PLOS ONE Supporting Information Appendix S1

Article title: Comparing Completion Time, Accuracy, and Satisfaction in Virtual Reality vs. Desktop Implementation of the Common Coordinate Framework Registration User Interface (CCF RUI)

Authors: Andreas Bueckle[1,2], Kilian Buehling[3], Patrick C. Shih[4], Katy Börner[1,2]

[1] Department of Information and Library Science, Luddy School of Informatics, Computing, and Engineering, Indiana University
[2] Department of Intelligent Systems Engineering, Luddy School of Informatics, Computing, and Engineering, Indiana University
[3] Research Group Knowledge and Technology Transfer, Fakultät Wirtschaftswissenschaften, Technische Universität Dresden
[4] Department of Informatics, Luddy School of Informatics, Computing, and Engineering, Indiana University

The following Supporting Information is available for this article:

- Supplementary text
- Figure S1
- Tables S1 to S2

Other supplementary materials for this manuscript include:

- Qualtrics survey PDF: https://github.com/cns-iu/rui-tissue-registration/blob/main/RUI_user_study_questionnaires.pdf
- Video demonstrations of three setups (2D Desktop, VR Tabletop, VR Standup):
  - 2D Desktop
  - VR Tabletop
  - VR Standup

Power Analysis

We performed a power analysis to estimate a sample size that would give us a high chance of statistically significant results. The sample size calculation was performed using G*power (1), an open-source statistical power analysis tool developed at the University of Duesseldorf in Germany. Among other functionality, G*power allows the user to set a range of metrics to retrieve a sample size for a given research design as expressed through the metrics outlined below. Before being able to use G*power, we needed to parametrize our study.
An essential factor when parametrizing this research design for sample size calculation is that it contains two separate user studies that differ by the number of groups, number of measurements, and thus the required sample size, see Table S1.

The statistical power calculation also needs other parameters from the researcher:

- **Test family**
  - We choose “F tests” as the type of test that allows to perform ANOVA for a comparison of repeated measures between groups of subjects
- **Statistical test**
  - The concrete test we choose to perform is ANOVA repeated measures, between factors as we aim to analyze between subjects/users (as opposed to just within subjects, i.e., just within users)
- **Type of power analysis**
  - G*power allows us to computer a variety of power analysis metrics. We choose “A priori: Compute required sample size” in order to determine the number of subjects needed for this study
- **Effect size**
  - The effect size is defined as the quotient of the mean difference over the standard deviation. In G*power, the effect size can range from 0 to 50. It is always positive or zero. The default effect size in G*power is set to 0.25. We choose 0.3 for our effect size calculation.
  - Literature (2, 3) suggests that it is common to distinguish between small (0.2), medium (0.5) and high (0.8) effect sizes in power analysis, so choosing 0.3 seems reasonably conservative (as sample size is negatively correlated with effect size, ceteris paribus). Effect size is commonly seen as either small, medium, or large. Ideally, the effect size can be estimated by looking at pilot study data. The pilot study data at hand, however, may not be sufficient in order to produce an accurate result for this estimate.
- **Alpha error probability**
  - Probability of Type I error
  - Our assumed alpha error probability of 0.05, which is a standard value.
- **Power**
  - 1 - Probability of Type II error (beta)
  - Standard value for beta is usually set to four times the alpha error probability
  - We set the power metric to 0.8.

Given these values for the metrics required by G*power, we obtain the results shown in Table S1 and Figure S1. Note that we conducted a separate user study described in a forthcoming paper where we compared several cohorts of 2D Desktop, VR Tabletop, and VR Standup users. This is why we chose 6 as the “number of groups” for the Plateau phase. We took the resulting 84 subjects and divided them by 2. The users in this study (84/2 = n = 42) then served as control cohort for the forthcoming study.

Table S1. Metrics for sample size calculation by phase. Notice that all the values are identical except the number of groups and measurements.

Metrics	**Ramp Up**	**Plateau**
Effect size	0.3	0.3
Alpha error probability	0.05	0.05

Power	0.8	0.8
Number of groups	3	6
Number of measurements	14	30
Sample size	63	84

Table S1 shows that the two studies contained in our RUI VR research design require two different sample sizes to achieve the targeted effect of size 0.3. Since we are conducting both studies together, we need to aim for the higher number, which is **84 subjects, or 14 subjects per group per setup** (Control_VR-Standup, Control_VR-Tabletop, Control_2D_Desktop, Experiment_VR-Standup, Experiment_VR-Tabletop, Experiment_2D_Desktop).

Figure S1 contains screenshots of the G*power interface with the values shown in Table S1 entered into the appropriate fields.

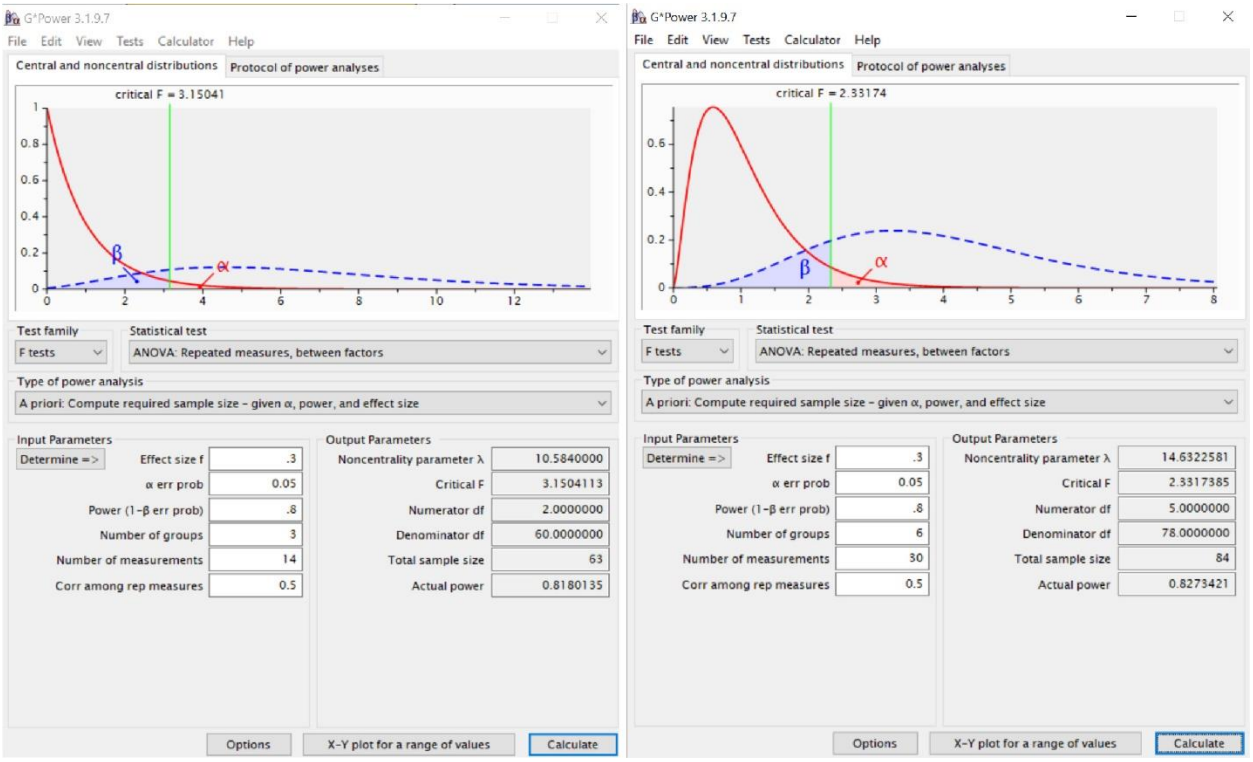

*Figure S1. G*power results for the Complexity phase (left) and the Plateau phase (right) using the metrics in Table S1.*

Data Collected from Unity

Table S2. Telemetry data logged during experiment.

Name of metric	**Definition**	**Setup**
elapsedTime	number of seconds since the beginning of the experiment	All
headsetX, headsetY, headsetZ	Position of HMD, from the origin of the virtual scene (VR only), in meters	VR Tabletop, VR Standup
controllerLeftX, controllerLeftY, controllerLeftZ, controllerRightX, controllerRightY, controllerRightZ,	Positions of controllers, from the origin in the virtual scene (VR only), in meters	VR Tabletop, VR Standup
mousepositionX, mousepositionY	the user's 2D mouse movements, in pixels	2D Desktop
distance	distance between the centroids of the two blocks, in meters (VR)/ Unity scene units (2D Desktop)	All
currentTissueBlock.transform.position.x, currentTissue block.transform.position.y, currentTissue block.transform.position.z	Current position of the tissue block	All
objectXLength, objectYLength, objectZLength	Current dimensions of tissue block (and target)	All
currentObjectRotationX, currentObjectRotationY, currentObjectRotationZ	Current x, y, z rotation of tissue block	All
targetRotationX, targetRotationY, targetRotationZ	Current x, y, z rotation of target block	All
angle	Current rotational difference between tissue block and target block, in degrees (0-180)	All
button	Current button pressed ("grab" and "menu" in VR, "left" or "right" mouse button on 2D Desktop), or function called ("reset position", "reset rotation", etc.)	All
side	"Left" or "right" hand (VR), "left" or "right" mouse button (2D Desktop)	VR
status	"Down" (pressed), "up" (released), "dragging" (2D Desktop only)	All

References

1. Faul F, Erdfelder E, Lang A-G, Buchner A. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods. 2007;39(2):175-91.
2. Brysbaert M. How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. J Cogn. 2019;2(1):1-38.
3. Cohen J. Statistical power analysis for the behavioral sciences. New York City, NY, USA: Academic Press; 2013.