Journal of Visualized Experiments

Evaluating Usability Aspects of A Mixed Reality Solution for Immersive Analytics in Industry 4.0 Scenarios

--Manuscript Draft--

Article Type:	Invited Methods Article - JoVE Produced Video
Manuscript Number:	JoVE61349R2
Full Title:	Evaluating Usability Aspects of A Mixed Reality Solution for Immersive Analytics in Industry 4.0 Scenarios
Corresponding Author:	Ruediger Pryss Julius-Maximilians-Universitat Wurzburg Würzburg, GERMANY
Corresponding Author's Institution:	Julius-Maximilians-Universitat Wurzburg
Corresponding Author E-Mail:	ruediger.pryss@uni-wuerzburg.de
Order of Authors:	Burkhard Hoppenstedt
	Thomas Probst
	Manfred Reichert
	Winfried Schlee
	Klaus Kammerer
	Myra Spiliopoulou
	Johannes Schobel
	Michael Winter
	Anna Felnhofer
	Oswald D. Kothgassner
	Ruediger Pryss
Additional Information:	
Question	**Response**
Please indicate whether this article will be Standard Access or Open Access.	Standard Access (US$2,400)
Please indicate the **city, state/province, and country** where this article will be **filmed**. Please do not use abbreviations.	Ulm, BW, Germany
Please confirm that you have read and agree to the terms and conditions of the author license agreement that applies below:	I agree to the Author License Agreement
Please specify the section of the submitted manuscript.	Engineering
Please provide any comments to the journal here.	

Powered by Editorial Manager® and ProduXion Manager® from Aries Systems Corporation

Manuscript Click here to access/download;Manuscript;61349_R2.docx

TITLE:

Evaluating Usability Aspects of a Mixed Reality Solution for Immersive Analytics in Industry 4.0 Scenarios

AUTHORS AND AFFILIATIONS:

Burkhard Hoppenstedt[1], Thomas Probst[2], Manfred Reichert[1], Winfried Schlee[3], Klaus Kammerer[1], Myra Spiliopoulou[4], Johannes Schobel[1], Michael Winter[1], Anna Felnhofer[5], Oswald D. Kothgassner[6], Rüdiger Pryss[7]

[1]Institute of Databases and Information Systems, Ulm University, Ulm, Germany
[2]Department for Psychotherapy and Biopsychosocial Health, Danube University Krems, Krems an der Donau, Austria
[3]Department of Psychiatry and Psychotherapy, University of Regensburg, Regensburg, Germany
[4]Faculty of Computer Science, Otto von Guericke University Magdeburg, Magdeburg, Germany
[5]Department of Pediatrics and Adolescent Medicine, Medical University of Vienna, Vienna, Austria
[6]Department of Child and Adolescent Psychiatry, Medical University of Vienna, Vienna, Austria
[7]Institute of Clinical Epidemiology and Biometry, University of Würzburg, Würzburg, Germany

KEYWORDS:

Immersive Analytics, Mixed Reality, Spatial Sounds, Visual Analytics, Smartglasses, Usability, Stress Level, Learnability.

SUMMARY:

This protocol delineates the technical setting of a developed mixed reality application that is used for immersive analytics. Based on this, measures are presented, which were used in a study to gain insights into usability aspects of the developed technical solution.

ABSTRACT:

In medicine or industry, the analysis of high-dimensional data sets is increasingly required. However, available technical solutions are often complex to use. Therefore, new approaches like immersive analytics are welcome. Immersive analytics promise to experience high-dimensional data sets in a convenient manner for various user groups and data sets. Technically, virtual-reality devices are used to enable immersive analytics. In Industry 4.0, for example, scenarios like the identification of outliers or anomalies in high-dimensional data sets are pursued goals of immersive analytics. In this context, two important questions should be addressed for any developed technical solution on immersive analytics: First, is the technical solutions being helpful or not? Second, is the bodily experience of the technical solution positive or negative? The first question aims at the general feasibility of a technical solution, while the second one aims at the wearing comfort. Extant studies and protocols, which systematically address these questions are still rare. In this work, a study protocol is presented, which mainly investigates the usability for immersive analytics in Industry 4.0 scenarios. Specifically, the protocol is based on four pillars. First, it categorizes users based on previous experiences. Second, tasks are presented, which can be used to evaluate the feasibility of the technical solution. Third, measures are presented, which

quantify the learning effect of a user. Fourth, a questionnaire will be presented that measures and evaluates the stress level when performing tasks. Based on these pillars, a technical setting was implemented that uses mixed reality smartglasses to apply the study protocol. The results of the conducted study show the applicability of the protocol on the one hand and the feasibility of immersive analytics in Industry 4.0 scenarios on the other. The presented protocol includes a discussion of limitations.

INTRODUCTION:

Virtual-reality solutions (VR solutions) are increasingly important in different fields. Often, with VR solutions (including Virtual Reality, Mixed Reality, and Augmented Reality), the accomplishment of many daily tasks and procedures shall be eased. For example, in the automotive domain, the configuration procedure of a car can be supported by the use of Virtual Reality[1] (VR). Researcher and practitioners have investigated and developed a multitude of approaches and solutions in this context. However, studies that investigate usability aspects are still rare. In general, the aspects should be considered in the light of two major questions. First, it must be evaluated whether a VR solution actually outperforms an approach that does not make use of VR techniques. Second, as VR solutions are mainly relying on heavy and complex hardware devices, parameters like the wearing comfort and mental effort should be investigated more in-depth. In addition, the mentioned aspects should always be investigated with respect to the application field in question. Although many extant approaches see the needs to investigate these questions[2], less studies exist that have presented results. However, the development of proper study measures is being recognized by recent presented works[3].

A research topic in the field of VR, which is currently topical, is denoted with immersive analytics. It is derived from the research field visual analytics, which tries to include the human perception into analytics tasks. This process is also well-known as visual data mining[4]. Immersive analytics includes topics from the fields of data visualization, visual analytics, virtual reality, computer graphics, and human-computer interaction[5]. Recent advantages in head-mounted displays (HMD) led to improved possibilities for exploring data in an immersive way. Along these trends, new challenges and research questions emerge like the development of new interaction systems, the need to investigate user fatigue or the development of sophisticated 3D visualizations[6]. In a previous publication[6], important principles of immersive analytics are discussed. In the light of big data, methods like immersive analytics are more and more needed to enable a better analysis of complex data pools. Similar to VR solutions in general, less studies exist that investigate usability aspects of immersive analytics solutions. Furthermore, the domain or field in question should also be considered in such studies. In this work, an immersive analytics prototype was developed, and based on that, a protocol, which investigates the developed solution for Industry 4.0 scenarios. The protocol thereby exploits the experience method[2], which is based on subjective, performance, and physiological aspects. In the protocol at hand, the *subjective aspects* are measured through perceived stress of the study users. Performance, in turn, is measured through the required time and errors that are made to accomplish analysis tasks. Finally, a skin conductance sensor measured physiological parameters. The first two measures will be presented in this work, while the measured skin conductance requires further efforts to be evaluated.

The presented study involves several research fields, particularly including neuroscience aspects and information systems. Interestingly, considerations on neuroscience aspects of information systems have recently garnered attention by several research groups[7,8], showing the demand to explore the use of IT systems also from a cognitive viewpoint. Another field that is relevant for this work constitutes the investigation of human factors of information systems[9,10,11]. In the field of human-computer interaction, instruments exist to investigate the usability of a solution. Note that the System Usability Scale is mainly used in this context[12]. Thinking Aloud Protocols[13] are another widely used study technique to learn more about the use of information systems. Although many approaches exist to measure usability aspects of information systems, and some of them have been presented very long ago[14], still questions emerge that require to investigate new measures or study methods. Therefore, research in this field is very active[12,15,16].

In the following, the reasons will be discussed why two prevalently used methods have not been considered in this work. First, the System Usability Scale was not used. The scale is based on ten questions[17] and its use can be found in several other VR studies[18] as well. As this study mainly aims at the measurement of stress[19], a stress questionnaire was more appropriate. Second, no Thinking Aloud Protocol[20] was used. Although this protocol type has shown its usefulness in general[13], it was not used here as the stress level of study users might increase only due to the fact that the think aloud session must be accomplished in parallel to the use of a heavy and complex VR device. Although these two techniques have not been used, results of other recent studies have been incorporated in the study at hand. For example, in previous works[21,22], the authors distinguish between novices and experts in their studies. Based on the successful outcome of these studies, this protocol makes use of this presented separation of study users. The stress measurement, in turn, is based on ideas of the following works[15,19,21,22].

At first, for the study, a suitable Industry 4.0 scenario must be found for accomplishing analysis tasks. Inspired by another work of the authors[23], two scenarios (i.e., the analysis tasks) have been identified, (1) Detection of Outliers, and (2) Recognition of Clusters. Both scenarios are challenging on one hand, and highly relevant in the context of the maintenance of high-throughput production machines on the other. Based on this decision, six major considerations have driven the study protocol presented in this work:

1. The solution developed for the study will be technically based on mixed reality smartglasses (see **Table of Materials**) and will be developed as a mixed reality application.
2. A suitable test must be developed, which is able to distinguish novices from advanced users.
3. Performance measures must be developed considering time and errors.
4. A desktop application must be developed, which can be compared to the immersive analytics solution.
5. A test must be developed to measure the perceived stress level.
6. In addition to the latter point, features shall be developed to mitigate the stress level while a user accomplishes the procedure of the two mentioned analysis tasks.

Based on the six mentioned points, the study protocol incorporates the following procedure. Outlier Detection and Cluster Recognition Analysis tasks have to be accomplished in an immersive way using mixed reality smartglasses (see **Table of Materials**). Therefore, a new application is developed. Spatial sounds shall ease the performing of the analysis tasks without increasing the mental effort. A voice feature shall ease the navigation in the developed application for the mixed reality smartglasses (see **Table of Materials**). A mental rotation test shall be the basis to distinguish novices from advanced users. The stress level is measured based on a questionnaire. Performance is evaluated based on the (1) time a user requires for the analysis tasks, and based on the (2) errors that were made by a user for the analysis tasks. The performance is compared with the accomplishment of the same tasks in a newly developed and comparable 2D desktop application. In addition, a skin conductance device is used to measure the skin conductance level as a possible indicator for stress. Results to this measurement are subject to further analysis and will not be discussed in this work. The authors revealed in another study with the same device that additional considerations are required[24].

Based on this protocol, the following five research questions (RQs) are addressed:
RQ1: Do spatial imagination abilities of the participants affect the performance of tasks significantly?
RQ2: Is there a significant change of task performance over time?
RQ3: Is there a significant change of task performance when using spatial sounds in the immersive analytics solution?
RQ4: Is the developed immersive analytics perceived stressful by the users?
RQ5: Do users perform better when using an immersive analytics solution compared to an 2D approach?

Figure 1 summarizes the presented protocol with respect to two scales. It shows the developed and used measures and their novelty with respect to the level of interaction. As the latter is the most important aspect when using features in a VR setting, this scale was used to show the novelty of the entire protocol shown in this work. Although the evaluation of the aspects within the two scales is subjective, the overall evaluation is based on the current related work in this context as follows: One principle is the use of abstractions of the environment for a natural interaction, in which the user has become attuned. It can be referred to this principle, as the visualization of point clouds seems to be intuitive for users and the recognition of patterns in such clouds has been recognized as a manageable task in general. Another principle is to overlay affordances[6]. Hereby, the use of spatial sounds, which correlate to the proximity to a searched object, is a primary example. The authors recommend to tune the representations in a way that most information is located in the intermediate zone, which is most important for the human perception. The authors did not include this principle, as they wanted to encourage the user to find the best spot by themselves as well as to try to orientate themselves in a data visualization, which is too large to be shown at once. In the presented approach, no further considerations of the characteristics of the 3D data to be shown were made. For example, if a dimension is assumed to be temporal, scatterplots could be shown[25]. The authors consider this kind of visualization generally interesting in the context of Industry 4.0. However, it has to been focused on a set of visualizations for the study of this work. Moreover, a previous publication[25] focused on the

collaborative analysis of data. In this work, this research question was excluded due to complexity of addressed issues in this study. In the presented setup here, the user is able to explore the immersive space by walking around. Other approaches[26] offer controllers to explore the virtual space. In their study[26], the focus is set on usability via the System Usability Scale (SUS)[17]. A previous publication[27], in turn, has conducted a study for economic experts for immersive analytics, but with virtual reality headsets. They complain about the limited field of view for other devices, such as the used mixed reality smartglasses in this work (see **Table of Materials**). Their findings show that beginners in the field of VR were able to use the analytics tool efficiently. This matches with the experiences of this study, although in this work beginners were not classified with VR or gaming experience. In contrast to most VR solutions, mixed reality is not fixed to a position as it allows to track the real environment. VR approaches such as[28] mention the use of special chairs for a 360° experience to free the user from his desktop. The authors of[29] indicate that perception issues influence the performance of immersive analytics; for example, by using shadows. For the study at hand, this is not feasible, as the used mixed reality smartglasses (see table of materials) are not able to display shadows. A workaround could be a virtual floor, but such a setup was out of the scope of this study. A survey study in the field of immersive analytics[30] identified 3D scatterplots as one of the most common representations of multi-dimensional data. Altogether, the aspects shown in **Figure 1** cannot be found currently as a protocol to investigate usability aspects of immersive analytics for Industry 4.0 scenarios.

PROTOCOL:

All materials and methods were approved by the Ethics Committee of Ulm University, and were carried out in accordance with the approved guidelines. All participants gave their written informed consent.

1. Establish Appropriate Study Environment

NOTE: A controlled environment was conducted to cope with the complex hardware setting. The used mixed reality smartglasses (see **Table of Materials**) and the laptop for the 2D application were therefore explained to the study participants.

1.1. Check the technical solution before each participant; set in default mode. Prepare the questionnaires and place next to a participant.

1.2. Have participants solve tasks from the use cases outlier detection and cluster recognition in one session (i.e., average time was 43 min). Furthermore, start half of them with the outlier detection, while starting the other half of them with the cluster recognition task. Randomly determine the initial use case. After that, alternate uses cases.

1.3. Start the study by welcoming the participants and introducing the goal of the study, as well as the overall procedure.

1.4. Have participants using the skin conductance measurement device (see **Table of**

Materials) adhere to a short resting phase, to receive a baseline measure.

1.5. Have all participants fill out the State-Trait Anxiety Inventory (STAI) questionnaire[31], prior to the start of the experiment.

1.5.1. Next, have the participants perform the mental rotation test (see **Figure 4**, this test evaluated the spatial imagination abilities), which was the basis to distinguish high from low performers (high performers are advanced users, while low performers are novices), followed by the spatial sound test to measure spatial hearing abilities of a participant.

NOTE: A median split of the test scores in the mental rotation test[32] was used to distinguish low from high performers.

1.6. Randomly separate participants into two groups; to start either with the task on outlier detection or cluster recognition, while continuing with the other use case afterwards. Further note that for the cluster recognition task, half of the participants first started with the used mixed reality smartglasses (see **Table of Materials**), and then used the 2D application, while the other half first started with the 2D application, and then used the mixed reality smartglasses (see **Table of Materials**). Make the decision for the group assignment as before (i.e., first decision randomly, then alternate).

NOTE: For the outlier's detection task, note that one group received sound support, while the other part of the group received no sound support. The decision was determined like before (i.e., first decision randomly, then alternate).

1.7. Concluding the session, have participants answer the State-Trait Anxiety Inventory (STAI) questionnaire[31] again, as well as the self-developed questionnaire, and a demographic questionnaire. Altogether, a session took about 40 to 50 minutes.

1.8. Store the data, which was automatically recorded by each application, on the laptop's storage after the session was accomplished.

2. Study Protocol for Participants

2.1. Prepare the experiment (see **Figure 2** for the room of the experiment) for each participant. Present the desktop PC, the used mixed reality smartglasses, and hand out the questionnaires.

2.2. Inform the participants that the experiment will take 40 to 50 minutes.

2.3. Decide for the first participant whether a skin conductance measurement is done. Alternate this decision for the next participants based on the first decision. Prepare the skin conductance measurement device[33] and inform the participant to put on the device. Request a short resting phase from participants to receive a baseline measure for their stress level.

2.4. Request participants to fill out the State-Trait Anxiety Inventory (STAI) questionnaire[31] and inform them that it measures the perceived stress during the experiment.

2.5. Conduct a mental rotation test.

2.5.1. Inform participants that their mental rotation capabilities are evaluated and usher them in front of a desktop computer. Inform participants about the test procedure. Note that they had to identify similar objects that had different positions in a simulated 3D space.

2.5.2. Inform participants that only two of the five shown objects are similar and that they will have 2 minutes for the entire test. Inform participants that seven tasks could be accomplished within the given 2 minutes and tell them that performance measures are recorded for each accomplished task.

2.6. Evaluate spatial sound abilities.

2.6.1. Inform participants that their spatial sound abilities are evaluated and usher them in front of a desktop computer. Inform participants about the test procedure. Explain to participants that six sound samples must be detected, which will be played for 13 seconds each.

2.6.2. Inform participants that they have to detect the direction (analogously to the four compass directions) of which the sound is coming from.

2.7. Evaluate outlier detection skills.

2.7.1. Request participants to put on the mixed reality smartglasses. Explain to them that outliers must be found within the world created for the mixed reality smartglasses.

2.7.2. Further inform them that an outlier is a red-marked point, all other points are white-marked. Explain then to them that they must direct their gaze to the red-colored point to detect it.

2.7.3. Further inform the participants that not only visual help is provided, additionally environmental sounds support them to find outliers. Provide the information to the participants that they have to accomplish 8 outlier tasks, meaning that 8 times within the virtual world, the red-colored point has to be found.

2.7.4. Tell participants which information will be recorded: required time for each task, length of walking, and how their final moving position is looking like related to their starting position. Finally tell participants that the red-marked point changes to green if it was detected (see **Figure 3**).

2.8. Evaluate cluster recognition skills.

2.8.1. Decide for the first participant whether firstly to use the mixed reality smartglasses or to usher the participant to a desktop computer. Alternate the decision for the next participants based on the first decision.

2.8.2. Request participants to put on the mixed reality smartglasses. Inform participants how to find clusters within the world created with the used mixed reality smartglasses. Emphasize to the participants that they had to distinguish between overlapping clusters by moving around them.

2.8.3. Explain to participants that they can navigate in the virtual world and around the clusters using voice commands. Finally tell participants that they had to detect six clusters.

2.8.4. Request participants to remove the used mixed reality smartglasses. Usher participants to a desktop computer and tell them to use the software shown on the screen of the desktop computer. Inform them that the same type of clusters like shown in the used mixed reality smartglasses had to be detected using the software on the desktop computer (see **Figure 7** and **Figure 8**).

2.9. Request participants to fill out three questionnaires, namely the State-Trait Anxiety Inventory (STAI) questionnaire[31], a self-developed questionnaire to gather subjective feedback, and a demographic questionnaire to gather information about them.

2.10. Request participants to remove the skin conductance measurement device[33] if they were requested in the beginning to put it on.

2.11. Relieve participants from the experiment by saying thanks for the participation.

REPRESENTATIVE RESULTS:

Setting up Measures for the Experiment

For the outlier detection task, the following performance measures were defined: time, path, and angle. See **Figure 6** for the measurements.

Time was recorded until a red-marked point (i.e., the outlier) was found. This performance measure indicates how long a participant required to find the red-marked point. Time is denoted as time (in milliseconds) in the results.

While participants tried to find the red-marked point, their walking path length was determined. The basis of this calculation was that the used mixed reality smartglasses (see **Table of Materials**) collect the current position as a 3D vector relatively to the starting position at a frame rate of 60 frames per second. Based on this, the length of path a participant had walked could be calculated. This performance measure indicates whether participants walked a lot or not. Path is denoted as Pathlength in the results. Based on the Pathlength, three more performance measures were derived: PathMean, PathVariance, and BoundingBox. PathMean denotes the average speed of participants in meter per frame, PathVariance the erraticness of a movement, and BoundingBox

denotes whether participants had intensively used their bounding box. The latter is determined based on the maximum and minimum positions of all movements (i.e., participants that often change their walking position revealed higher BoundingBox values).

The last value that was measured is denoted with AngleMean and constitutes a derived value of the angle, which is denoted with AngleMean. The latter denotes the rotation between the current position and the starting position of a participant at a frame rate of 60 per second. Based on this, the average rotation speed in degrees per frame was calculated. Derived on this value, the erraticness of the rotation using the variance was calculated, which is denoted as AngleVariance.

To summarize the purposes of the calculated path and angle values, the path indicates whether users walk much or not. If they are not walking much, it might indicate their lack of orientation. The angle, in turn, should indicate whether participants make quick or sudden head movements. If they are doing sudden head movements multiple times, again, this might indicate a lack of orientation.

The state version of State-Trait Anxiety Inventory (STAI) questionnaire[31] was used to measure the state anxiety, a construct similar to state stress. The questionnaire comprises 20 items and was handed out before the study started, as well as afterwards to evaluate the changes in the state anxiety. For the evaluation of this questionnaire, all positive attributes were flipped (e.g., an answer '4' becomes a '1'), and all answers are summed up to a final STAI score. The skin conductance was measured for 30 randomly selected participants by using the skin conductance measurement device (see **Table of Materials**)[33].

For the cluster detection task, the following performance measures were defined: time and errors. Time was recorded until the point in time at which participants reported how many clusters they have detected. This performance measure indicates how long participants needed to find clusters. Time is denoted as Time (in milliseconds). Errors are identified in the sense of a binary decision (true/false). Either the number of reported clusters was correct (true) or not correct (false). Errors are denoted with errors.

The state version of State-Trait Anxiety Inventory (STAI) questionnaire[31] was used to measure the state anxiety, a construct similar to state stress. The questionnaire comprises 20 items and was handed out before the study started, as well as afterwards to evaluate the changes in the state anxiety. For the evaluation of this questionnaire, all positive attributes were flipped (e.g., an answer '4' becomes a '1'), and all answers are summed up to a final STAI score. The skin conductance was measured for 30 randomly selected participants by using the skin conductance measurement device (see **Table of Materials**)[33].

After the two task types have been accomplished, a self-developed questionnaire was handed out at the end of the study to ask for participant's feedback. The questionnaire is shown in **Table 1**. Furthermore, a demographic questionnaire asked about gender, age, and education of all participants.

Overall Study Procedure and Study Information

The overall conducted study procedure is illustrated in **Figure 9**. 60 participants joined the study. The participants were mostly recruited at Ulm University and software companies from Ulm. The participating students were mainly from the fields of computer science, psychology, and physics. Ten were female and 50 were male.

Based on the mental rotation pretest, 31 were categorized as low performers, while 29 were categorized as high performers. Specifically, 7 females and 24 males were categorized as low performers, while 3 females and 26 males were categorized as high performers. For the statistical evaluations, 3 software tools were used (see **Table of Materials**).

Frequencies, percentages, means, and standard deviations were calculated as descriptive statistics. Low and high performers were compared in baseline demographic variables using Fisher's exact tests and t-Tests for independent samples. For RQ1 -RQ5, linear multilevel models with the full maximum likelihood estimation were performed. Two levels were included, where level one represents the repeated assessments (either in outlier detection or cluster recognition), and level two the participants. The performance measures (except errors) were the dependent variables in these models. In RQ 1, also Fisher's exact tests for the error probabilities was used. In RQ3, the time differences when using spatial sounds versus no sounds were investigated. The STAI scores were evaluated using t-Tests for dependent samples for RQ4. In RQ5, the effect of the 2D application versus the used mixed reality smartglasses (see table of materials) was investigated, using McNemar's test for the error probability. All statistical tests were performed two tailed; the significance value was set to $P<.05$.

The skin conductance results have not been analyzed and are subject to future work. Importantly, the authors revealed in another study with the same device that additional considerations are required[24].

For the mental rotation test, the differences between participants were used to distinguish low from high performers. For the spatial ability test, all participants showed good scores and therefore were all categorized to high performers with respect to their spatial abilities.

At first, important results of the participants are summarized: Low and high performers showed no differences in their baseline variables (gender, age, and education). Descriptively, the low performers had a higher percentage of female participants than high performers and high performers were younger than low performers. **Table 2** summarizes the characteristics about the participants.

Regarding results for RQ1, for the cluster recognition task, low and high performers did not differ significantly for the 2D application (4 errors for low and 2 errors for high performers) and the 3D approach (8 errors for low and 2 errors for high performers). For the outlier's detection task, high performers were significantly faster than low performers. In addition, high performers required a shorter walking distance to solve the tasks. For the outlier's task, **Table 3** summarizes the

detailed results.

Regarding results for RQ2, significant results emerged only for the outlier's detection task. The BoundingBox, the Pathlength, the PathVariance, the PathMean, the Angle-Variance, and the AngleMean increased significantly from task to task (see **Table 4**). The recorded time, in turn, did not change significantly from task to task using the mixed reality smartglasses (see **Table of Materials**).

Regarding results for RQ3, based on the spatial sounds, the participants were able to solve the tasks in the outlier detection case quicker than without using spatial sounds (see **Table 5**).

Regarding results for RQ4, at the pre-assessment, the average state on the STAI scores were M = 44.58 (SD = 4.67). At post-assessment, it was M = 45.72 (SD = 4.43). This change did not attain statistical significance (p = .175). Descriptive statistics of the answers in the self-developed questionnaire are presented in **Figure 10**.

Regarding results for RQ5, the mixed reality smartglasses (see **Table of Materials**) approach indicates significantly faster cluster recognition times than using a desktop computer (see T**able 6**). However, the speed advantage when using the mixed reality smartglasses (see **Table of Materials**) was rather small (i.e., in a milliseconds range).

Finally note that the data of this study can be found at[36].

FIGURE AND TABLE LEGENDS:

Figure 1: Investigated Aspects on the scale Interaction versus Novelty. The figure shows the used measures and their novelty with respect to the interaction level.

Figure 2: Pictures of the study room. Two pictures of the study room are presented.

Figure 3: Detected Outlier. The screenshot shows a detected outlier.

Figure 4: Example of the mental rotation test. The screenshot shows the 3D-objects participants were confronted with; i.e., two out of five objects in different positions with the same object structure had to bet detected. This figure has been modified based on this work[35].

Figure 5: Setting for the Spatial Ability Test. In (**A**), the audio configuration for the task Back is shown, while, in (**B**), the schematic user interface of the test is shown. This figure has been modified based on this work[35].

Figure 6: Illustration of the Setting for the Task Outlier's Detection. Three major aspects are shown. First, the outliers are illustrated. Second, performance measures are shown. Third, the way how the sound support was calculated is shown. This figure has been modified based on this work[35].

Figure 7: Illustration of the Setting for the Task Cluster Recognition. Consider the scenarios A-C for a better impression, participants had to change their gaze to identify clusters correctly. This figure has been modified based on this work[35].

Figure 8: Illustration of the Setting for the Task Cluster Recognition in Matlab. The figure illustrates clusters provided in Matlab, which was the basis for the 2D desktop application.

Figure 9: Overall Study Procedure at a Glance. This figure presents the steps participants had to accomplish, in their chronological order. This figure has been modified based on this work[35].

Figure 10: Results of the self-developed questionnaire (see Table 1). The results are shown using box plots. This figure has been modified based on this work[35].

Table 1: Self-developed questionnaire for user feedback. It comprises 7 question. For each question, participants had to determine a value within a scale from 1-10, whereby 1 means a low value (i.e., bad feedback), and 10 a high value (i.e., a very good feedback).

Table 2: Participant sample description and comparison between low and high performers in baseline variables. The table shows data to the three demographic questions on gender, age, and education. In addition, the results of the two pretests are presented.

Table 3: Results of the Multilevel Models for RQ1 (Outlier Detection Using the Smartglasses). The table shows statistical results of RQ1 for the outlier's detection task (for all performance measures).

Table 4: Results of the Multilevel Models for RQ2 (Outlier Detection Using the Smartglasses). The table shows statistical results of RQ2 for the outlier's detection task (for all performance measures).

Table 5: Results of the Multilevel Models for RQ3 (Outlier Detection Using the Smartglasses). The table shows statistical results of RQ3 for the outlier's detection task (for all performance measures).

Table 6: Results of the Multilevel Models for RQ5 (Cluster Recognition Using the Smartglasses). The table shows statistical results of RQ5 for the cluster recognition task (for all performance measures).

DISCUSSION:

Regarding the developed mixed reality smartglasses (see **Table of Materials**) application, two aspects were particularly beneficial. The use of spatial sounds for the outlier's detection task was positively perceived on one hand (see the results of RQ3). On the other, the use of voice commands was also perceived positively (see **Figure 10**).

Regarding the study participants, although the number of recruited participants was rather small

for an empirical study, the number is competitive compared to many other works. Therefore, a larger-scale study is planned based on the shown protocol. However, as it showed its feasibility for 60 participants, more participants are expected to reveal no further challenges. It was discussed that the selection of participants could be broader (in the sense of the fields the participants are coming from) and that the number of baseline variables to distinguish between high and low performers could be higher. On the other, if these aspects are changed to higher numbers, the protocol itself has not to be changed profoundly.

In general, the revealed limitations do not affect the conduction of a study based on the protocol shown in this work, they only affect the recruitment and the used questions for the demographic questionnaire. However, one limitation of this study is nevertheless important: the overall required time to finish the experiment for one participant is high. On the other hand, as the participants did not complain about the wearing comfort or that the test device is burdening them too much, the time of conducting the overall protocol for one participant can be considered acceptable. Finally, in a future experiment, several aspects have to be added to the protocol. In particular, the outlier detection task should also be evaluated in the 2D desktop application. Furthermore, other hardware devices like the used mixed reality smartglasses (see **Table of Materials**) must be also evaluated. However, the protocol seems to be beneficial in a broader sense.

The following major insights were gained for the presented protocol. First, it showed its feasibility for evaluating immersive analytics for a mixed-reality solution. Specifically, the used mixed reality smartglasses (see **Table of Materials**) revealed their feasibility to evaluate immersive analytics in a mixed-reality application for Industry 4.0 scenarios. Second, the comparison of the developed used mixed reality smartglasses (see **Table of Materials**) application with a 2D desktop application was helpful to investigate whether the mixed-reality solution can outperform an application that does not make use of VR techniques. Third, the measurement of physiological parameters or vital signs should be always considered in such experiments. In this work, stress was measured using a questionnaire and a skin conductance device. Although the latter worked technically properly, the authors revealed in another study with the same device that additional considerations are required[24]. Fourth, the spatial ability test and the separation of high and low performers was advantageous. In summary, although the presented protocol seems to be complex at a first glance (see **Figure 9**), it showed its usefulness technically. Regarding the results, it also revealed its usefulness.

As the detection of outliers and the recognition of clusters are typical tasks in the evaluation of many high-dimensional data sets in Industry 4.0 scenarios, their use in an empirical study is representative for this field of research. The protocol showed that these scenarios can be well-integrated in a usability study on immersive analytics. Therefore, the used setting can be recommended for other studies in this context.

As the outcome of the shown study showed that the use of a mixed-reality solution based on the used smartglasses (see **Table of Materials**) is useful to investigate immersive analytics for Industry 4.0 scenarios, the protocol might be used for other usability studies in the given context

as well.

ACKNOWLEDGMENTS:
The authors have nothing to acknowledge.

DISCLOSURES:
The authors have nothing to disclose.

REFERENCES:
1. Korinth, M., Sommer-Dittrich, T., Reichert, M., Pryss, R. Design and Evaluation of a Virtual Reality-Based Car Configuration Concept. In Science and Information Conference, 169-189 Springer, Cham. (2019).
2. Whalen, T. E., Noël, S., Stewart, J. Measuring the human side of virtual reality. In IEEE International Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems, 2003. VECIMS'03. IEEE 8-12 (2003).
3. Martens, M. A. et al. It feels real: physiological responses to a stressful virtual reality environment and its impact on working memory. *Journal of Psychopharmacology*. **33** (10), 1264-1273 (2019).
4. Keim, D. A. Information visualization and visual data mining. *IEEE transactions on Visualization and Computer Graphics*. **8** (1), 1-8 (2002).
5. Dwyer, T. et al. Immersive analytics: An introduction. In Immersive analytics, 1-23. Springer, Cham (2018).
6. Moloney, J., Spehar, B., Globa, A., Wang, R. The affordance of virtual reality to enable the sensory representation of multi-dimensional data for immersive analytics: from experience to insight. *Journal of Big Data*. **5** (1), 53 (2018).
7. Davis, F. D., Riedl, R., Vom Brocke, J., Léger, P. M., Randolph, A. B. *Information Systems and Neuroscienc*e. Springer (2018).
8. Huckins, J. F. et al. Fusing mobile phone sensing and brain imaging to assess depression in college students. *Frontiers in Neuroscience*. **13**, 248 (2019).
9. Preece, J. et al. Human-computer interaction. Addison-Wesley Longman Ltd (1994).
10. Card, S. K. The psychology of human-computer interaction. CRC Press (2018).
11. Pelayo, S., Senathirajah, Y. Human factors and sociotechnical issues. *Yearbook of Medical Informatics*. **28** (01), 078-080 (2019).
12. Bangor, A., Kortum, P., Miller, J., Determining what individual SUS scores mean: adding an adjective rating scale. *Journal of Usability Studies*. **4** (3), 114-123 (2009).
13. Krahmer, E., Ummelen, N. Thinking about thinking aloud: A comparison of two verbal protocols for usability testing. *IEEE Transactions on Professional Communication*. **47** (2), 105-117 (2004).
14. Hornbæk, K. Current practice in measuring usability: Challenges to usability studies and research. *International Journal of Human-Computer Studies*. **64** (2), 79-102 (2006).
15. Peppa, V., Lysikatos, S., Metaxas, G. Human-Computer interaction and usability testing: Application adoption on B2C websites. *Global Journal of Engineering Education*. **14** (1), 112-118 (2012).
16. Alwashmi, M. F., Hawboldt, J., Davis, E., Fetters, M. D. The iterative convergent design for

617 mobile health usability testing: mixed-methods approach. *JMIR mHealth and uHealth*. **7** (4),
618 e11656 (2019).
619 17. Affairs, Assistant Secretary for Public. System Usability Scale (SUS). 6. September 2013, /how-
620 to-and-tools/methods/system-usability-scale.html (2013).
621 18. Fang, Y. M., Lin, C. The Usability Testing of VR Interface for Tourism Apps. *Applied Sciences*. **9**
622 (16), 3215 (2019).
623 19. Pryss, R. et al. Exploring the Time Trend of Stress Levels While Using the Crowdsensing Mobile
624 Health Platform, TrackYourStress, and the Influence of Perceived Stress Reactivity: Ecological
625 Momentary Assessment Pilot Study. *JMIR mHealth and uHealth*. **7** (10), e13978 (2019).
626 20. Zugal, S. et al. Investigating expressiveness and understandability of hierarchy in declarative
627 business process models. *Software & Systems Modeling*. **14** (3), 1081-1103 (2015).
628 21. Schobel, J. et al. Learnability of a configurator empowering end users to create mobile data
629 collection instruments: usability study. *JMIR mHealth and uHealth*. **6** (6), e148 (2018).
630 22. Schobel, J., Probst, T., Reichert, M., Schickler, M., Pryss, R. Enabling Sophisticated Lifecycle
631 Support for Mobile Healthcare Data Collection Applications. *IEEE Access*. **7**, 61204-61217 (2019)..
632 23. Hoppenstedt, B. et al. Dimensionality Reduction and Subspace Clustering in Mixed Reality for
633 Condition Monitoring of High-Dimensional Production Data. *Sensors*. **19** (18), 3903 (2019).
634 24. Winter, M., Pryss, R., Probst, T., Reichert, M. Towards the Applicability of Measuring the
635 Electrodermal Activity in the Context of Process Model Comprehension: Feasibility Study. *Sensors*
636 **20**, 4561 (2020).
637 25. Butscher, S., Hubenschmid, S., Müller, J., Fuchs, J., Reiterer, H. Clusters, trends, and outliers:
638 How immersive technologies can facilitate the collaborative analysis of multidimensional data. In
639 Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1-12 (2018).
640 26. Wagner Filho, J.A., Rey, M.F., Freitas, C.M.D.S., Nedel, L. Immersive analytics of dimensionally-
641 reduced data scatterplots. In 2nd Workshop on Immersive Analytics (2017).
642 27. Batch, A. et al. There is no spoon: Evaluating performance, space use, and presence with
643 expert domain users in immersive analytics. *IEEE Transactions on Visualization and Computer*
644 *Graphics*. **26** (1), 536-546 (2019).
645 28. Cliquet, G., Perreira, M., Picarougne, F., Prié, Y., Vigier, T. Towards hmd-based immersive
646 analytics. *HAL*. https://hal.archives-ouvertes.fr/hal-01631306 (2017).
647 29. Luboschik, M., Berger, P., Staadt, O. On spatial perception issues in augmented reality based
648 immersive analytics. In Proceedings of the 2016 ACM Companion on Interactive Surfaces and
649 Spaces. 47-53 (2016).
650 30. Fonnet, A., Prié, Y. Survey of Immersive Analytics. *IEEE Transactions on Visualization and*
651 *Computer Graphics* (2019).
652 31. Spielberger, C.D., Gorsuch, R.L., Lushene, R.E. STAI Manual for the Stait-Trait Anxiety
653 Inventory (self-evaluation questionnaire), Palo Alto, CA, USA: Consulting Psychologist. **22**, 1-24
654 (1970).
655 32. Vandenberg, S.G., Kuse, A.R. Mental rotations, a group test of three-dimensional spatial
656 visualization. *Perceptual Motor Skills*. **47** (2), 599-604 (1978).
657 33. Härtel, S., Gnam, J.-P., Löffler, S., Bös, K. Estimation of energy expenditure using
658 accelerometers and activity-based energy models-Validation of a new device. *European Review*
659 *of Aging and Physical Activity*. **8** (2), 109-114 (2011).
660 34. Gautier, L. RPY2: A Simple and Efficient Access to R from Python. Accessed: Feb. 09, 2020.

661 [Online]. Available: https://sourceforge.net/projects/rpy/ (2020).
662 35. Hoppenstedt, B. et al. Applicability of immersive analytics in mixed reality: Usability study.
663 *IEEE Access*. **7**, 71921-71932 (2019).
664 36. Burkhard Hoppenstedt. Applicability of Immersive Analytics in Mixed Reality: Usability Study.
665 IEEE Dataport. http://dx.doi.org/10.21227/6mme-0526 (2019).

Figure 1

Click here to access/download;Figure;Figure1.pdf

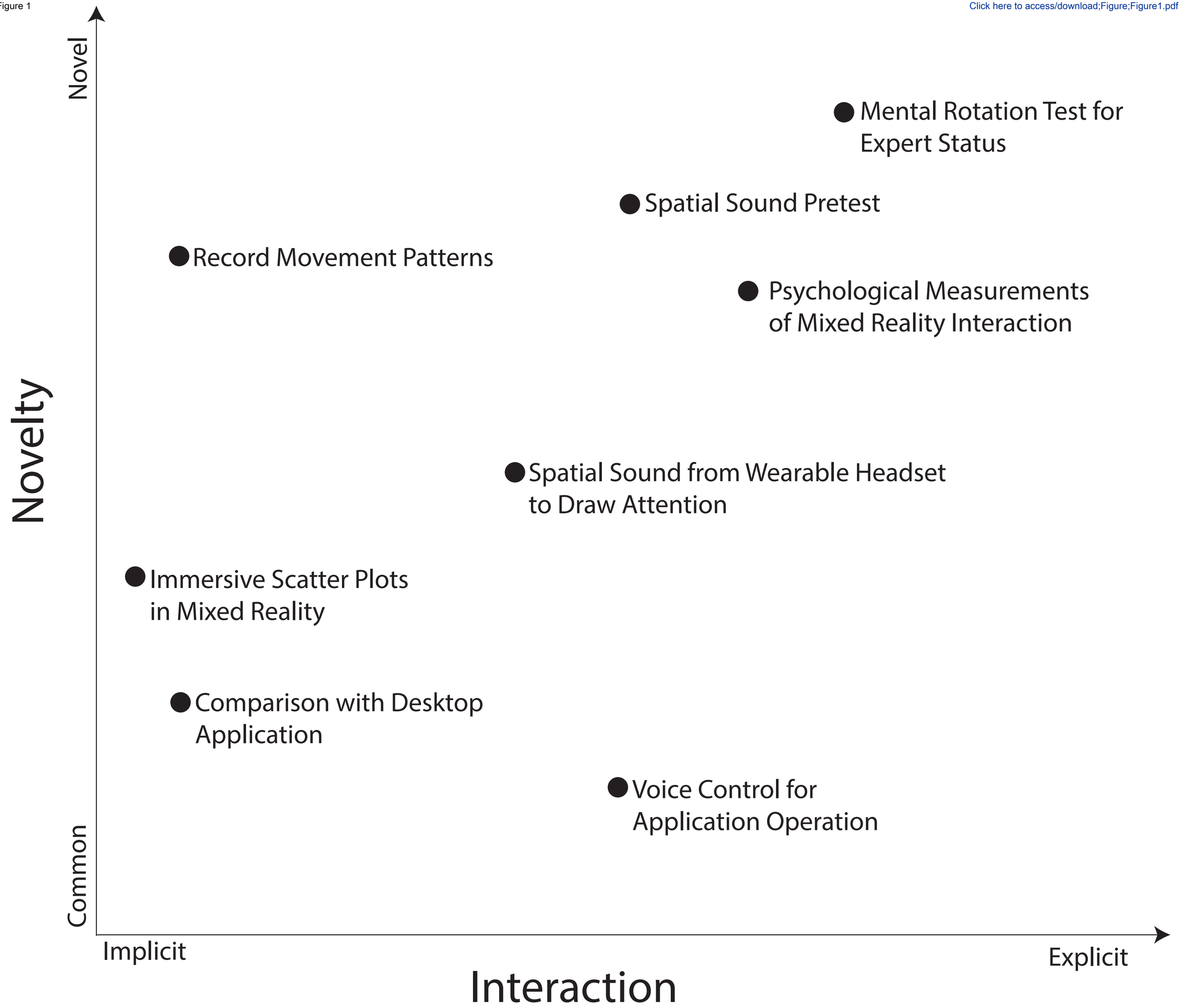


Figure 2

Click here to access/download;Figure;Figure2.pdf

Click here to access/download;Figure;Figure3.pdf

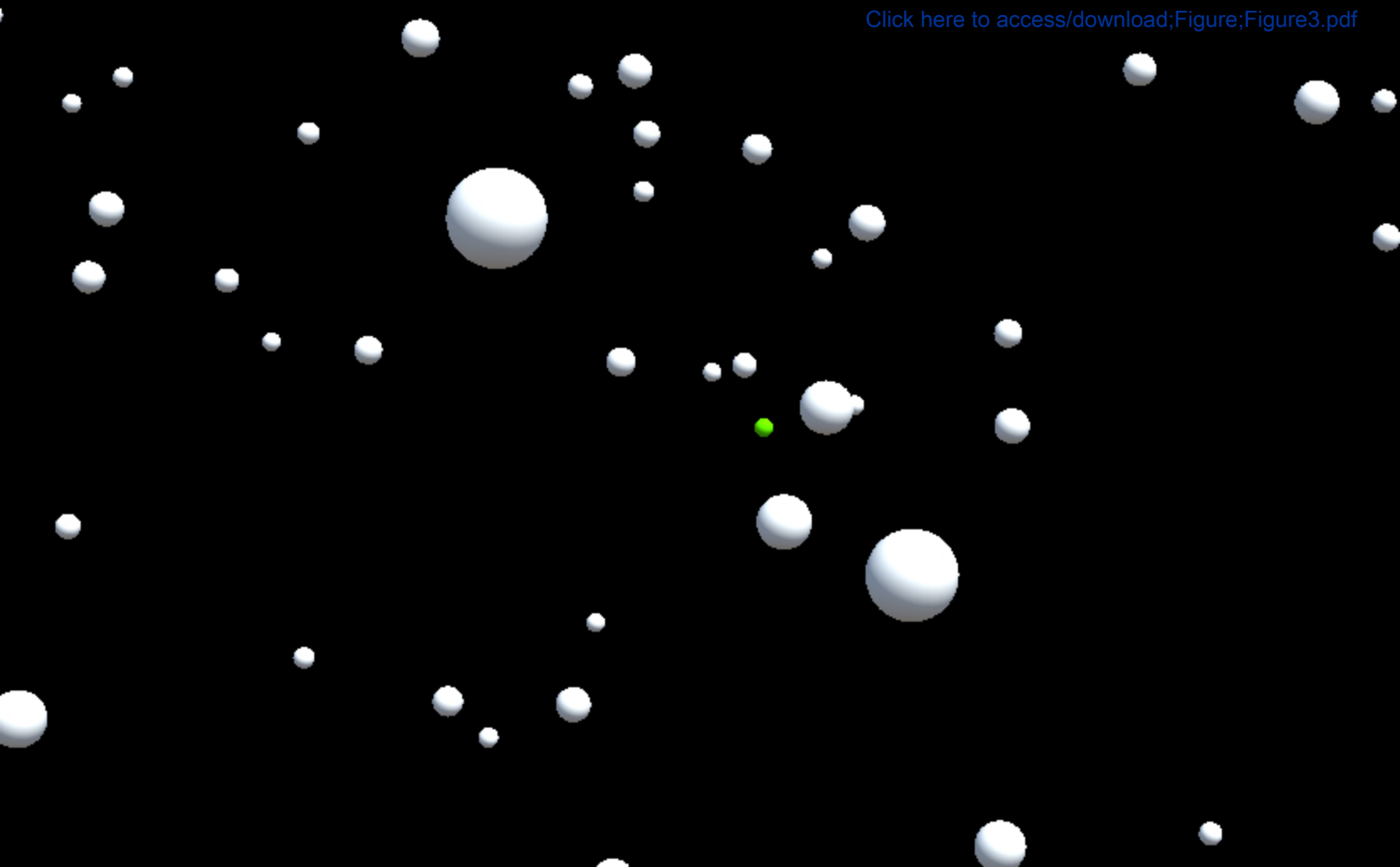


Figure 4

Click here to
access/download;Figure;Figure4.pdf

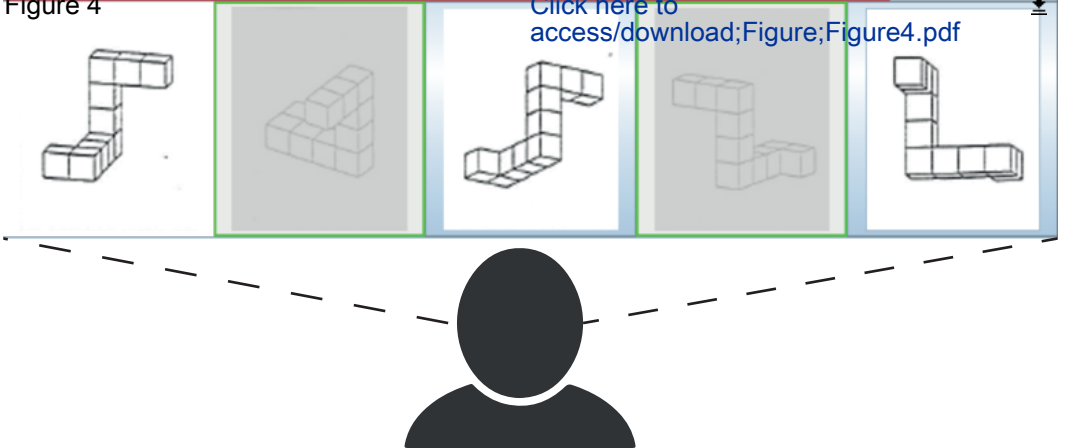
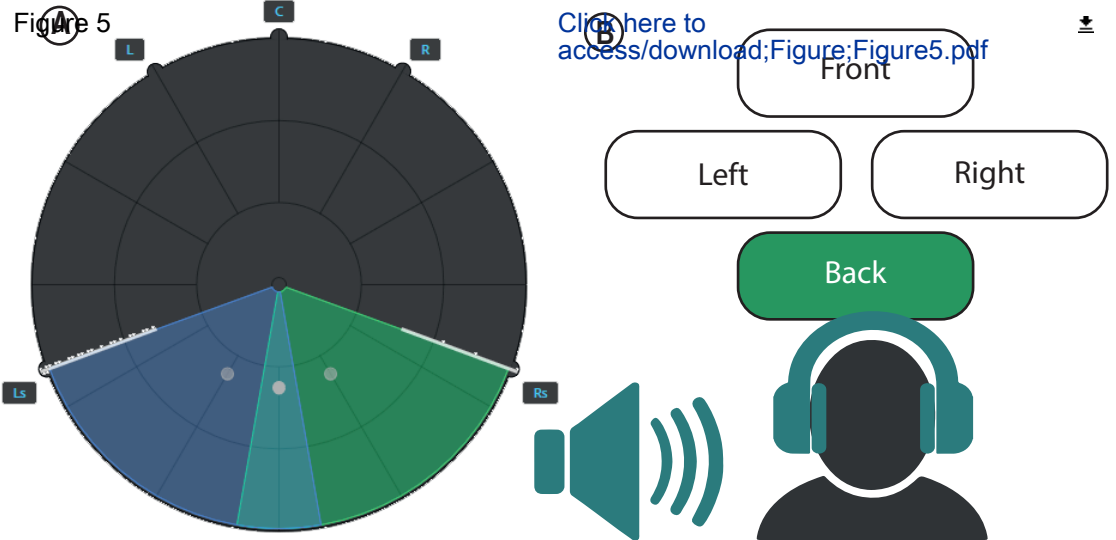


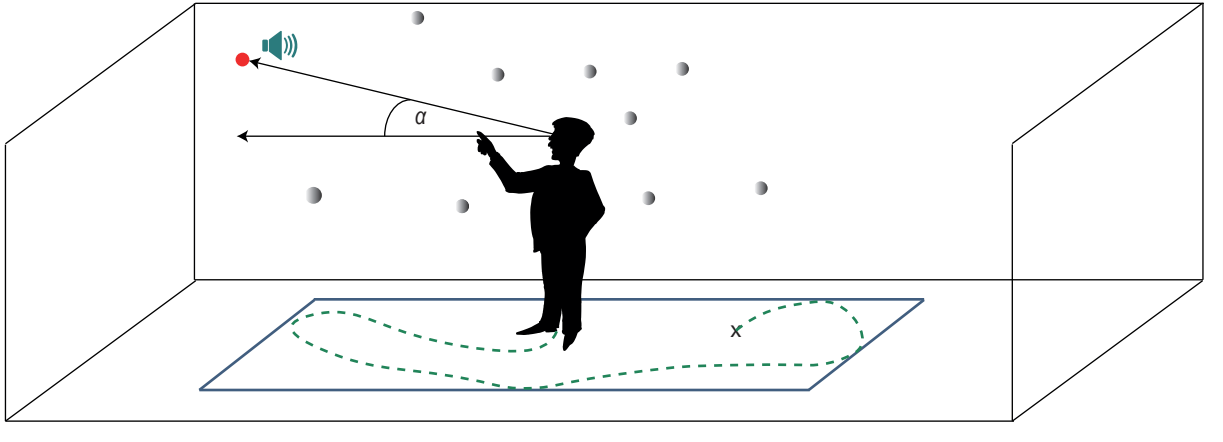
Figure 5



Click here to access/download;Figure;Figure5.pdf

Figure 6

Click here to access/download;Figure;Figure6.pdf



Anomaly Point

Random Point

α Angle of View

Spatial Sound

X Start Position

Walking Path

Bounding Box

Figure 7

Click here to access/download;Figure;Figure7.pdf

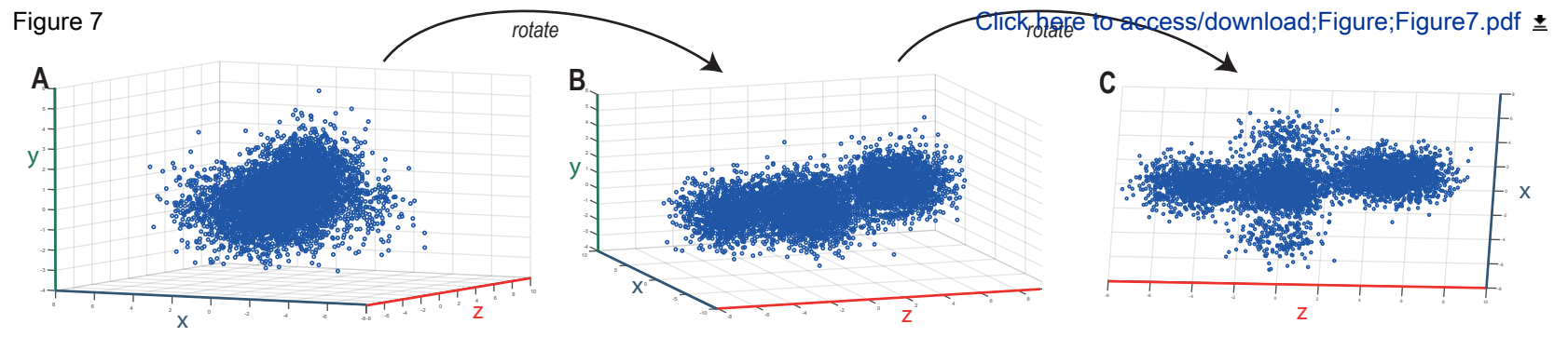


Figure 8

Click here to access/download;Figure;Figure8.pdf

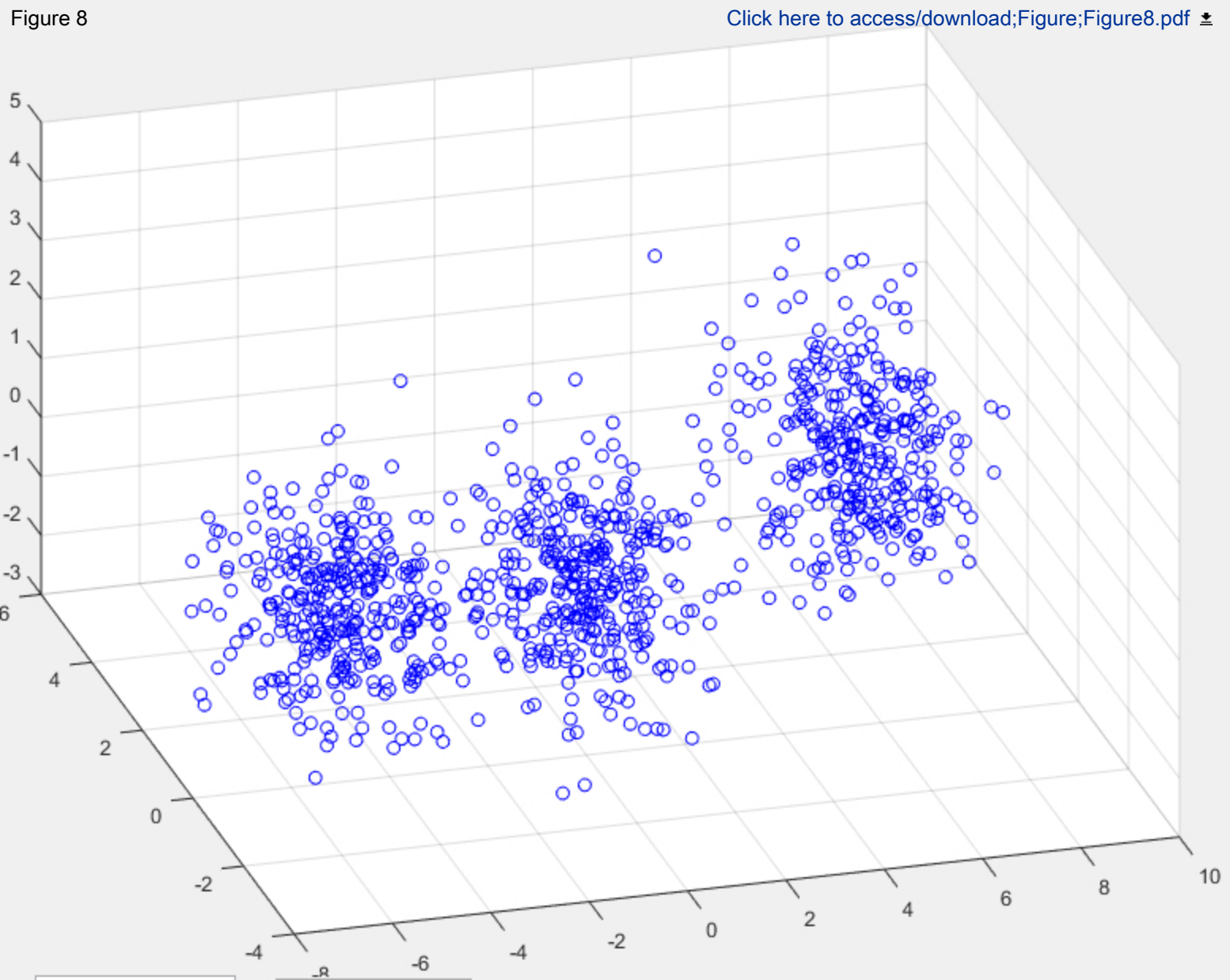


Figure 9

Click here to access/download;Figure;Figure9.pdf

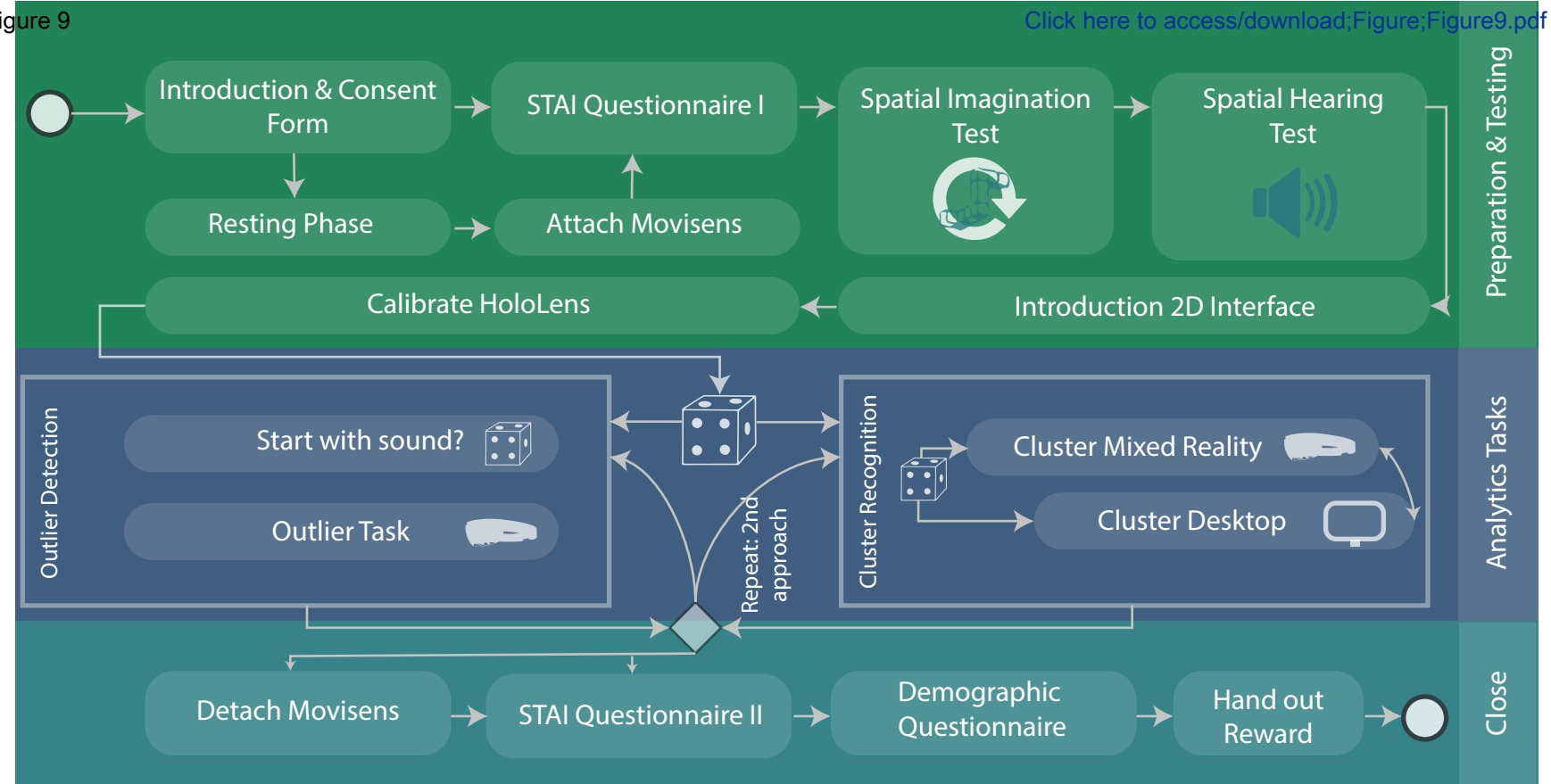
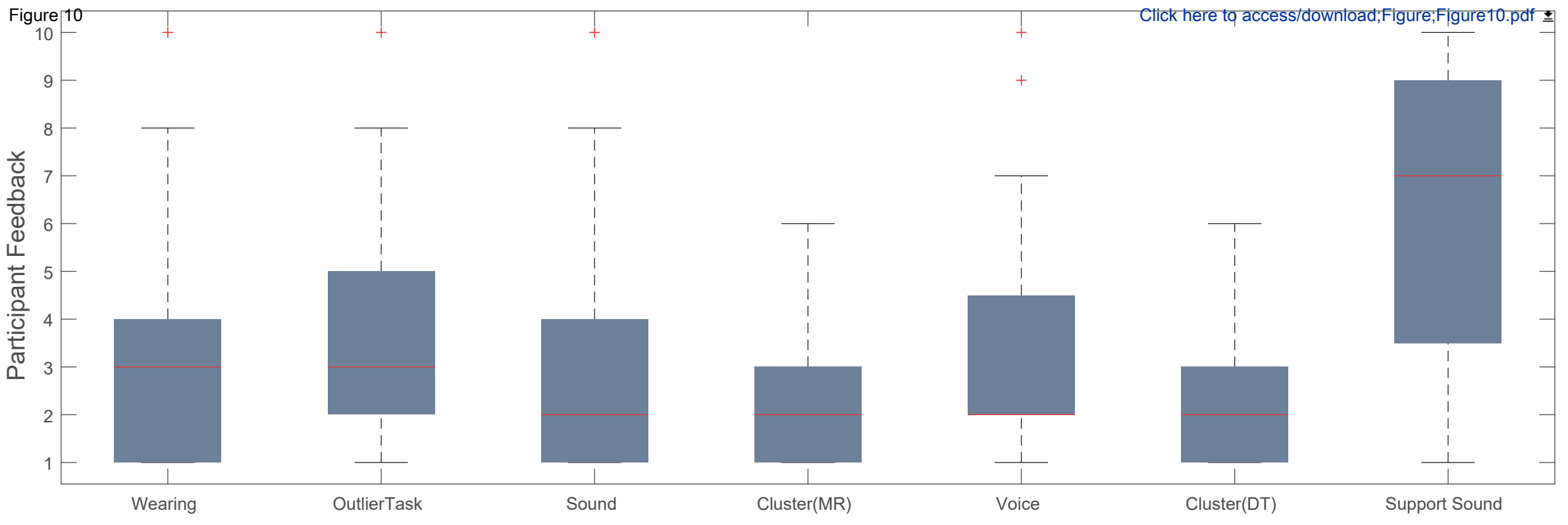


Figure 10

Click here to access/download;Figure;Figure10.pdf



Table

Click here to access/download;Table;Table1.xlsx

#Question	Question
1	As how stressful did you experience wearing the glasses?
2	How stressful was the outlier's task?
3	As how stressful did you experience the spatial sounds?
4	How stressful was the task finding clusters in Mixed Reality?
5	How stressful was the task finding clusters in the desktop approach?
6	How stressful was the usage of the voice commands?
7	Did you feel supported by the spatial sounds?

Target	Scale	Meaning
Wearing	1-10	10 means high, 1 means low
Outliers	1-10	10 means high, 1 means low
Sound	1-10	10 means high, 1 means low
Cluster MR	1-10	10 means high, 1 means low
Cluster DT	1-10	10 means high, 1 means low
Voice	1-10	10 means high, 1 means low
Sound	1-10	10 means high, 1 means low

Table

Click here to access/download;Table;Table2.xlsx

Variable	Low performer (n=31) (n=31)
Gender, n(%)	
Female	7 (23%)
Male	24 (77%)
Age Category, n(%)	
<25	1 (3%)
25-35	27 (87%)
36-45	0 (0%)
46-55	1 (3%)
>55	2 (6%)
Highest Education, n(%)	
High School	3 (10%)
Bachelor	7 (23%)
Master	21 (68%)
Mental Rotation Test, Mean (SD)	
Correct Answers	3.03 (1.40)
Wrong Answers	2.19 (1.47)
Spatial Hearing Test, Mean (SD) ©	
Correct Answers	4.39 (1.09)
Wrong Answers	1.61 (1.09)

a:Fisher's Exact Test
b:Two-sample t-test
c: SD Standard Deviation

High performer (n=29)	P Value
3 (10%)	
26 (90%)	.302 (a)
5 (17%)	
21 (72%)	
2 (7%)	
0 (0%)	
1 (3%)	.099 (a)
5 (17%)	
6 (21%)	
18 (62%)	.692 (a)
5.31 (0.76)	.001 (b)
1.21 (0.56)	.000 (b)
4.31 (1.00)	.467 (b)
1.69 (1.00)	.940 (b)

Table

Click here to access/download;Table;Table3.xlsx

Variable	Estimate
BoundingBox for low performer across tasks	2,224
Alteration of BoundingBox for high performer across tasks	+.131
Time for low performer across tasks	20,919
Alteration of Time for high performer across tasks	-3,863
Pathlength for low performer across tasks	5,637
Alteration of Pathlength for high performer across tasks	-1,624
PathVariance for low performer across tasks	4.3E-4
Alteration of PathVariance for high performer across tasks	+4.3E-6
PathMean for low performer across tasks	.0047
Alteration of PathMean for high performer across tasks	+3.8E-5
AngleVariance for low performer across tasks	.0012
Alteration of AngleVariance for high performer across tasks	-2.7E-5
AngleMean for low performer across tasks	.015
Alteration of AngleMean for high performer across tasks	-3.0E-4

(a) SE = Standard Error

SE (a)	Result
.438	t(60.00) = 5.08; p<.001
.630	t(60.00) = .21; p=.836
1,045	t(60.00) = 20.02; p<.001
1,503	t(60.00) = -2.57; p=.013
.613	t(60.00) = 9.19; p<.001
.882	t(60.00) = -1.84; p=.071
4.7E-5	t(65.15) = 9.25; p<.001
6.7E-5	t(65.15) = .063; p=.950
5.3E-4	t(60.00) = 8.697; p<.001
7.7E-4	t(60.00) = .05; p=.960
7.3E-5	t(85.70) = 16.15; p<.001
1.0E-4	t(85.70) = -.26; p=.796
.001	t(60.00) = 14.27; p<.001
1.5E-3	t(60.00) = -.20; p=.842

Table

Click here to access/download;Table;Table4.xlsx

Variable	Estimate	SE (a)
BoundingBox at first task	.984	.392
Alteration of BoundingBox from task to task	+.373	.067
Time at first task	19,431	1,283
Alteration of Time from task to task	-.108	.286
Pathlength at first task	3,903	.646
Alteration of Pathlength from task to task	+.271	.131
PathVariance at first task	3.1E-4	3.7E-5
Alteration of PathVariance from task to task	+3.5E-5	4.5E-6
PathMean at first task	.0033	4.2E-4
Alteration of PathMean from task to task	+4.1E-4	5.2E-5
AngleVariance at first task	.001	5.7E-5
Alteration of AngleVariance from task to task	+4.1E-5	6.5E-6
AngleMean at first task	.0127	8.1E-4
Alteration of AngleMean from task to task	+6.1E-4	9.0E-5

(a) SE = Standard Error

Result
t(138.12) = 2.51; p=.013
t(420.00) = 5.59; p<.001
t(302.08) = 15.11; p<.001
t(420.00) = -.37; p=.709
t(214.81) = 6.05; p<.001
t(420.00) = 2.06; p=.040
t(117.77) = 8.43; p<.001
t(455.00) = 7.90; p<.001
t(88.98) = 7.66; p<.001
t(420.00) = 7.81; p<.001
t(129.86) = 17.92; p<.001
t(541.75) = 6.34; p<.001
t(82.17) = 15.52; p<.001
t(420.00) = 6.86; p<.001

Table

Click here to access/download;Table;Table5.xlsx

Variable	Estimate	SE (a)
BoundingBox without sound across tasks	2,459	.352
Alteration of BoundingBox with sound across tasks	-.344	.316
Time without sound across tasks	20,550	1,030
Alteration of time with sound across tasks	-2,996	1,319
Pathlength without sound across tasks	5,193	.545
Alteration of Pathlength with sound across tasks	-.682	.604
PathVariance without sound across tasks	.0004	3.5E-5
Alteration of PathVariance with sound across tasks	+1.3E-5	2.2E-5
PathMean without sound across tasks	.005	4.0E-4
Alteration of PathMean with sound across tasks	+1.4E-4	2.5E-4
AngleVariance without sound across tasks	.0012	5.4E-5
Alteration of AngleVariance with sound across tasks	+3.3E-5	3.1E-5
AngleMean without sound across tasks	.0145	7.8E-4
Alteration of AngleMean with sound across tasks	+6.0E-4	4.3E-4

(a) SE = Standard Error

Result
t(93.26) = 6.98; p<.001
t(420.00) = -1.09; p=.277
t(161.17) = 19.94; p<.001
t(420.00) = -2.27; p=.024
t(121.81) = 9.54; p<.001
t(420.00) = -1.13; p=.260
t(79.74) = 12.110; p<.001
t(429.20) = .592; p=.554
t(73.66) = 11.35; p<.001
t(420.00) = .56; p=.575
t(101.32) = 21.00; p<.001
t(648.56) = 1.07; p=.284
t(70.17) = 18.51; p<.001
t(420.00) = 1.39; p=.166

Table

Click here to access/download;Table;Table6.xlsx

Variable	Estimate	SE (a)	Result
Time with desktop across tasks	10,536	.228	t(156.43) = 46
Alteration of time with Hololens across tasks	-.631	.286	t(660.00) = -2.

(a) SE = Standard Error

.120; p<.001
206; p=.028

Click here to access/download;Table of Materials;JoVE_Materials.xls

Name of Material/ Equipment	Company	Catalog Number
edaMove	movisens	
HoloLens	Microsoft	
Matlab R2017a	MathWorks	
RPY2	GNU General Public License v2 or later (GPLv2+) (GPLv2+)	
SPSS 25.0	IBM	

Comments/Description

https://pypi.org/project/rpy2/

Rebuttal Letter
Click here to access/download;Rebuttal Letter;CV17082020.docx

Julius-Maximilians-
UNIVERSITÄT
WÜRZBURG

Institute of Clinical Epidemiology and Biometry (ICE-B)
Director: Prof. Dr. P. Heuschmann

Institute of Clinical Epidemiology and Biometry
Josef-Schneider-Str.2 / D7, 97080 Würzburg

Prof. Dr. Rüdiger Pryss
Professor of Medical Informatics
Tel.: 0931 / 201-46471
Fax: 0931 / 201-647310
ruediger.pryss@uni-wuerzburg.de
www.epidemiologie.uni-wuerzburg.de

Editor
Journal of Visualized Experiments

Würzburg, 17th of August, 2020

Submission of Manuscript: "*Evaluating Usability Aspects of A Mixed Reality Solution for Immersive Analytics in Industry 4.0 Scenarios*" after the first review round.

Dear Editor,

please, find attached our manuscript entitled with „Evaluating Usability Aspects of A Mixed Reality Solution for Immersive Analytics in Industry 4.0 Scenarios". We have mostly rewritten the first submission entitled with "Evaluating the Usability of A Mixed Reality Setting for Immersive Analytics Based on Performance Measures and Perceived Stress." We adhered to the author guidelines of JoVE and we have addressed the comments of the five reviewers. We are thankful for their valuable comments, which we have addressed in our revision. We think that the paper is interesting for a broad audience and hope to convince the reviewers in this statement.
Importantly, in this revision, we have mostly rewritten the entire paper for the sake of readability and the mentioned reviewer comments.

Changes in this revision:

The manuscript will benefit from thorough language revision as there are a number of grammatical errors throughout. Please thoroughly review the manuscript and edit any errors. Themost common types of errors in your manuscript are: awkward phrasing, grammatical errors from translation into English, run-on sentences.	Manuscript was completely rewritten.
Textual Overlap: Significant portions show significant overlap with previously published work. Please re-write the text on lines to 172-178, 188-202, 209-216, 327-332, 343-347, 365-375, 383-358, 399-409 avoid this overlap.	Most of the parts were rewritten; however, as we were invited based on an existing paper, some parts must slightly overlap since the results cannot be changed.

Introduction: Please use paragraph style throughout, avoid the numbered lists.	Changed
Protocol Language: The JoVE protocol should be almost entirely composed of numbered short steps (2-3 related actions each) written in the imperative voice/tense (as if you are telling someone how to do the technique, i.e. "Do this", "Measure that" etc.). Any text that cannot be written in the imperative tense may be added as a brief “Note” at the end of the step(please limit notes). Please re-write your ENTIRE protocol section accordingly. Descriptive sections of the protocol can be moved to Representative Results or Discussion. The JoVE protocol should be a set of instructions rather a report of a study. Any reporting should be moved into the representative results.	The part with the instructions was rewritten and highlighted.
Please note that your protocol will be used to generate the script for the video, and must contain everything that you would like shown in the video. **Please add more specific details (e.g. button clicks for software actions, numerical values for settings, etc) to your protocol steps.** There should be enough detail in each step to supplement the actions seen in the video so that viewers can easily replicate the protocol. Some examples: Line 219: What are the contents of the questionnaire.	
Protocol Numbering: 1) Please add numbering to yourprotocol section. Subheadings should be labeled 1. This should be followed by 1.1. and then 1.1.1. if necessary and all steps should be lined up at the left margin with no indentations. 2) Add a one-line space between each protocol step.	Added.
Protocol Highlight: After you have made all of the recommended changes to your protocol (listed above), please re-evaluate the length of your protocol section. Please highlight ~2.5 pages or less of text (which includes headings and spaces) in yellow, to identify which steps should be visualized to tell the most cohesive story of your protocol steps.	Done and changed.

1) Please ensure that the title best represents the highlighted portion of the protocol. 2) The highlighting must include all relevant details that are required to perform the step. For example, if step 2.5 is highlighted for filming and the details of how to perform the step are given in steps 2.5.1 and 2.5.2, then thesub-steps where the details are provided must be included in the highlighting. 3) The highlighted steps should form a cohesive narrative, that is, there must be a logical flow from one highlighted step to the next. 4) Please highlight complete sentences (not parts of sentences). Include sub-headings and spaces when calculating the final highlighted length. 5) Notes cannot be filmed and should be excluded from highlighting.	
Discussion: JoVE articles are focused on the methods and the protocol, thus the discussion should be similarly focused. Please ensure that the discussion covers the following in detail and in paragraph form (3-6 paragraphs): 1) modifications and troubleshooting, 2) limitations of the technique, 3) significance with respect to existing methods, 4) future applications and 5) critical steps within the protocol.	Discussion section was restructured.
References: 1) Please spellout journal names. 2) Please make sure that your references comply with JoVE instructions for authors. Citation formatting should appear as follows: (For 6 authors or less list all authors. For more than 6 authors, list only the first author then *et al.*): [Lastname, F.I., LastName, F.I., LastName, F.I. Article Title. *Source*. **Volume** (Issue), FirstPage – LastPage, (YEAR).]	All checked and changed.
Commercial Language: JoVE is unable to publish manuscripts containing commercial sounding language, including trademark or registered trademark symbols (TM/R) and the mention of company brand names before	Changed.

an instrument or reagent. Examples of commercial sounding language in your manuscript are Microsoft HoloLens, Matlab R2017a, RPY223 , and SPSS 25.0 1) Please use MS Word's find function (Ctrl+F), to locate and replace all commercial sounding language in your manuscript with generic names that are notcompany-specific. All commercial products should be sufficiently referenced in the table of materials/reagents. You may use the generic term followed by "(see table of materials)" to draw the readers' attention to specific commercial names.	
Table of Materials: 1) Please revise the table of the essential supplies, reagents, and equipment. The table should include the name, company, and catalog number of all relevant materials/software in separate columns in an xls/xlsx file. Please include items such as 2) Sort the list alphabetically. • If your figures and tables are original and not published previously or you have already obtained figure permissions, please ignore this comment. If you are re-using figures from a previous publication, you must obtain explicit permission to re-use the figure from the previous publisher (this can be in the form of a letter from an editor or a link tothe editorial policies that allows you to re-publish the figure). Please upload the text of the re-print permission (may be copied and pasted from an email/website) as a Word document to the Editorial Manager site in the "Supplemental files (as requested by JoVE)" section. Please also cite the figure appropriately in the figure legend, i.e. "This figure has been modified from [citation]."	Not relevant.
Reviewer 1: The protocol seems to be very specific for the application of the authors. Many immersive analytics applications do not utilize spatial sound.	We have changed the title and revised the Introduction.
Reviewer 1: Having the instructor decide randomly for/against measurements is not guaranteeing randomness.	Statement was refined.

Reviewer 1: It is not clear why skin conductance is not measured for all participants when the instrument is available and used for some.	Because of the handling, but we also stated that in needs a further experiment.
Reviewer 1: I do not understand why a generic test protocol should include a mental rotation test as well as a spatial sound abilities test.	See Figure 1 and the revised Introduction.
Reviewer 1: Why is the outlier detection test not done with a PC (but only with the HL)?	We assumed that the experience for the outliers is not like the clusters and did nor evaluate this. We added a sentence to the outlook.
Reviewer 1: The study protocol measures the usability of very specific tasks (outlier detection, cluster detection). It is not clear why this is a generic protocol for measuring the usability of immersive analytics applications.	Title was changed and we made clear that we focus on Industry 4.0 scenarios.
Reviewer 1: When the stress test doesn't show any statistically significant difference before/after, why include it in a generic protocol to test immersive analytics usability?	To evaluate indicators for the mental workload, which heavily depends on stress.
Reviewer 1: I think that the discussion section over-generalizes the study. The HL has severe limitations for immersive analytics applications (FOV, performance) when comparedto VR setups. The immersive visualizations that were evaluated are IMHO only a small glimpse of all potential immersive analytics applications. The spatial sound test contributions is not clear.	We changed the title, the introduction, and the discussion and made the focus clearer and narrower.
Reviewer 1: Why was only the HL used for outlier detection?	We assumed that the experience for the outliers is not like the clusters and did nor evaluate this. We added a sentence to the outlook.
Reviewer 1: The 2D software UI isn't included in the material. Hence, it is very hard to evaluate if the 3D/immersive aspect is important or if some other design flaw of the 2D UI was the reason for its poor performance	See Figure 8.
Reviewer 1: The manuscript often talks about virtual reality but then uses the MS HoloLens (which is AR/MR).	We changed many statements.
Reviewer 1: The bounding box metrics is not clear. What does it mean that a bounding box is heavily used.	Explanation was changed.
Reviewer 1: Using the median to split between High and low performers is ok. But stating that high perfromers are experts is a stretch. Especially in cases when the actual results	We changed the term to advanced users; very good point.

of low performers and high performers is very similar.	
We say thanks to Reviewer 1 for the valuable comments.	
Reviewer 2: The manuscriptlacks detailed information about the dataset used in the clustering and outlier detection tasks. In order to make the protocol reproducible, it is important to share the dataset publicly or to give guidelines to generate a similar dataset. The assumption that protocol would work with any dataset with an outlier or any dataset that could be clustered, seems not to be correct. The dataset is a key aspect to evaluate certain aspects that the authors explore in their pilot study.	See Reference [36]
Reviewer 2: The paper refers to the commercial product movisens sensor but it does not provide enough information about the characteristics of a generic sensor that would be recommended to reproduce the protocol.	Because of the handling, but we also stated that in needs a further experiment.
Reviewer 2: No software or data repository was provided. Having access to open source code would help other developers to reproducethe protocol and also to understand in detail how the metrics are computed and processed. This would be a great contribution to the field.	See Reference [36]
We say thanks to Reviewer 2 for the valuable comments.	
Reviewer 3: It was not clear about the hypothesis and control group. Is the hypothesis that immersive analytics is more effective? Was this compared with a 2D version of the task?	We have revised the introduction and discussion. Yes, the hypothesis was that immersive analytics outperforms a 2D solution
Reviewer 3: Analytic tasks were limited to spatial discrimination, like positional identification (visual & audio) within 3D scatterplots. Data viz involves much more, like comparison and causal inference.	We added a new section to the introduction to explain our thoughts on this. We added also Figure 1 for this purpose.
Reviewer 3: Somedetails of study results were reported, but results are not summarized and assessed. Did this preliminary study show that immersive analytics is effective?	We have rewritten the summary as well as many parts to present that our results show that immersive analytics is beneficial for Industry 4.0 scenarios.
Reviewer 3: There was no control for the effect of the chosen h/w solution (HoloLens). Would other VR/AR devices elicit different behaviors?	Very good point; due the basically complex setting, this will be addressed in future work.

Reviewer 3: Needs better literature background about the field of immersive analytics. Citations of similar research? Definition of and motivation for IA?	We have added a new section to the Introduction.
Recommendation: - Accept. The authors have made a disciplined effort to design, conduct and analyze their experimental protocol. - Quality lab experiments in immersive analytics is very important to the advancement of data viz research. The VR/AR/xR technology has rapidly evolved so that creative solutions to complex data viz can now be both feasible and effective. Hence, this type of disciplined research should be encouraged. The field is beyond simplistic statements that 2D data viz issuperior to 3D. We need to know when and why.	We say warmly thanks to Reviewer 3 for the valuable comments.
Reviewer 4: Why using augmented reality and not only virtual reality for this study? Indeed, the objective of the study is to visually analyze a 3D scatter plot. The real environment is not used and even not described. The use of AR in this context, and more generally for immersive analytics, is not well justified and limits the interest of the study.	We have changed the entire Introduction for this purpose and also added Figure 1.
Reviewer 4: A screenshot and a real illustration of the AR and desktop application would be very helpful to better understand the study. How was the room (visual environment, color of the walls, luminosity) in which theexperimentation was run?	See Figures 2,3, and 8
Reviewer 4: The skin conductance was measured only for a part of the subjects. Why?	Because of the handling, but we also stated that in needs a further experiment.
Reviewer 4: The choice of the path and angle measures are not well motivated. These measures can be interesting to understand differences in behavior but authors claim that they reflect performance, why? These measures are not really discussed then.	We added a paragraph explaining these aspects.
Reviewer 4: In general, results could be more discussed and authors can try to propose some hypothesis to explain them.	We added a new section to the Introduction and revised the discussion as well as the results part.
We say thanks to Reviewer 4 for the valuable comments.	

Reviewer 5: Please elaborate on the sound use. On page 5 it is stated "To better find the red-marked points, a constant sound of 44100 Hz was produced". How is this high frequency sound, well above the maximum audible sound frequency (~20000 Hz) useful? Also, what are the 6 audio samples mentioned in line 232 (frequency, level, etc. attributed)? Please elaborate. Figure 2 does not provide much information that would allow replication of the experiment.	The 44100 HZ refer to the sampling rate (https://en.wikipedia.org/wiki/44,100_Hz), while the used tone pitch was 250HZ
Reviewer 5: The abstract talks about immersive analytics and virtual-reality settings. However, the study uses Microsoft HoloLens that is not fully immersive and use mixed-reality settings. Please better align the abstract with the content of the paper. Similarly, Introduction section does not even mentioned mixed reality and related research. Similarly, page 11 (lines 459-460) claims feasibility for virtual reality based environment which was not the case for the conductedexperiment. Please be consistent.	Very good points, the Title, the Introduction, the Abstract and the Discussion were entirely changed for this purpose.
Reviewer 5: Spatial imagination ability is mention on page 4 (RQ1) and in Figure 5. However, there is not clear description how is spatial imaging ability defined and measured? Is that the Spatial Ability Test? Outlier detection test? Mental rotation test?	The mental rotation test is shown in Figure 5, this test was used for evaluating the Spatial imagination abilities. In addition, we added a sentence and a reference to the figure that shows the test.
Reviewer 5: Industry 4.0 is mentioned in the abstract and on page 3 as "an emerging application field that is highly interested in analyzing data using immersive analytics." Could you elaborate and expand on this example to make a stronger case for practical use of immersive analytics?	Very good point, the Title, the Introduction, the Abstract and the Discussion were entirely changed for this purpose.
Reviewer 5: What is the impact of limited field of view of Microsoft HoloLens device given the immersive analytics focus?	We added a section in the Introduction as well as a sentence in the outlook that further devices have to be investigated.
Reviewer 5: Could you at least elaborate on what is expected from skin conductance data? What are the physiological factors you hope to measure? Also, the term Electro Dermal Activity (EDA) should be used instead of skin conductance.	See our statement for the need of a further experiment.

Reviewer 5: Figure6 shows sound and support sound. What is the difference?	The two aspects refer to this questionnaire after the study (https://docs.google.com/forms/d/e/1FAIpQLScoe7CyxlLfxeghicVI6rleU-cN40I1KHE94tn1oSbTaS__kg/viewform) Sound means the answer to the question: How stressful was the sound that supported you to find outliers? Support Sound means the answer to the question: How helpful was the sound support to find outliers?
Reviewer 5: Please provide more discussion about RQ1-RQ5 findings on page 10 (just referring to tables makes it more difficult to understand the results).	We added more information to these points.
We say thanks to Reviewer 5 for the valuable comments.	

Yours sincerely,

Prof. Dr. Rüdiger Pryss
Corresponding author
Institute of Clinical Epidemiology and Biometry
University of Würzburg
ruediger.pryss@uni-wuerzburg.de