest voice enabled technologies such as Amazon's Alexa, Google's Assistant etc. These bots are female voice enabled (Siri for iPhone, Microsoft's Cortana) as the basis of conveying information. Furthermore, previous research confirms that using only vocal cues within a machine is enough to bring sex based stereotype responses even though the environment and circumstances in which the robot operates could be different [8], [35]. When robots are given some human-like attributes, people can easily relate to them. Even though we consider robots/bots as machines we still tend to use he/she when addressing them. Therefore, we should consider gender (both of the robot and the user) when designing interfaces in order to balance these effects.

III. Research Methodology

Our study was a between-group (male and female) study. Our experiment design included both qualitative and quantitative data capture and analysis. The study was conducted in three phases: A pre-task phase (survey), the actual task session where the user tele-operated the robot (video-taped), and a post-task session (survey) where users were asked to reflect on the task.

A. Experimental Setup

The study took place in a research lab setting. During the experiment the lab was quiet, and only the researcher and the participant were present in the experimental area where the robot was. The participant was placed about four feet in front of the robot (facing it) so that the robot could not accidentally hit them. The object for the task was placed by the experimenter on a chair directly in front of the robot. Participants were asked to tele-operate the robot with two joysticks, one for each robot arm. They operated the robot while facing it, requiring mirrored hand movements. I.e., the robot's left hand was controlled by the participant's left hand, but from the participant's point of view the left hand was on the right side in their field of view (see Figure 1). While mirroring probably caused additional cognitive load, this did enable the participants to see both the robot and the manipulation space at the same time.

During the task participants were asked to think-aloud [15]. The average time to complete the entire study was less than an hour.

1) PR2 Robot: We used a Personal Robot (PR2) from Willow Garage in our experiment (shown in Figure 1). The PR2 is a humanoid robot (mobile and with two arms) that was developed in part for human-robot interaction studies.

2) Participants: Participants were recruited via fliers placed on and around campus. We did not restrict our study participation by age group or gender, however, only students responded to the flier. Communication and scheduling for the study was done through email. Participants were given $15 to complete the study. Overall 12 participants (6 males and 6 females) took part in the study. An additional 5 participants were used in a pilot study; their data is not included in this paper.

3) Objects and Task Description: Participants were asked to use a joy-stick to tele-operate the PR2 robot in order to manipulate the four objects shown in Figure 1 (bowl, rose, glass, and scarf).

Object	*Task*	*Directive*
Bowl	Pick it up and put it down	Glass/fragile
Rose	Pick it up by touching stem	Real flower
Glass	Have the robot drink from it	Glass/fragile
Scarf	Fold it	Favorite

TABLE I
OBJECT AND TASK DESCRIPTIONS.

A pilot study was used to narrow an initial six objects and tasks down to four. The criteria was that the tasks had to be difficult, but doable. The objects themselves were chosen to have a mix of material properties (soft versus hard, fragile versus robust, compliant versus rigid). Although we could not use actual glass/water for the task, we compensated for this by explicitly adding a directive to the task (3rd column in Table I) such as the bowl is fragile, don't drop or crush it, and don't spill the water.

B. Study Procedure

1) Pre-task session
    a) Participant enters the study room and the researcher introduces them to the robot.
    b) Consent process.
    c) The participant fills out the pre-task questionnaire.
2) Task session
    a) Researcher provides tutorial handout for controlling the robot.
    b) Researcher gives the participant the joystick controllers. The participant was free to "play" with the controllers and ask questions about their use until comfortable.
    c) Researcher places the first object on the chair in front of the robot.
    d) Researcher explains the task and directives.
    e) The participant attempts the task.
    f) Repeat c-d until all four tasks are attempted.
3) Post-task session: Participant fills out post-task questionnaire.

IV. Data Analysis and Coding

In this section we provide details on the data collected in each phase and how it was analyzed or coded.

A. Pre-task Session Data

There were three pre-task questionnaires, one for each of the factors. In addition to this, we collected basic demographic data (age, gender, experience with video games, field of study). For gender, the field was left blank; for our group, all participants self-identified as male or female.

1) Self-efficacy questionnaire (SE): We used the standard self-efficacy test questionnaire proposed by Compeau and Higgins [9] with slight modifications to the wording to make it applicable to the robot tele-operation task. We used the same 10 point Likert-scale (1 indicates "Not at all confident",

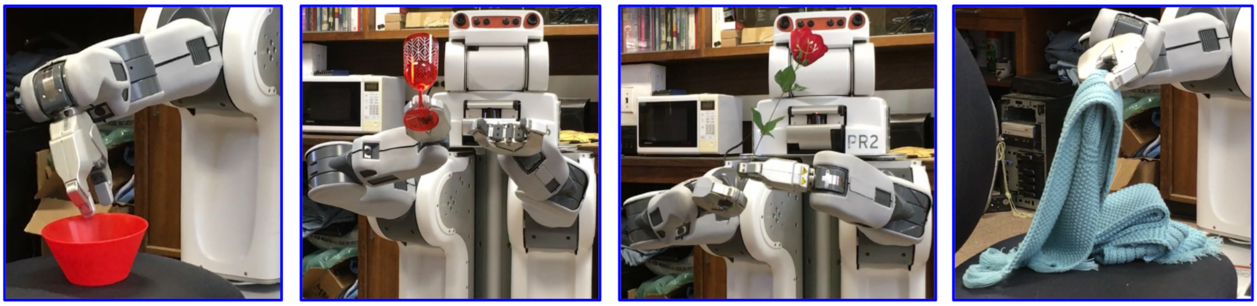

Fig. 2. Examples of task success, from left to right. Putting the bowl down, drinking from a glass, passing a rose, folding a scarf.

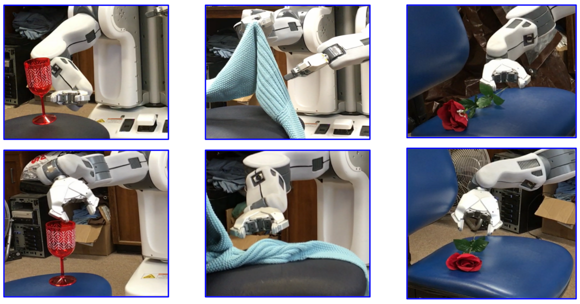

Fig. 3. Tinkering while grasping the wine glass (object-stem-left, object-rim-left), scarf (gripper-both, gripper-horizontal-left), rose (gripper-horizontal-left, gripper-vertical-left).

Code	*Query type*
Strategy	Preferred process to do something
Feature	Particular feature functionality, eg. buttons
Functionality	Robot arm/wrist/body function

TABLE II

CODING STRATEGY FOR INFORMATION NEED OPEN-ENDED QUESTION (INO).

5 indicates "Moderately confident", and 10 indicates "Totally confident"). There were a total of 10 questions

Responses were summed to give a single self-efficacy score between 0 and 100 for each participant.

2) Information-need open-ended question (INO): We asked the following open-ended question: "*Imagine that you are asked to move a delicate expensive flower vase on the table using the PR2 robot. What information do you need to complete the task using robot?*"

We categorized information-need by: i) strategy, ii) feature and iii) functionality, based on [19]. We coded responses using the criteria given in Table II, and counted both the number in each category and the number of categories that were seen.

3) Tinkering open-ended question (TinO): We asked participants the following open-ended question: "*How often do you play with a new hardware device? Why? Why not?*"

Responses were qualitatively coded as playful/not playful using two rules: 1) playful: when a response contains keywords such as "fun","play","enjoy", "tinker", "love trying" etc. 2) not-playful: when a response demonstrates a negative approach about playfulness towards a new device (eg. "not much", "not often").

B. Task Session Data

We collected three measurements from the task session: A task score (how well they completed the task), time to complete the task, and a tinkering score (coded based on the video data).

1) Task score (TS): We qualitatively coded the task score from the video data. The Task Score was 100 for successfully completing a task, 80 for completing the task but with a fault such as breaking the flower stem or tipping the bowl over, 20 for attempting a task but not completing it, and 0 if they were unable to pick up the object (see Figure 2).

2) Task Completion Time (TT): Total task completion time (in minutes) was computed from the video data, starting when the participant first moved the joy stick. Completion time was determined by when the participants said they were done.

3) Tinkering count from video (TinVID): We used the video data to count the number of times a participant showed evidence of tinkering or playful experimentation. More specifically, we counted the number of times a participant tried a different approach, defined as follows: i) Switching from left to right hand or from one hand to both (and vice-versa) ii) Approaching the object from a different direction, top, bottom, or side iii) Changing the orientation of the gripper by rotating it at the wrist, and iv) picking up the object from a different point (eg top of stem versus bottom). Several playful experimentations are shown in Figure 3.

C. Post-task Session Data

1) Information processing questionnaire (IPQ): We used a standard information processing questionnaire from the literature [37]. It consists of ten 7-point likert scale questions. The first 5 questions test for comprehensive information processing style and the remaining test for systematic information processing. We modified the questionnaire slightly so the questions fit the robot interaction task.

The comprehensive score was calculated by summing up the score for the first 5 questions, the selective score from the sum of the last 5 questions (max 35 in each case).

V. Results

A. Demographics

Of our twelve participants, half self-identified as female, half as males. Participant ages averaged 20.5 years old, with a minimum of 18 and a maximum of 24. The male participants were slightly younger than the female (19.5 years versus 21.5).

Statistical analysis of our background data on video gaming experience showed significant gender differences, with males averaging 9.5 years of experience, and females reporting 0-10 years (average 2.1) (Mann-Whitney $W = 3$, $p = 0.0185$).

We qualitatively coded field of study as follows: Liberal Arts = 1, Engineering = 2. Statistical analysis revealed that there was no significant gender differences related to field of study (Mann-Whitney $W = 9$, $p = 0.0705$).

B. Statistical tests and validity

In all of our ANOVA analysis, we used gender as the factor. We rank transformed our data before applying the ANOVA. We also used Fisher's exact test for small-category data. For research questions 2 and 3 we also looked at the effect of video game experience on the results, using an ANCOVA with video game experience as the covariate and gender as the factor.

For research questions 1 and 3 we use two data sources; this helps mitigate the fact that those results use qualitative data that was coded by a single experimenter [17], [22], [24], [33].

C. RQ1: Information processing style

We use two sources of information to answer the information processing research question; the open-ended pre-test question (INO) and the post-test information processing style questionnaire (IPQ). Similar to previous studies, we find that females tend to ask more questions, more *types* of questions, and employ a comprehensive information processing style. All individual responses are shown in Figure 4, with aggregate numbers in Table III.

1) INO question response analysis: We split the analysis into two pieces: Types of questions asked (by category) and total number of questions asked. Refer to Table II for a definition of the three categories. Table III summarizes the number of participants who asked each *type* of question, split by gender. Example quotes:

- Strategy: 83% of all participants asked at least one strategy question.

  "How do I turn it on? How do I control it?" - P02
  "How do I move the arm? How do I open and close the gripper? How do I rotate the gripper?" - P06

- Feature: 50% of all participants asked for feature-related information.

  "What keys do I use? Which button does each task?" - P11
  "What does it use to grab stuff?" - P05

Gen	*Strat*	*Feat*	*Func*	*Comp*	*Sel*
M	4/6	1/6	2/6	61/210	125/210
F	6/6	5/6	6/6	111/210	128/210
Total	10/12	6/12	8/12	172/420	253/420

TABLE III

Left (INO): How many people asked at least one question in each category (Strategy, Feature, Functionality). Right (IPQ): Comprehensive and Selective scores, summed for all participants. Results reported by gender.

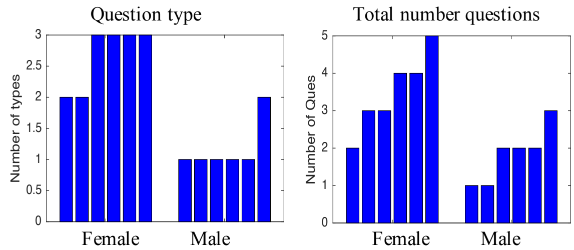

Fig. 4. Left: Number of *types* of questions asked, by gender (out of 3). Right: Total number of questions asked, by gender (INO).

- Functionality: 67% of all participants asked at least one function-related question.

  "How fast does it rotate?" - P01
  "How sensitive are the controls?" - P11

A one-way ANOVA of the INO data in Table III ($F(1, 10) = 32.472$, $p = 0.0002$) revealed significant gender difference in the number of categories of question asked.

We also compared the total number of questions asked (Figure 4). Again, the ANOVA revealed a significant gender difference in the total number of questions asked ($F(1, 10) = 10.796$, $p = 0.008$).

2) IPQ response analysis: We summed up the total number of questions asked by gender (Table III). These numbers are also statistically significant (Fisher exact test $p < 0.001$).

3) Summary: Our findings are in-line with existing work that shows that females tend to use a more comprehensive processing style, with more information-gathering up-front, while males tend to use a selective processing style with fewer questions [7].

D. RQ2: Self-efficacy and Task Completion

Figure 5 shows all self-efficacy scores from our SE questionnaire, with female participants having a lower SE score than male ones (Table IV). A one-way ANOVA ($F(1, 10) = 31.304$, $p = 0.0002$) revealed significant gender differences, congruent with previous studies in various computational problem solving settings such as math and debugging [3], [28].

We have two quantitative measures of the participant's ability to complete the task, Task Time (TT) and Task Score (TS), summarized in Table IV. Both showed statistically significant differences, with males performing the task both better and faster (TS - ANOVA $F(1, 10) = 11.538$, $p = 0.0068072$), (TT - $F(1, 10) = 5.16$, $p - value = 0.049$).

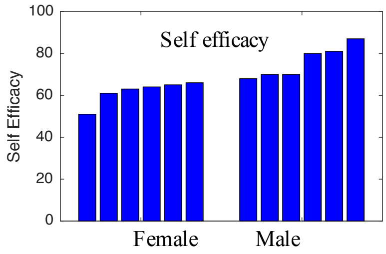

Fig. 5. Self-efficacy scores (SE) by gender.

Gender	*SE score*	*Task score*	*Time*	*Tinker*
F	61.67	228.3	21.56	4.8
M	76.0	350	12.45	7.7

TABLE IV

SELF-EFFICACY SCORE (SE) (OUT OF 100), TOTAL TASK SCORE (TS) (OUT OF 400), TASK TIME (TT) IN MINUTES, AND NUMBER OF OBSERVED TINKERING EVENTS. NUMBERS REPRESENT AVERAGES BY GENDER.

1) Video-game experience and Self-efficacy: An obvious question to ask is if previous video game experience is the underlying factor in these results, rather than gender. For this analysis we turned to an ANCOVA with gender as the categorical variable, and video game experience as the predictor.

Our results show that video game experience has significant effect ($p = 0.0226$) on self-efficacy. This finding is similar to previous research of performance in debugging and introductory computer courses, which showed increases in experience may increase self-efficacy for females [3], [32].

Video gaming experience also showed significant effect ($F(1, 10) = 6.61, p = 0.030$) on task time with gender as the factor. This can be explained by the fact that the users who played video games are probably more comfortable using joysticks, and therefore completed the task faster. This fact was also mentioned by several users while performing the task.

Interestingly, there was no correlation between self-efficacy and task score ($p = 0.13$), and no correlation between video game experience and task score ($p = 0.09$). A summary of the correlations is given in Table V.

E. RQ3: Tinkering

We analyze tinkering through the open-ended tinkering question (TinO) and coding of the video (TinVid). Both

Factor	*Task score*	*Total time*	*VG Exp.*
Tinkering	0.006*	0.003*	0.006*
Self-efficacy	0.27	0.27	0.04*

TABLE V

p-VALUE OF SPEARMAN'S CORRELATION COEFFICIENT TEST OF SELF-EFFICACY AND TINKERING W.R.T. TASK SCORE, TASK TIME AND VIDEO GAMING EXPERIENCE. HERE * INDICATES SIGNIFICANT CORRELATION.

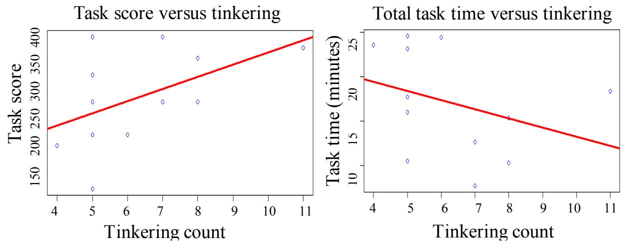

Fig. 6. Total task score and task completion time versus tinkering.

show significant gender differences, with males tinkering more than females. Again, these findings are congruent with existing work on problem-solving.

1) TinO response analysis: Example quotes of playful/tinkering versus not when using a new device:

- playful:

  "I buy new hardware to program (Ardunio) about once every month or two. A new phone 3-4 years. Love buying VR, tablets, smart watches." - P08

- not-playful:

  "I don't play with new hardware device, I am a people's person." - P07

Fisher exact test on our coded data revealed significant ($p = 0.0021$) gender differences when asked about playing with new devices.

2) TinVid data analysis: Average tinkering events (as counted in the video) are summarized in Table IV. Differences are significant ($F(1, 10) = 13.158$, $p = 0.004$), with males trying over half again as many approaches as females.

3) Video game experience and Tinkering: As before, an ANCOVA with tinkering as the dependent variable, gender as the independent variable, and video gaming experience as the covariate reveals a significant (p <2e-16 ***) effect.

Unlike self-efficacy, there is a positive correlation (0.7319251) between tinkering and task score (non-parametric Spearman's rho test, $S = 76.669$ and $p = 0.006807$). Figure 6 shows a linear model fit to the data.

Willingness to abandon a strategy and try another probably accounts for this correlation. In the video, participants who tried multiple approaches in quick succession were more likely to find a strategy that was successful than those who only tried one or two.

There was also a significant negative correlation (-0.7724873) between tinkering and task time (non-parametric Spearman's rho test $S = 506.93$, $p = 0.0032$, with 95% confidence interval). Figure 6 shows a linear model fit to the data.

Interestingly, task time goes *down* as tinkering increases, indicating that tinkerers tended to switch more quickly to another approach than non-tinkerers.

As in self-efficacy, there was a strong positive correlation (0.7345532) between video game experience and tinkering ($S = 75.918$ and $p = 0.0065$).

VI. Discussion

Tele-operating a robot fundamentally involves two skill sets; controlling the gripper via the joy sticks and understanding *how* to grab and manipulate the object with the gripper (what grips will work, when it will slip out, where to grab the object to obtain the best grasp, etc.). We hypothesize that video game experience helps with the former, leading to faster completion times. Manipulation with a gripper, however, was a relatively novel task for all participants, and here the key to success was simply trying several different approaches in rapid succession.

Our findings for gender-bias in information processing, self-efficacy, and tinkering were in line with existing studies, indicating that problem-solving in the physical domain with a physical tele-op interface shares many of the same underlying factors. This implies that existing approaches for reducing bias in software interfaces (supporting a comprehensive information processing style, encouraging tinkering) are probably applicable to the tele-op domain as well and requires further research in varied settings. We request the research community to further investigate in this direction.

VII. Threats and Limitations

Although we attempted to recruit from a broader demographic, our participants ended up being students, with the attendant biases associated with that group. Our study population was also small.

We have identified two possible confounds that may explain differences in task performance; the first is video game experience, the second is a known gender difference in spatial or mental rotation. It is possible that task score was partially influenced by the need to mentally mirror the actions of the robot.

In the following subsections, we will discuss internal, external, and conclusion threats to our study.

a) External Validity: External validity refers to the fact that whether findings from our experiment is generalizable or not [5], [11], [33], [39]. To avoid demographic bias we made our study open to all. But eventually only students reached out and showed interest to participate in our study. Thus, our result may not be generalizable across different demographics (age, education etc.) due to small sample size.

b) Internal Validity: It refers to the risks that may hinder the causal relationship between dependent and independent variables [5], [11], [33], [39]. In our study, males and females had significantly different video gaming experience. We used video gaming experience as covariates in all of our ANCOVA analysis to understand its possible influence. We cross validated our results using methodological triangulation.

c) Statistical Conclusion Validity: Conclusion validity refers to the threat associated with the incorrect assumptions made before we statistically test a relationship between two variables [5], [11], [33], [39]. We rank transformed our data before applying ANOVA. We validated each relationship with non-parametric tests and methodological triangulation. Moreover, we used covariates in all our ANCOVA. All these certainly validates our research findings.

VIII. Conclusion

In this study, we compared information processing style, self-efficacy and tinkering, and how they influence task success and task completion time when tele-operating a robot. Our results show that existing identified gender differences in these factors also exist in the tele-operation realm, with similar effects on task. In future, we would like to extend our study by incorporating more complex tasks with wide range of population. Furthermore, robots gender and how it influence males and females task behavior would be interesting to research as well.

IX. Acknowledgments

We would like to thank all of our participants for their time and support.

References

[1] John Baer and James C Kaufman. 2008. Gender differences in creativity. *The Journal of Creative Behavior* 42, 2 (2008), 75–105.

[2] Albert Bandura. 1989. Human agency in social cognitive theory. *American psychologist* 44, 9 (1989), 1175.

[3] Laura Beckwith, Margaret Burnett, Susan Wiedenbeck, Curtis Cook, Shraddha Sorte, and Michelle Hastings. 2005. Effectiveness of end-user debugging software features: Are there gender issues?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 869–878.

[4] Susan Bergin and Ronan Reilly. 2005. The influence of motivation and comfort-level on learning to program. In *Proceedings of the PPIG*, Vol. 17. Citeseer, 293–304.

[5] Marilynn B Brewer, H Reis, and C Judd. 2000. Research design and issues of validity. *Handbook of research methods in social and personality psychology* (2000), 3–16.

[6] Margaret Burnett, Anicia Peters, Charles Hill, and Noha Elarief. 2016. Finding gender-inclusiveness software issues with GenderMag: a field investigation. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2586–2598.

[7] Patricia Cafferata and Alice M Tybout. 1989. Gender Differences in Information Processing: A Selectivity Interpretation, Cognitive and Affective Responses to Advertising. (1989).

[8] C CDATA-Nass, Youngme Moon, and Nancy Green. 1997. Are computers gender-neutral? Gender stereotypic responses to computers. *Journal of Applied Social Psychology* 27, 10 (1997), 864–876.

[9] Deborah R Compeau and Christopher A Higgins. 1995. Computer self-efficacy: Development of a measure and initial test. *MIS quarterly* (1995), 189–211.

[10] Kerstin Dautenhahn, Michael Walters, Sarah Woods, Kheng Lee Koay, Chrystopher L Nehaniv, A Sisbot, Rachid Alami, and Thierry Siméon. 2006. How may I serve you?: a robot companion approaching a seated person in a helping context. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. ACM, 172–179.

[11] Ellen A Drost. 2011. Validity and reliability in social science research. *Education Research and Perspectives* 38, 1 (2011), 105.

[12] Valentina Grigoreanu, Margaret Burnett, Susan Wiedenbeck, Jill Cao, Kyle Rector, and Irwin Kwan. 2012. End-user debugging strategies: A sensemaking perspective. *ACM Transactions on Computer-Human Interaction (TOCHI)* 19, 1 (2012), 5.

[13] Valentina Grigoreanu, Jill Cao, Todd Kulesza, Christopher Bogart, Kyle Rector, Margaret Burnett, and Susan Wiedenbeck. 2008. Can feature design reduce the gender gap in end-user software development environments?. In *Visual Languages and Human-Centric Computing, 2008. VL/HCC 2008. IEEE Symposium On*. IEEE, 149–156.

[14] Charles Hill, Shannon Ernst, Alannah Oleson, Amber Horvath, and Margaret Burnett. 2016. GenderMag experiences in the field: The whole, the parts, and the workload. In *Visual Languages and Human-Centric Computing (VL/HCC), 2016 IEEE Symposium on*. IEEE, 199–207.

[15] Riitta Jääskeläinen. 2010. Think-aloud protocol. *Handbook of translation studies* 1 (2010), 371–373.

[16] Beverley Jane. 2006. Online Reflections about Tinkering in Early Childhood: A Socio-Cultural Analysis.. In *Asia-Pacific Forum on Science Learning and Teaching*, Vol. 7. ERIC, n1.

[17] Todd D Jick. 1979. Mixing qualitative and quantitative methods: Triangulation in action. *Administrative science quarterly* 24, 4 (1979), 602–611.

[18] Frank R Kardes, Maria L Cronley, James J Kellaris, and Steven S Posavac. 2004. The role of selective information processing in price-quality inference. *Journal of Consumer Research* 31, 2 (2004), 368–374.

[19] Cory Kissinger, Margaret Burnett, Simone Stumpf, Neeraja Subrahmaniyan, Laura Beckwith, Sherry Yang, and Mary Beth Rosson. 2006. Supporting end-user debugging: what do users want to know?. In *Proceedings of the working conference on Advanced visual interfaces*. ACM, 135–142.

[20] I Han Kuo, Joel Marcus Rabindran, Elizabeth Broadbent, Yong In Lee, Ngaire Kerse, RMQ Stafford, and Bruce A MacDonald. 2009. Age and gender factors in user acceptance of healthcare robots. In *Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on*. IEEE, 214–219.

[21] Maarten H Lamers, Fons J Verbeek, and Peter WH van der Putten. 2013. Tinkering in scientific education. In *Advances in computer entertainment*. Springer, 568–571.

[22] Nancy L Leech and Anthony J Onwuegbuzie. 2007. An array of qualitative data analysis tools: A call for data analysis triangulation. *School psychology quarterly* 22, 4 (2007), 557.

[23] Joseph J Martocchio and Jane Webster. 1992. Effects of feedback and cognitive playfulness on performance in microcomputer software training. *Personnel Psychology* 45, 3 (1992), 553–578.

[24] Sandra Mathison. 1988. Why triangulate? *Educational researcher* 17, 2 (1988), 13–17.

[25] Judith L Meece and M Gail Jones. 1996. Gender differences in motivation and strategy use in science: Are girls rote learners? *Journal of Research in Science Teaching* 33, 4 (1996), 393–406.

[26] Joan Meyers-Levy and Barbara Loken. 2015. Revisiting gender differences: What we know and what lies ahead. *Journal of Consumer Psychology* 25, 1 (2015), 129–149.

[27] Tatsuya Nomura, Takayuki Kanda, and Tomohiro Suzuki. 2006. Experimental investigation into influence of negative attitudes toward robots on human–robot interaction. *Ai & Society* 20, 2 (2006), 138–150.

[28] Frank Pajares and M David Miller. 1994. Role of self-efficacy and self-concept beliefs in mathematical problem solving: A path analysis. *Journal of educational psychology* 86, 2 (1994), 193.

[29] Tuomas Pekkarinen. 2012. Gender differences in education. *Nordic Economic Policy Review* 1 (2012), 165–195.

[30] Aaron Powers and Sara Kiesler. 2006. The advisor robot: tracing people's mental model from a robot's physical attributes. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. ACM, 218–225.

[31] Aaron Powers, Adam DI Kramer, Shirlene Lim, Jean Kuo, Sau-lai Lee, and Sara Kiesler. 2005. Eliciting information from people with a gendered humanoid robot. In *Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on*. IEEE, 158–163.

[32] Karsten Rex and Roberta M Roth. 1998. The relationship of computer experience and computer self-efficacy to performance in introductory computer literacy courses. *Journal of research on computing in education* 31, 1 (1998), 14–24.

[33] Per Runeson, Martin Host, Austen Rainer, and Bjorn Regnell. 2012. *Case study research in software engineering: Guidelines and examples*. John Wiley & Sons.

[34] Paul Schermerhorn, Matthias Scheutz, and Charles R Crowell. 2008. Robot social presence and gender: Do females view robots differently than males?. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*. ACM, 263–270.

[35] Jean Scholtz. 2003. Theory and evaluation of human robot interactions. In *System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on*. IEEE, 10–pp.

[36] Mikey Siegel, Cynthia Breazeal, and Michael I Norton. 2009. Persuasive robotics: The influence of robot gender on human behavior. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*. IEEE, 2563–2568.

[37] Chris MR Smerecnik, Ilse Mesters, Math JJM Candel, Hein De Vries, and Nanne K De Vries. 2012. Risk perception and information processing: The development and validation of a questionnaire to assess self-reported information processing. *Risk Analysis* 32, 1 (2012), 54–66.

[38] MARJA VAN DEN HEUVEL-PANHUIZEN. 2004. Girls and Boys problems. *International perspectives on learning and teaching mathematics* (2004), 237–52.

[39] C Wohlin, P Runeson, M Host, MC Ohlsson, B Regnell, and A Wesslen. 2000. Experimentation in software engineering: an introduction. 2000. (2000).