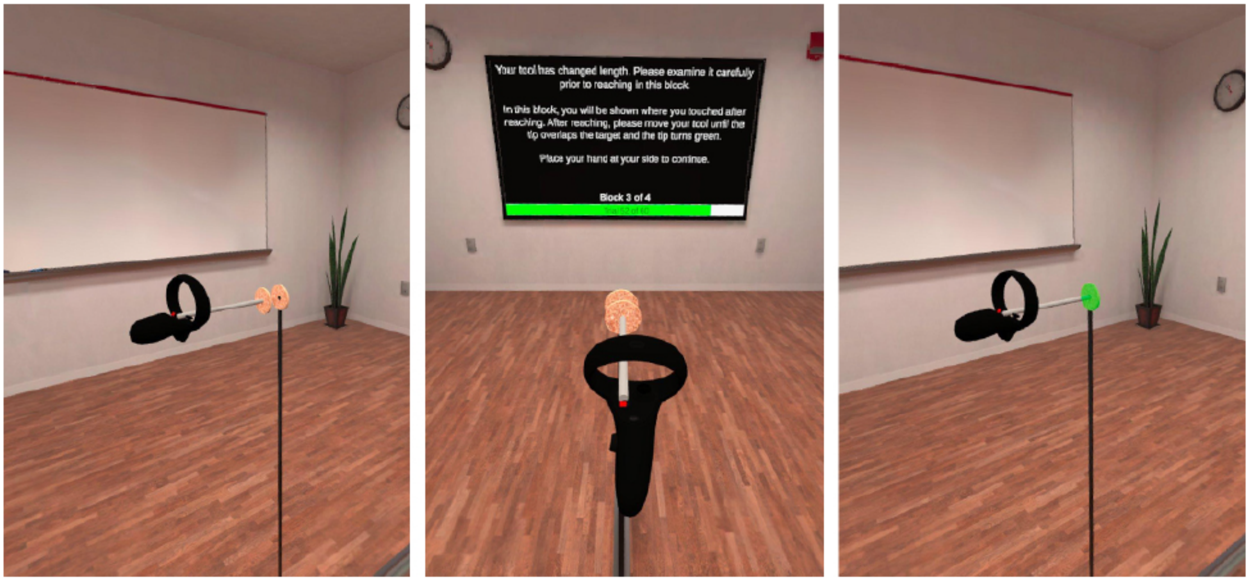
Objects May Be Farther Than They Appear: Depth Compression Diminishes Over Time with Repeated Calibration in Virtual Reality

Kristopher Kohm, Sabarish V. Babu, Christopher Pagano, and Andrew Robb



Your tool has changed length. Please examine it carefully prior to reaching in this block.

In this block, you will be shown where you touched after reaching. After reaching, please move your tool until the tip overlaps the target and the tip turns green.

Place your hand at your side to continue.

Block 3 of 4

Trial 52 of 60

Fig. 1: This figure shows the virtual environment during the calibration block of the depth perception experiment. The image on the left shows the participant getting feedback after their blind reach. The image in the middle is the same as the image on the left but from the participant's point of view. The image on the right shows the participant correcting their reach based on the feedback they received. Feedback was provided by showing participants where the true location of the target was. Once shown, participants could correct their error by placing the tip of their tool on the target location. When they successfully corrected their error (i.e. calibrated) the tip of the tool glowed green as shown in the right-most image.

Abstract—Prior research in depth perception and perceptuo-motor calibration have primarily focused on participants completing experiments in single sessions and therefore do not empirically evaluate changes over time. Further, these studies do not typically take into account the amount of experience that the participants have in virtual reality (VR) prior to participation, the role of experience during participation, or calibration that may occur throughout the experiment session. In this contribution, we conducted a novel empirical evaluation of how calibration affects perception-action coordination over time. We recruited novice VR users and they completed eight sessions of a depth perception reaching experiment over the course of 12 weeks. During these experiments, we examined how participants' ability to estimate depth in a virtual environment changed as they gradually gained experience. While previous literature has shown that participants tend to underestimate distances, we found that this underestimation diminished over time as they gained experience in the virtual environment. Our study highlights the need for carrying out VR studies over time and the influence that longitudinal calibration can have on spatial perception in long-term VR experiences.

Index Terms—Distance estimation, calibration, perception, longitudinal, virtual reality

1 Introduction

Distance compression, where objects are judged to be closer than they actually are, is a well established phenomenon in virtual reality (VR) head-mounted displays (HMDs), particularly when judging distances to objects in the near field [25,49]. Despite extensive research, it is not entirely clear what causes distance compression in HMDs. Proposed factors include field-of-view (FOV) [12,28], the absence of a visible body [32,43], the accommodation-vergence conflict [6], graphical quality [42], display resolution [10], geometric distortions [32,55], and the absence of known reference objects [2,41].

It has been demonstrated that *calibration* can reduce the effect of distance compression, potentially even eliminating it [22]. Calibration is the process by which people use feedback to scale their action capabilities to the affordances present in their environment after changes to their body schema or while wielding tools [13]. Calibration has been used to improve distance estimation under normal reaching circumstances [18], when a user's reach boundary is extended via a tool [19],

- *Kristopher Kohm is with Clemson University, USA. E-mail: kckohm@clemson.edu.*
- *Sabarish V. Babu is with Clemson University, USA. E-mail: sbabu@clemson.edu.*
- *Christopher Pagano is with Clemson University, USA. E-mail: cpagano@clemson.edu.*
- *Andrew Robb is with Clemson University, USA. E-mail: arobb@clemson.edu.*

Manuscript received 11 March 2022; revised 11 June 2022; accepted 2 July 2022. Date of publication 01 September 2022; date of current version 03 October 2022. Digital Object Identifier no. 10.1109/TVCG.2022.3203112

and when a user's reach boundary is extended by modifying the bodily proportions of their self-avatar [13]. The efficacy of calibration can be further enhanced by repeating the calibration process until errors in distance judgments have almost disappeared [22].

While it has been shown that repeated calibration can improve distance estimation, this effect has not been studied in a longitudinal context. What information we have about calibration and distance perception in VR has generally come from single-session experiments, typically lasting 30 minutes to an hour. It is unclear whether the gains established by calibration will continue to improve distance perception in future uses of VR or whether distance compression will reassert itself and thus require re-calibration each time VR is used. This is a particularly important question given that the ever-increasing availability of VR devices and applications means that most people who use VR applications will likely have done so in the past or will do so again in the future. If repeated calibration does affect distance estimation and depth compression in the long-term, it is important to understand how this effect manifests to better understand how distance judgments may differ between users with different levels of experience in VR.

To examine how repeated calibration over time influences distance estimation, we conducted a study where participants without prior VR experience were given an Oculus Quest to use for 12 weeks. During these 12 weeks, participants completed a near-field distance estimation activity eight times allowing us to measure the magnitude of the change in depth compression over time. During each session, participants completed four blocks of reaching trials with tools of various lengths. All reaches were blind reaches and the target distance was always within their maximum reach boundary. Feedback was provided to participants in one of the blocks after they had performed their reach so that participants could calibrate their reaching behavior.

Our hypotheses include:

H1: The overall effect of depth compression will decrease with time as participants repeatedly re-calibrate their distance judgments each time they use an HMD.

H2: Participants will be more accurate with a tool that is a length they have previously calibrated to than with a tool of a different length over time.

H3: With feedback, participants will reach full calibration (i.e. where their performance is no longer improving) faster over time.

- **H3a**: Participants' baseline performance will improve over time, explaining why they calibrated faster.
- **H3b**: Participants' rate of calibration will improve over time, explaining why they calibrated faster.

H4: There will be a difference in accuracy between the pre- and post-calibration blocks due to carryover effects, and the effects will diminish over time.

2 Related Works

2.1 Depth Perception in VR

As mentioned, several works have studied distance estimation in virtual environments. Gagnon et al. examined how reaching out and up with feedback could improve distance estimation in immersive virtual environments (IVEs) [22]. The researchers found that participants initially overestimated distances but that their estimations became more accurate with feedback across several blocks. Overestimation of distances is an uncommon finding in work related to near-field depth perception in VR. Witmer and Kline analyzed how observers estimated distances in virtual environments when moving and when stationary [63]. They found that participants underestimated distance in the real and virtual environment, but that the underestimation was more profound in the virtual environment. Napieralski et al. compared near-field distance estimation in the real world against an IVE using physical reaches and verbal reports of measures [39]. They found that participants consistently underestimated distances in both conditions, but that verbal responses in the virtual world underestimated distance less than in the real world. This work also showed that verbal reports for distance estimation was less accurate than action-based measurements. Based on this work, our study used reaching as an action based measurement for near-field distance estimation.

Altenhoff et al. had participants give verbal and reach-based estimates in an IVE in a pre-calibration, calibration, and post-calibration block [1]. They showed that depth perception estimations were more accurate after the calibration block suggesting that the effects of calibration carried over between the blocks. This experimental design is similar to the one we are presenting in this paper, but we repeated these blocks over a prolonged period of time and added a block before pre-calibration to measure carryover effects of calibration between sessions.

2.2 Calibration

In the ecological approach to perception, the environment is part of the perception-action system. The organism's interaction with the environment lawfully generates information which specifies the organism-environment relationship [8]. The organism, or in this case VR user, has to perceive what opportunities for action are available to them so that they can act; Gibson described these opportunities as "affordances" [23]. Calibration allows the user to remap their actions to match the affordances of the environment when their own capabilities for action are altered. An alteration of the whole environment, such as a novel experience in an IVE, could also require calibration.

Bingham and Romack investigated the effects of calibration on participants' ability to adapt to displacement prisms which skewed their vision while completing a task requiring reaches across several sessions and days [5]. They found that there was no rate of change in calibration during the blocks but that later trials required less calibration because participants were more accurate at the start of subsequent blocks. This result suggests that the calibration to the displacement prisms in the experiment may have carried over between sessions, even if sessions were completed over an extended period of time (in this instance three days). This would imply that a "carryover" effect of calibration was present. Bingham and Pagano investigated the influence of calibration on correcting underestimation in distance perception caused by an alteration to the participant's viewing capability [7]. They found that feedback from reaching could eliminate the influence of the altered viewing capability by reducing the underestimation of distance.

A study by Day et al. examined how an altered virtual avatar affected action capabilities when a calibration phase was and was not present [13]. They reported that participants were successfully able to calibrate to thee altered avatar when feedback was present. Ebrahimi et al. studied carryover effects of calibration when negative and positive gain was applied to the movement of a virtual stylus [17]. The researchers concluded that VR users can calibrate to visual feedback even if it conflicts with physical movement. Ebrahimi et al. had also examined reaches during closed-loop feedback co-located with the physical location of a tracked stylus [16]. They found that visuo-motor calibration was present since participants improved their reaching accuracy over trials.

Calibration in virtual environments throughout these studies is shown to be effective in improving distance estimation. Feedback can even help bridge the gap between motor capabilities in the real world and virtual world [18]. The missing aspect from these studies is an evaluation of calibration on distance estimation in IVEs over time.

2.3 Long-Term Exposure VR Studies

There are several different kinds of studies related to long-term exposure in VR that have been conducted over several different time frames. A study by Lin et al. examined if participants could calibrate to audio reverberation for distance perception [33]. They found similar patterns of underestimation as found in some of the works discussed above. They also found that participants were able to calibrate to the reverberation and that the calibration persisted. In sessions held one and six months after the first the authors found that the calibration from the first session still impacted participants' distance estimation.

Steinicke and Bruder exposed a participant to an IVE for 24 hours straight (with short breaks allowed during the session) [54]. They reported that the participant experienced varying levels of sickness throughout the course of the experiment based on the SSQ that was not mitigated by time. A similar study by Nordahl et al. exposed two participants to prolonged exposure in HMD's for 12 hours [40]. They found that participants' reported sickness went up and down inconsistently throughout the study. Zielasko also self-reported sensitivity to simulator sickness, but instead of prolonged exposure to VR in one session they discussed their seven year exposure to VR systems [64].

A survey of studies related to simulator sickness by Dużmańska et al. found that duration of sickness and a threshold time for maximum sickness varies greatly between studies [15]. Porter and Robb found evidence that suggested the way consumers talk about simulator sickness and other factors in VR in online forums may change as they gain experience over prolonged periods of exposure [46]. Baileson and Yee studied users in a collaborative virtual environment across 10 weeks in 15, 45 minute sessions [3]. They found that users reported less sickness as the participants gained experience in the virtual environment, but also that participants looked at each other less over time. Porter et al. also found a small effect related to simulator sickness over time through a study comparing immersive and non-immersive virtual environments in 6, 45 minute sessions over 3 weeks [45].

Many longitudinal studies in VR are system usability studies where the researchers study participants' behavior in the context of particular VR systems over time. This work spans studies related to training (e.g. [50,53,62]), therapeutic applications (e.g. [14,20,21,27,58]), education (e.g. [24,57]), and computer-supported collaboration (e.g. [26,37,59]). The primary focus of these studies is usually on specific experiences in VR and not general changes in interactions as they gain experience.

Based on these related works, there is evidence to suggest that VR users can calibrate their near-field depth perception capabilities, that the effect of this calibration can carryover, and that longitudinal studies provide unique insight for VR studies outside of single experimental sessions. Our work examines how calibration over a longer period of time and repeated across several sessions influences depth-perception in open-loop responses. It also explores how long the carryover effects of calibration may last and what influences carryover effects in reaching estimations.

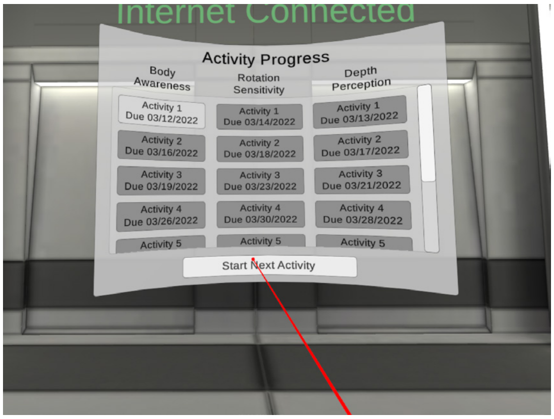
3 Methods

3.1 Participants

We recruited 22 participants via an email sent to undergraduate and graduate students enrolled at Clemson University. Participants were required to have normal or corrected-to-normal vision and to have had less than one hour of prior experience using VR. Due to the longitudinal nature of the study, we experienced a high dropout rate. Six participants completed enough sessions of the experiment to be included in the analyses. Of these six participants, four were male and two were female. Ages ranged from 19 to 25, with a median age of 21 years. Two participants were left-handed. Participants' arm lengths ranged from 67.6 cm to 96.3 cm. Participants received a $50 gift card when they started the study and a $75 gift card at the end of the study if they completed at least half of the experiment modules. This study was approved by the IRB office at Clemson University.

3.2 Procedure

Participants were loaned an Oculus Quest to use for the duration of the 12 week study. Participants were asked to spend at least five hours per week using the headset including time spent completing the study activities and any applications of interest to participants (e.g. games, movies, social applications). In addition to the distance estimation activity, participants were also asked to complete two other experimental activities, one dealing with sensitivity to rotational gains and another with sensitivity to offsets applied to their virtual hands. While the focus of this paper is on the results of the distance estimation activity, we briefly explain the other activities to provide context for everything participants did during the 12 week experiment.



Internet Connected

Activity Progress

Body Awareness	Rotation Sensitivity	Depth Perception
Activity 1 Due 03/12/2022	Activity 1 Due 03/14/2022	Activity 1 Due 03/13/2022
Activity 2 Due 03/16/2022	Activity 2 Due 03/18/2022	Activity 2 Due 03/17/2022
Activity 3 Due 03/19/2022	Activity 3 Due 03/23/2022	Activity 3 Due 03/21/2022
Activity 4 Due 03/26/2022	Activity 4 Due 03/30/2022	Activity 4 Due 03/28/2022
Activity 5	Activity 5	Activity 5

Fig. 2: On starting our application, participants were greeted by the schedule for their activities. They could press "Start Next Activity" to begin the next scheduled activity.

Participants completed each of the three activities before repeating them again. The order in which activities were completed was randomly assigned to each participant, but was held constant across the weeks for a given participant. The number of activities assigned each week varied, with activities completed more frequently at the beginning of the experiment when participants were less familiar with VR. The application we developed enforced this schedule and reminded participants when they were supposed to complete different activities (see Figure 2). Participants completed each activity twice in the first week and once in the second, fourth, sixth, eighth, tenth, and twelfth weeks. Participants were asked to refrain from using the Quest for games and other personal use before completing each of the different experiments once to establish a baseline for each. Participants were also asked to submit a brief journal online each week reflecting on their experience while using the Quest during that week to encourage engagement. Participants were given detailed instructions on how to set up the Quest and how to install our custom application using the Sidequest[1] application.

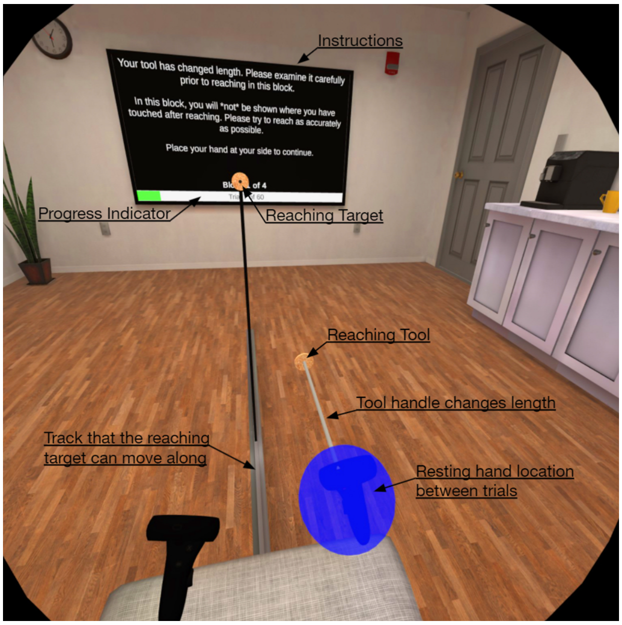
The "hand offset" and "rotational gain" activities are outside of the scope of this paper, however we provide a short description of each activity here. In the hand offset activity, participants were instructed to repeatedly stack blocks on top of each other while seated at a virtual table. Blocks were grasped using a pinching gesture, which was tracked using the Oculus Quest's hand tracking system. While stacking blocks, an offset was sometimes applied to the visible location of participants' hands. Participants were asked to identify when an offset was present each time they stacked a set of blocks. In the rotational gain activity, participants stood in the center of a virtual garage and performed a series of rotations to the left and the right, with a positive or negative gain applied to one of the rotations. After turning to the left and the right, participants were asked to indicate whether they believed a gain had been applied to the first or the second rotation.

Figure 3 shows an annotated example of what participants saw during the distance estimation activity. As participants completed these activities outside of the lab, detailed instructions were recorded and shown to participants prior to beginning each activity. Instructions were also provided during the activity about the specific tasks they should complete. Participants were instructed to sit in a stationary chair in a location where the Quest guardian bounds were not visible. After pressing a button to indicate they had done this, they were re-centered in the virtual environment such that they were aligned with a virtual chair. In the first session of the depth perception experiment, participants were asked to measure their maximum arm reach. To do this, a blue line appeared slightly below eye height and participants were asked to reach as far along the blue line as was possible to do comfortably as shown in the left-most image in Figure 1. A video was played during these

[1] https://sidequestvr.com/

instructions illustrating this process and how to reach. Participants were allowed to bend at the waist while reaching as there was no way to physically constrain this motion outside of a laboratory. Participants were next asked to rest their hand comfortably on their leg and pull the trigger to capture this position. This was used as the resting position all trials started and ended at. The resting position was indicated to participants with a translucent blue sphere that was always visible.

Once participants had completed the introduction, they began the first of four reaching blocks: the carryover block, the pre-calibration block, the calibration block, and the post-calibration block. In each of these blocks, participants were provided with a tool that extended their reach boundary. Targets appeared in front of participants at the same height as the blue line that was used to measure participants maximum arm reach. Participants were asked to reach out and place the tip of their tool in the same location these targets appeared. Participants' vision was blinded at the start of their reach, as determined by when their hand left the resting position measured at the start of the first session. Once participants believed the tool was aligned, they pulled the trigger to record the position of the tool. In the calibration block, participants' vision was restored at this point and they were instructed to move their tool until the tip was correctly aligned with the target and to then return their hand to the resting location. In the other blocks, participants' vision was only restored once they had returned their hand to the resting position. The next trial began once participants returned their hands to the resting position. Participants completed a total of 60 trials in each block. At the end of each session participants were asked to complete the Igroup Presence Questionnaire [48].


Your tool has changed length. Please examine it carefully prior to reaching in this block.

In this block, you will *not* be shown where you have touched after reaching. Please try to reach as accurately as possible.

Place your hand at your side to continue.

Instructions
Bl of 4
Tria of 60
Progress Indicator
Reaching Target
Reaching Tool
Tool handle changes length
Track that the reaching target can move along
Resting hand location between trials

Fig. 3: The depth perception environment. Annotations in the image indicate salient aspects of the simulation.

3.3 Apparatus

Data collected during each activity was automatically saved to the device and uploaded to a remote SQL server. In the event that the Quest was not connected to the Internet when an activity was completed, the locally saved data would be uploaded the next time our application was started and the Quest was connected to the Internet. Our custom application was configured to automatically track participants' application usage on the Quest using Android's Usage Stats API, however this data was not consistently recorded due to several participants failing to allow our application the permissions required to access this portion of the API. In future experiments, additional steps will need to be taken to ensure that all participants properly grant permission to our application to access the Android Usage Stats API.

Each block in each session had a different tool length from the previous block as shown in Table 1. We wanted to avoid making the tool too short (like an unaided reach) or too long (entering medium field distance). We also wanted each tool to be a distinct length so we could compare calibration to specific tools over time vs. general calibration. Finally, we wanted the length of the calibration tool to be alternatively shorter or longer than the pre-calibration tool so participants would not learn that it was always longer or shorter. The pre- and post-calibration blocks used the same tool length so we could compare performance in an open-loop setting before and after closed-loop calibration with a tool of a different length (e.g. [16, 17, 39]). These constraints limited the space of possible tool lengths appropriate for this experiment. We ultimately settled on four tool lengths for the pre- and post-calibration blocks ranging from 20 to 50 cm. The tool in the calibration block was either increased or decreased by 12.5 cm, the maximum amount we judged possible without resulting in tools that were too short or too long. The carryover tool was the same length as the calibration tool from the previous session, allowing us to assess the extent to which closed-loop calibration to a specific tool length persisted across sessions. In the first session, the carryover tool length was arbitrarily chosen since there was no calibration block before it.

The tool lengths used in the first four sessions were repeated in the final four sessions. We would have preferred to use unique tool lengths in each session. However, attempts to do this resulted in different sessions using tools that were nearly indistinguishable from other tools due to the constraint that all reaches should remain within the near-field. As such, we settled on two repetitions of four variations of the tool lengths.

As is common practice in reaching experiments, the target distances were scaled based on percentages of each participants' maximum arm reach (e.g. [7, 39]). Target reaching distances were scaled based on a participant's unique reach boundary, which was determined by their unaided maximum arm reach plus the length of their current tool. The targets were placed between 33% and 84% (in 6% increments) of the participant's reach boundary in the pre-calibration and post-calibration blocks and between 33% and 87% (in 6% increments) in the carryover and calibration block. In all four blocks, the target distances were presented in a random order. We wanted the target distances to differ between the pre- and post- and carryover and calibration blocks which is why their ranges are offset by 3%. We did not want participants to reach to the same distances four times in a row in a given session. Since the target distances were scaled based on percentages of maximum reach, if we only changed the tool length but left the distances the same then the participants would have seen the same actual distances between blocks. By offsetting these ranges, the distances that participants reach to alternate with every other block.

Table 1: Tool lengths used between sessions and blocks. The target distance was scaled to a percentage of each participant's maximum arm reach plus the tool length.

Session	Carryover	Pre-Calibration	Calibration	Post-Calibration
1	15.0 cm	30.0 cm	17.5 cm	30.0 cm
2	17.5 cm	20.0 cm	32.5 cm	20.0 cm
3	32.5 cm	50.0 cm	37.5 cm	50.0 cm
4	37.5 cm	40.0 cm	52.5 cm	40.0 cm
5	52.5 cm	30.0 cm	17.5 cm	30.0 cm
6	17.5 cm	20.0 cm	32.5 cm	20.0 cm
7	32.5 cm	50.0 cm	37.5 cm	50.0 cm
8	37.5 cm	40.0 cm	52.5 cm	40.0 cm

3.4 Analyses

We used linear-mixed models (LMMs) to analyze the results of this experiment. The models were created in R [47] first using the 'buildmer' [60] and 'lme4' [4] R packages. Buildmer automatically tests different possible models based on a set of independent variables and uses the likelihood-ratio test and minimum Bayesian information criterion to select the model that best fits the observed data [52]. The models suggested by buildmer was then considered from a theoretical perspective; in our case, all suggested models were theoretically sound. We then checked for violations of assumptions and made any necessary modifications to the suggested model formula based on the results. Once the final model was specified, we fit it to the data using the lmer command provided by lme4. The lmerTest package [31] was used to estimate p-values using the Satterthwhaite degrees of freedom method [51] for the models generated by lmer. Figures were generated using ggplot2 [61].

Each of the models was checked for normality of residuals, normality of random effects, a linear relationship, homogeneity of variance, and multicollinearity (see the 'check_model' function in the 'performance' R package [34]). These checks led to the alteration of some of the model formulas output by buildmer, primarily due to violations of collinearity. In particular, we removed predictors with variance inflation factors (VIFs) greater than five. These moderate to high levels of collinearity were often caused by interaction effects in the models. We also grand mean centered the continuous variable log(day) in the calibration block model. Grand mean centering of continuous independent variables prior to computing the interaction terms are typically conducted in multiple regression and hierarchical multilevel linear modeling analyses. Centering minimizes or eliminates high correlations between the individual independent variables or predictors and the interaction terms derived from them [29]. If a predictor was removed from a model to correct a violated assumption, the new model's conditional r^2 value was computed and compared against the previous model to assess the change's impact on model fit. In our case, all model adjustments resulted in negligible changes to model fit.

As mentioned previously, only six participants completed enough of the experiment to be included in our analysis. Of those six participants, five completed all eight sessions and one participant completed the first six sessions. The 'check_outliers' function of the 'performance' R package [34], with the outlier threshold set to a z-score of three, was used to identify and exclude 412 out of 11,040 data points. Outliers were excluded on a per-participant basis to account for inter-participant variability in error scores. Outliers were excluded prior to the construction of the LMMs.

3.5 Interpreting Linear Mixed-Effect Models

LMMs are becoming a widely used approach to statistical analyses when working with quantitative data in psychological research [36], but we have yet to see their widespread use in papers related to VR research. Thus, we briefly discuss their advantages and how to interpret their results here.

LMMs are particularly well suited to analyze data involving correlations between conditions or measurements (e.g. when a single participant's behavior is measured multiple times [30]). When compared with repeated-measures ANOVAs, LMMs are more robust to assumption violations and also perform better when analyzing datasets with missing or imbalanced data. Additionally, LMMs are capable of handling data collected on a variable schedule, making them practical for use with human-subject research and its natural variability. Taken together, these features make LMMs especially appropriate when analyzing longitudinal data [11].

LMMs model both *fixed* and *random* effects [36]. Fixed effects model the influence of predictors on the measured quantities, while random effects model the unexplained variability associated with individual differences between participants. To do this, LMMs independently model each participants' response to the data. Random effects can be modeled as *random intercepts*, where the intercept of each participant's individual model is allowed to vary based on the best fit to that participant's data. Random effects can also be modeled as *random slopes*, where the slope of each participant's individual model is also allowed to vary based on the best fit to their particular data. Including random effects in the model improves LMMs' abilities to precisely model the response of fixed effects independent of the innate variability between participants.

The fitted model includes an intercept, which describes the predicted value of the modeled variable when all predictors are set to 0. The model also includes fitted beta values for each fixed effect, which describe how much the modeled variable is predicted to change when the fixed effect unit increases by one. Beta values are highly sensitive to the scale of a given dependent variable, such that dependent variables spanning a small range of values produce smaller beta values and vice versa. This makes it difficult to judge the relative importance of a fixed effect from the size of the beta values alone. Therefore, we also provide standardized beta values (abbreviated std. beta in our tables), which are scaled based on the individual distribution of each dependent variable. This makes it easier to determine the relative contribution each fixed effect has on the independent variable. Instead of determining the change in the dependent variable when the independent variable increases by one unit (holding all other variables constant), the standardized beta values tell us the standard deviation change in the dependent variable when the independent variable changes by one standard deviation [9].

We also provide both marginal and conditional r^2 values. Marginal r^2 values (abbreviated m. r^2 in our tables) report the amount of variance explained by the fixed effects alone and conditional r^2 values report the amount of variance explained by both the fixed and random effects [38]. Thus, the conditional r^2 values tell us the explanatory power for the entire model.

4 Results

4.1 Dependent and Independent Variables

Our analyses include several independent variables and one primary dependent variable, *signed-error*. Signed-error is the difference between a participant's judgment and the actual distance for a given trial. A negative-signed error indicates distance underestimation, and a positive-signed error indicates distance overestimation. In addition to our analysis of signed-error, we also include a graph visualizing the original measures, target distance and reach estimate, in Figure 4.

Our primary independent variable is time, given that it is the main focus of this work. We represent time in our analyses as the natural log of days since the start of the study (represented as *log(days)*). Work by Mazur and Hastie has shown that changes in human performance over time, including performance related to perception and motor skills, are rarely linear and can be better modeled as a negative exponential or logarithmic relationship [35]. As such, a log transform was applied to days in order to more appropriately model the relationships in our data using LMMs.

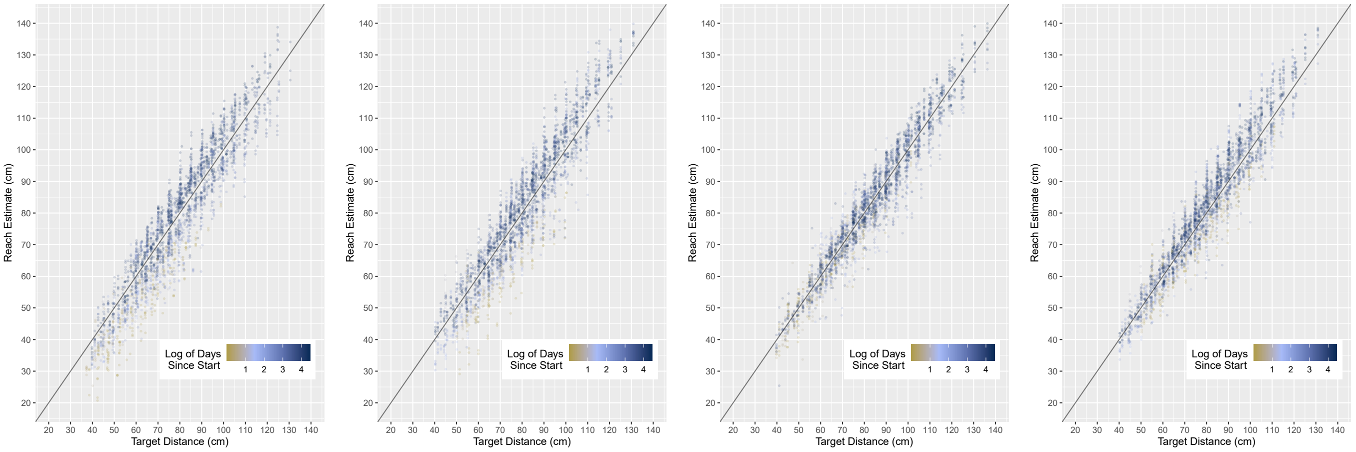
We considered two other secondary independent variables when constructing our models. These variables were included primarily to control for unwanted effects that may make it more difficult to determine whether a relationship between signed-error, log(days), and block was present in our data. The variables included the 1) trial number within a block and 2) the target distance as a percentage of a participant's current reach boundary.

4.2 Hypotheses

H1 was tested by examining participants' performance in the pre-calibration block. Participants used a tool with an unfamiliar length during this block and had not yet received feedback about their performance. As such, a reduction in underestimation in this block would indicate that the effect of depth compression was decreasing. We expected a similar change to be visible in all four blocks if depth compression decreased. However, this can be most directly tested using the pre-calibration block for the reasons explained above.

H2 was tested by comparing the results of the pre-calibration block with the carryover block. If the effects of calibrating to a tool of a specific length persisted, rather than the calibration affecting the overall system of depth perception, we expected to see a difference in performance between the carryover block (which used a familiar tool) and the pre-calibration block (which used a tool of a different length). In particular, we expected to see that performance in the carryover block was better than in the pre-calibration block (assuming participants' performance had actually improved in the previous session after calibration).

H3 was tested by evaluating participants' performance in the calibration block where they received feedback about their reaches. If



(a) Carryover block (b) Pre-Calibration block (c) Calibration block (d) Post-Calibration block

Fig. 4: This shows the relationship between the target distances and participants' reach distances for all four blocks across all sessions. The graphs show that as time went on, indicated by log(days), participants appear to underestimate distances less. The line in the center of the graph represents a slope of one, so if participants had perfect accuracy then the point would fall on the line. We also note that the data in the calibration and post-calibration blocks are more tightly grouped around the center line indicating a higher level of accuracy.

participants' baseline performance improved with time, we would expect to see a main effect of log(days) in the model, as this would correspond to improved performance even in the first trial. Alternatively, if the rate of calibration improves with time, we would expect to see a main effect of trial in the model (a higher rate of calibration equates to more improvement within a given trial).

Finally, H4 was tested by comparing participants' performance in the pre- and post-calibration blocks. Assuming participants calibrated to the specific tool length used in the calibration block, this calibration should continue to influence reaches in the post-calibration block where participants do not receive feedback. If the effect of calibration diminishes over time, we would expect to observe an interaction effect between block and log(days), where block (and thus the calibration that occurs within a block) would have less of an impact on error over time.

4.3 Signed Error in Carryover vs. Pre-calibration Blocks

To evaluate H1 and H2, we fit a model using the data from the carryover and pre-calibration blocks. The fitted linear model included log(day), block, trial, and target distance as fixed effects and participant ID as a random effect. Buildmer originally included the between log(day) and target distance but this interaction had a very large VIF (35) so it was removed from the model. The removal of this interaction had a negligible impact on the fit of the model. The model's total explanatory power (conditional r^2) is 0.43 and the explanatory power of the fixed effects alone (marginal r^2) is 0.27. The parameter values for the model are shown in Table 2.

Based on the model's intercept (when all predictors were zero), participants underestimated distances by 9.81 cm. Underestimation decreased by 3.21 cm for each unit of log(day). It also decreased by 1.26 cm when moving from the carryover to pre-calibration block, decreased by 0.03 cm for each trial completed, and increased as targets were placed farther away from the participant. All model parameters had significant effects on the predicted signed-error, with log(Day) having the largest impact as indicated by both the standardized beta and marginal r^2 values.

4.4 Signed Error in the Calibration Block

To evaluate H3, we fit a model to the data from the calibration block to understand how performance changed over time when participants were receiving feedback about their reaches. The fitted linear model included target distance, log(day), trial, and the interaction between log(day) and trial as fixed effects and participant ID as a random effect. The model's total explanatory power (conditional r^2) is 0.24 and the explanatory power of the fixed effects alone (marginal r^2) is 0.09. The parameter values for the model are shown in Table 3.

Unlike previous models that did not require data to be grand-mean centered, the intercept for this model does not accurately reflect the amount of error present when all predictors were zero and is thus

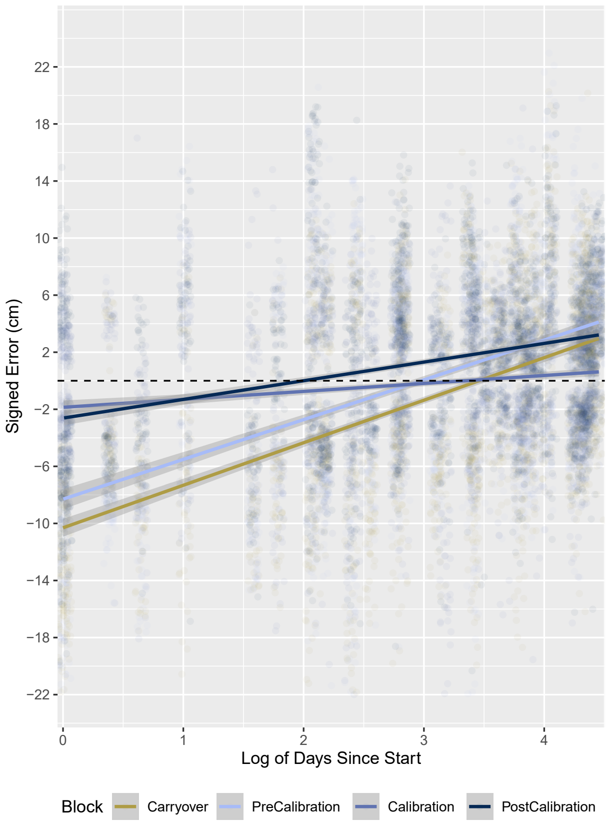


Fig. 5: The relationship between log(days) and signed-error in each of the four blocks is shown here. Depth compression, indicated by a negative signed-error, decreased with time.

Table 2: Carryover vs. Pre-Calibration Parameters

Fixed Effect	beta [95% CI]	Std. beta [95% CI]	t(5225)	p	m. r^2
(Intercept)	-9.81 [-12.61, -7.01]	–	-6.88	<0.001	–
log(Day)	3.21 [3.08, 3.35]	0.52 [0.50, 0.55]	47.49	<0.001	0.2552
Block	1.26 [0.93, 1.59]	0.16 [0.12, 0.20]	7.48	<0.001	0.0062
Trial	0.03 [0.02, 0.04]	0.07 [0.05, 0.09]	6.33	<0.001	0.0042
Target Distance	-3.39 [-4.70, -2.08]	-0.06 [-0.08, -0.03]	-5.08	<0.001	0.0028

more difficult to interpret. However, grand-mean centering does not affect the interpretation of the other effects in this model. The model suggests that participants would underestimate the distance to their reach boundary by 11.26 cm. Underestimation decreased by 1.15 cm per unit of log(day). Trial did not have a significant effect on signed-error in this model, indicating that trial number had little to no effect on signed-error. Finally, the interaction effect of trial on log(day) was significant. This interaction effect indicates that the main effect of log(day) diminished by 0.0195 cm with each trial. By the end of the 60 trials, the main effect of log(day) was reduced by 1.17 cm. As the main effect of log(day) was 1.15 cm, this suggests that log(day) no longer impacted signed-error by the end of the calibration block. Except for trial, all model parameters had a significant effect on signed-error. Target distance had the largest impact on signed-error in this model as judged by both the standardized beta and marginal r^2 values.

Table 3: Calibration Parameters

Fixed Effect	beta [95% CI]	Std. beta [95% CI]	t(2668)	p	m. r^2
(Intercept)	7.87 [5.81, 9.92]	–	7.516	<0.001	–
Target Distance	-11.62 [-13.07, -10.17]	-0.27 [-0.30, -0.24]	-15.71	<0.001	0.0703
log(Day)	1.15 [0.88, 1.43]	0.15 [0.11, 0.18]	8.35	<0.001	0.0206
Trial	0.00280 [-0.00779, 0.01]	0.0088 [-0.02, 0.04]	0.52	0.604	<0.001
log(Day):Trial	-0.0195 [-0.03, -0.01]	-0.08 [-0.12, -0.05]	-4.81	<0.001	0.0065

4.5 Signed Error in Pre-calibration vs. Post-calibration

To evaluate H4, we fit a model using the data from the pre-calibration and post-calibration blocks. The fitted linear model included log(day), block, trial, and target distance as fixed effects, and participant ID as a random effect. An interaction effect between log(day) and block was also included in the final model. The model's total explanatory power (conditional r^2) is 0.36 and the explanatory power of the fixed effects alone (marginal r^2) is 0.19. The parameter values for the model are shown in Table 4.

Based on the model's intercept (when all predictors were zero), participants underestimated distances by 8.16 cm. Underestimation decreased by 3.17 cm for each unit of log(day). Underestimation also decreased by 6.87 cm when moving from the pre-calibration to post-calibration block, decreased by 0.03 cm for each trial completed, and increased as target distance increased. The interaction effect between log(day) and block also increased distance underestimation, such that the difference between pre-calibration and post-calibration blocks decreased by 1.80 cm for each unit of log(day). There was a main effect of block, indicating that depth compression decreased in the post-calibration block compared to the pre-calibration block. Therefore, the interaction effect indicates that the magnitude of the main effect of block diminished with time. This can likely be attributed to the increase in baseline performance during the pre-calibration phase which limited the improvement possible for the post-calibration phase. As before, all model parameters had a significant effect on signed-error, and log(day) remained the parameter with the largest impact on signed-error, as judged by both the standardized beta and marginal r^2 values.

5 Discussion

Our analyses of the data provided evidence to support H1. A main effect of log(day) was observed in the model of the carryover and pre-calibration data where depth compression decreased by 3.21 cm for each unit of log(Day). An interaction effect was not observed between log(day) and block, which indicates that this improvement in performance was similar between blocks. We are primarily concerned with the improvement visible in the pre-calibration block, as participants used a tool of an unfamiliar length in this block. Given that participants had no experience using this tool prior to a given session, we can attribute the observed improvement to an overall improvement in distance judgments rather than to improvements with a specific tool.

Table 4: Pre-Calibration vs. Post-Calibration Parameters

Fixed Effect	beta [95% CI]	Std. beta [95% CI]	t(5315)	p	m. r^2
(Intercept)	-8.16 [-10.73, -5.58]	–	-6.21	<0.001	–
log(Day)	3.17 [2.99, 3.35]	0.59 [0.56, 0.62]	34.52	<0.001	<0.001
Block	6.87 [6.09, 7.66]	0.21 [0.17, 0.26]	17.19	<0.001	0.0356
Trial	0.03 [0.02, 0.04]	0.07 [0.04, 0.09]	5.81	<0.001	0.0040
Target Distance	-3.54 [-4.79, -2.28]	-0.06 [-0.08, -0.04]	-5.52	<0.001	0.0038
log(Day):Block	-1.80 [-2.04, -1.56]	-0.33 [-0.38, -0.29]	-14.70	<0.001	0.0261

Regarding H2, we did not observe any evidence that calibration to a tool of a given length persisted across sessions. In particular, we did not find an interaction effect between log(day) and block in the model of the carryover and pre-calibration data. An interaction effect would have suggested that participants used a new tool and a tool they had previously calibrated to differently. Instead, our findings suggest that participants performed worse with tools they had previously calibrated to. Specifically, depth compression was more evident in the carryover block (when they used the tool they had calibrated to in the previous session) than in the pre-calibration block (when they used a tool of an unfamiliar length).

It is surprising that participants would perform worse with the carryover tool than the pre-calibration tool. In the pre- vs. post-calibration model we observed a reduction in depth compression between the blocks, indicating that performance improved after calibration. If calibrating to a specific tool carried over between sessions, we would expect to see improved performance in the carryover block compared to the pre-calibration block. If calibrating to a specific tool did not carryover between sessions, we would expect to see no difference in performance between the blocks. Instead, we found the opposite of this. On further consideration, this effect may potentially be attributed to the small, but significant effect of trial in the carryover vs. pre-calibration model. While participants did not receive feedback in the carryover block, their performance improved by a small amount for each trial. Distance estimation research typically assumes that performance only changes in the presence of explicit feedback, but the ecological approach to perception posits that people constantly calibrate to invariants in the environment [8]. This means that performance can change minutely even when feedback is not explicitly present. It may be that the observed difference between the carryover and the pre-calibration blocks is due to this learning effect. This is made more plausible by an examination of the estimated model parameters. Each trial was estimated to reduce overestimation by 0.03 cm. Given that there were 60 trials in the carryover block, we would expect to see a reduction in overestimation of 1.8 cm by the end of the carryover block. This corresponds to the observed effect of block which estimated that underestimation was reduced by 1.26 cm when moving from the carryover to the pre-calibration block.

Regarding H3, the results in the calibration block model partially supported the hypothesis that participants would reach full calibration faster over time with feedback. The main effect of log(day) indicates that their baseline performance improved over time in the calibration block (H3a). Conversely, the lack of significant effect for trial suggests that the rate of calibration did not change over time (H3b). Trial moderated the effect of log(day), reducing its influence to basically zero at the end of the 60 trials in the calibration block. This is likely a reflection of how participants' reach estimates were highly accurate by the end of the calibration block, regardless of log(day). Their initial performance (i.e. the first trial) improved with time due to the main effect of log(day). However, log(day) ceased to matter in the final trials as participants had calibrated their performance based on the feedback they received.

Regarding H4, we observed a carryover effect of calibration indicated by the main effect of block in the model of pre- and post-calibration data. Both the pre- and post-calibration blocks employed tools of the same length, so the difference between these blocks can be attributed to the carryover effects from the calibration block where feedback was provided while participants reached with a different length tool. We also observed an interaction effect between log(days) and block. This interaction effect indicated that the effect of calibration diminished with time, such that reaches in the pre- and post-calibration block became more similar with time. This interaction effect can likely be attributed to the improved baseline performance. Specifically, as depth compression in general decreased, there was less potential (and less need) to calibrate in the calibration block.

While depth compression decreased in the earlier sessions, we did not find that signed error approached zero by the end of the experiment. Surprisingly, participants showed signs of *depth expansion*, or overestimation of distances by the end of the experiment. We fit a curvilinear function to check whether this was an artifact of fitting the data with a linear model. The curvilinear function also suggested that participants were overestimating distance by the end of the experiment. This observation was unexpected, so we suggest that future research explore this further.

The most unexpected finding from this study was the shift from underestimation to overestimation of distances that occurred by the end of the study. This was an unexpected result given the well-documented phenomenon of underestimation and the ability for calibration to minimize error. This is especially true given that our data was modeled using a linear fit. However, this may misrepresent the trend in the data if it actually follows a non-linear trajectory given the small sample size. That said, the application of a log transform to our time variable makes the use of a linear model less problematic, and we observed that applying a curvilinear fit to the data shown in Figure 5 also suggested that participants were overestimating distances by the end of the experiment. The linear models also demonstrated reasonably good measures of fit, with conditional r^2 values ranging between 0.24 and 0.43. This overestimation requires further investigation and would benefit from confirmation via additional studies.

5.1 Limitations

Despite recruiting 22 participants, only six completed enough sessions to be used in our analyses. As we expected significant dropout going into the study, we intentionally set the number of trials and sessions high so as to gather substantial amounts of data from the participants who completed the study. Further, spatial perception in individuals with no perceptual or motor deficits is not expected to differ based on race, gender, etc. in VR with strong depth cues (e.g. [44]) like we might expect of cognitive effects (e.g. [56]). Even with a relatively small participant pool, completing a relatively large number of trials provides sufficient power for the results to generalize to a larger population of normal sighted (or corrected to normal) and motor capable individuals. LMMs are also well suited for analyzing longitudinal data with few participants as they model the individual variability of each participant separately, which allows to overall trends to be better identified independent of the variability that may be present in small samples. That said, our final sample size was small and was limited to young, motor capable individuals. Future work that incorporated more diverse participants could help strengthen the evidence of the phenomena we observed.

A second limitation of this work was that the order of tool lengths were not randomized between participants. This choice was made to simplify the already complex analysis but given the limited sample size counter-balancing the lengths could have a significant impact on the models.

6 Conclusion

In this paper, we discussed the results of the first study to date that considered how depth perception in VR changes with time and experience. We found that repeated calibration reduced the effect of depth compression over time, potentially even to the point of depth expansion, where users began to overestimate distances in VR. Our results suggested that this occurred broadly, rather than being restricted to tools of specific lengths that have been used previously. Our findings also suggested that the effects of calibration that occurred within a given session diminished with time, likely due to reductions in depth compression prior to calibration. It is important to note that calibration can occur in any virtual environment that provides feedback to users about the accuracy of their reaches. In practice, this is essentially all virtual environments. A specific calibration block was employed in our experiment to examine the effects of feedback on reaching behavior, but such a block is not necessarily required in other activities where users do not engage in blind reaching. Therefore, we would expect that depth compression will diminish with time as users gain experience with VR, possibly generally across all applications they encounter and almost certainly when they repeatedly use the same application as in this experiment. While this work specifically focused on distance estimation and depth compression, its findings have implications for other perceptual phenomena affected by VR. Future work is needed to understand whether similar effects can be observed in users' abilities to detect perceptual distortions, both unintentional and otherwise.

Acknowledgments

The authors would like to thank the participants of our longitudinal study for their time and effort. This work was supported in part by the US National Science Foundation (CISE HCC) under Grants #1717937 and #2007435.

References

[1] B. M. Altenhoff, P. E. Napieralski, L. O. Long, J. W. Bertrand, C. C. Pagano, S. V. Babu, and T. A. Davis. Effects of calibration to visual and haptic feedback on near-field depth perception in an immersive virtual environment. In *Proceedings of the ACM symposium on applied perception*, pp. 71–78, 2012.

[2] C. Armbrüster, M. Wolter, T. Kuhlen, W. Spijkers, and B. Fimm. Depth perception in virtual reality: distance estimations in peri-and extrapersonal space. *Cyberpsychology & Behavior*, 11(1):9–15, 2008.

[3] J. N. Bailenson and N. Yee. A longitudinal study of task performance, head movements, subjective report, simulator sickness, and transformed social interaction in collaborative virtual environments. *Presence: Teleoperators and Virtual Environments*, 15(6):699–716, 2006.

[4] D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01

[5] G. Bingham and J. L. Romack. The rate of adaptation to displacement prisms remains constant despite acquisition of rapid calibration. *Journal of Experimental Psychology: Human Perception and Performance*, 25(5):1331, 1999.

[6] G. P. Bingham, A. Bradley, M. Bailey, and R. Vinner. Accommodation, occlusion, and disparity matching are used to guide reaching: a comparison of actual versus virtual environments. *Journal of experimental psychology: human perception and performance*, 27(6):1314, 2001.

[7] G. P. Bingham and C. C. Pagano. The necessity of a perception–action approach to definite distance perception: Monocular distance perception to guide reaching. *Journal of Experimental Psychology: Human Perception and Performance*, 24(1):145, 1998.

[8] J. J. Blau and J. B. Wagman. Introduction to ecological psychology: A lawful approach to perceiving, acting, and cognizing.

[9] J. Bring. How to standardize regression coefficients. *The American Statistician*, 48(3):209–213, 1994.

[10] L. Buck, R. Paris, and B. Bodenheimer. Distance compression in the htc vive pro: A quick revisitation of resolution. *Frontiers in Virtual Reality*, p. 157.

[11] A. Cnaan, N. M. Laird, and P. Slasor. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in medicine*, 16(20):2349–2380, 1997.

[12] S. H. Creem-Regehr, P. Willemsen, A. A. Gooch, and W. B. Thompson. The influence of restricted viewing conditions on egocentric distance perception: Implications for real and virtual indoor environments. *Perception*, 34(2):191–204, 2005.

[13] B. Day, E. Ebrahimi, L. S. Hartman, C. C. Pagano, A. C. Robb, and S. V. Babu. Examining the effects of altered avatars on perception-action in virtual reality. *Journal of Experimental Psychology: Applied*, 25(1):1, 2019.

[14] F. A. dos Santos Mendes, J. E. Pompeu, A. M. Lobo, K. G. da Silva, T. de Paula Oliveira, A. P. Zomignani, and M. E. P. Piemonte. Motor learning, retention and transfer after virtual-reality-based training in parkinson's disease–effect of motor and cognitive demands of games: a longitudinal, controlled clinical study. *Physiotherapy*, 98(3):217–223, 2012.

[15] N. Dużmańska, P. Strojny, and A. Strojny. Can simulator sickness be avoided? a review on temporal aspects of simulator sickness. *Frontiers in psychology*, 9:2132, 2018.

[16] E. Ebrahimi, B. Altenhoff, L. Hartman, J. A. Jones, S. V. Babu, C. C. Pagano, and T. A. Davis. Effects of visual and proprioceptive information in visuo-motor calibration during a closed-loop physical reach task in immersive virtual environments. In *Proceedings of the ACM Symposium on Applied Perception*, pp. 103–110, 2014.

[17] E. Ebrahimi, B. M. Altenhoff, C. C. Pagano, and S. V. Babu. Carryover effects of calibration to visual and proprioceptive information on near field distance judgments in 3d user interaction. In *2015 IEEE Symposium on 3D User Interfaces (3DUI)*, pp. 97–104. IEEE, 2015.

[18] E. Ebrahimi, S. V. Babu, C. C. Pagano, and S. Jörg. An empirical evaluation of visuo-haptic feedback on physical reaching behaviors during 3d interaction in real and immersive virtual environments. *ACM Transactions on Applied Perception (TAP)*, 13(4):1–21, 2016.

[19] E. Ebrahimi, A. Robb, L. S. Hartman, C. C. Pagano, and S. V. Babu. Effects of anthropomorphic fidelity of self-avatars on reach boundary estimation in immersive virtual environments. In *Proceedings of the 15th ACM Symposium on Applied Perception*, pp. 1–8, 2018.

[20] A. Faria, J. Couras, M. Cameirão, T. Paulino, G. Costa, and S. Bermúdez i Badia. Impact of combined cognitive and motor rehabilitation in a virtual reality task: an on-going longitudinal study in the chronic phase of stroke. *Proceedings of the 11th ICDVRAT*, 2016.

[21] D. Freeman, C. Thompson, N. Vorontsova, G. Dunn, L.-A. Carter, P. Garety, E. Kuipers, M. Slater, A. Antley, E. Glucksman, et al. Paranoia and post-traumatic stress disorder in the months after a physical assault: a longitudinal study examining shared and differential predictors. *Psychological medicine*, 43(12):2673–2684, 2013.

[22] H. C. Gagnon, T. Rohovit, H. Finney, Y. Zhao, J. M. Franchak, J. K. Stefanucci, B. Bodenheimer, and S. H. Creem-Regehr. The effect of feedback on estimates of reaching ability in virtual reality. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 798–806. IEEE, 2021.

[23] J. J. Gibson. The theory of affordances. *Hilldale, USA*, 1(2):67–82, 1977.

[24] T. Hill and H. du Preez. A longitudinal study of students' perceptions of immersive virtual reality teaching interventions. In *2021 7th International Conference of the Immersive Learning Research Network (iLRN)*, pp. 1–7. IEEE, 2021.

[25] J. Kelly, T. Doty, M. Ambourn, and L. Cherep. Distance perception in the oculus quest and oculus quest 2. 2022.

[26] N. Khojasteh and A. S. Won. Working together on diverse tasks: A longitudinal study on individual workload, presence and emotional recognition in collaborative virtual environments. *Frontiers in Virtual Reality*, 2:53, 2021.

[27] H.-J. Kim, S. Lee, D. Jung, J.-W. Hur, H.-J. Lee, S. Lee, G. J. Kim, C.-Y. Cho, S. Choi, S.-M. Lee, et al. Effectiveness of a participatory and interactive virtual reality intervention in patients with social anxiety disorder: longitudinal questionnaire study. *Journal of medical Internet research*, 22(10):e23024, 2020.

[28] J. M. Knapp and J. M. Loomis. Limited field of view of head-mounted displays is not the cause of distance underestimation in virtual environments. *Presence: Teleoperators & Virtual Environments*, 13(5):572–577, 2004.

[29] I. G. Kreft and J. De Leeuw. *Introducing multilevel modeling*. Sage, 1998.

[30] C. Krueger and L. Tian. A comparison of the general linear mixed model and repeated measures anova using a dataset with multiple missing data points. *Biological research for nursing*, 6(2):151–157, 2004.

[31] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1–26, 2017. doi: 10.18637/jss.v082.i13

[32] M. Leyrer, S. A. Linkenauger, H. H. Bülthoff, U. Kloos, and B. Mohler. The influence of eye height and avatars on egocentric distance estimates in immersive virtual environments. In *Proceedings of the ACM SIGGRAPH symposium on applied perception in graphics and visualization*, pp. 67–74, 2011.

[33] W.-Y. Lin, Y.-C. Wang, D.-R. Wu, R. Venkatakrishnan, R. Venkatakrishnan, E. Ebrahimi, C. Pagano, S. V. Babu, and W.-C. Lin. Empirical evaluation of calibration and long-term carryover effects of reverberation on egocentric auditory depth perception in vr. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 232–240. IEEE, 2022.

[34] D. Lüdecke, M. S. Ben-Shachar, I. Patil, P. Waggoner, and D. Makowski. performance: An r package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, 6(60), 2021.

[35] J. E. Mazur and R. Hastie. Learning as accumulation: a reexamination of the learning curve. *Psychological Bulletin*, 85(6):1256, 1978.

[36] L. Meteyard and R. A. Davies. Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, 112:104092, 2020.

[37] F. Moustafa and A. Steed. A longitudinal study of small group interaction in social virtual reality. In *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology*, pp. 1–10, 2018.

[38] S. Nakagawa, P. C. Johnson, and H. Schielzeth. The coefficient of determination r 2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, 14(134):20170213, 2017.

[39] P. E. Napieralski, B. M. Altenhoff, J. W. Bertrand, L. O. Long, S. V. Babu, C. C. Pagano, J. Kern, and T. A. Davis. Near-field distance perception in real and virtual environments using both verbal and action responses. *ACM Transactions on Applied Perception (TAP)*, 8(3):1–19, 2011.

[40] R. Nordahl, N. C. Nilsson, A. Adjorlu, E. Magalhaes, S. Willemsen, N. S. Andersson, J. Wang, and S. Serafin. 12 hours in virtual reality: Two cases of long-term exposure to consumer-grade virtual reality. In *Proc. of IEEE VR Workshop on Immersive Sickness Prevention*, 2019.

[41] L. Phillips, V. Interrante, M. Kaeding, B. Ries, and L. Anderson. Correlations between physiological response, gait, personality, and presence in immersive virtual environments. *Presence*, 21(2):119–141, 2012.

[42] L. Phillips, B. Ries, V. Interrante, M. Kaeding, and L. Anderson. Distance perception in npr immersive virtual environments, revisited. In *Proceedings of the 6th Symposium on Applied Perception in Graphics and Visualization*, pp. 11–14, 2009.

[43] L. Phillips, B. Ries, M. Kaeding, and V. Interrante. Avatar self-embodiment enhances distance perception accuracy in non-photorealistic immersive virtual environments. In *2010 IEEE virtual reality conference (VR)*, pp. 115–1148. IEEE, 2010.

[44] J. Ping, D. Weng, Y. Liu, and Y. Wang. Depth perception in shuffleboard: Depth cues effect on depth perception in virtual and augmented reality system. *Journal of the Society for Information Display*, 28(2):164–176, 2020.

[45] J. Porter III, M. Boyer, and A. Robb. Guidelines on successfully porting non-immersive games to virtual reality: a case study in minecraft. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*, pp. 405–415, 2018.

[46] J. Porter III and A. Robb. An analysis of longitudinal trends in consumer thoughts on presence and simulator sickness in vr games. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, pp. 277–285, 2019.

[47] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[48] H. Regenbrecht and T. Schubert. Real and illusory interactions enhance presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, 11(4):425–434, 2002.

[49] R. S. Renner, B. M. Velichkovsky, and J. R. Helmert. The perception of egocentric distances in virtual environments-a review. *ACM Computing Surveys (CSUR)*, 46(2):1–40, 2013.

[50] A. Ricca, A. Chellali, and S. Otrnane. The influence of hand visualization in tool-based motor-skills training, a longitudinal study. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 103–112. IEEE, 2021.

[51] F. E. Satterthwaite. An approximate distribution of estimates of variance components. *Biometrics bulletin*, 2(6):110–114, 1946.

[52] G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, pp. 461–464, 1978.

[53] S. J. Smith, S. Farra, D. L. Ulrich, E. Hodgson, S. Nicely, and W. Matcham. Learning and retention using virtual reality in a decontamination simulation. *Nursing education perspectives*, 37(4):210–214, 2016.

[54] F. Steinicke and G. Bruder. A self-experimentation report about long-term use of fully-immersive technology. In *Proceedings of the 2nd ACM symposium on Spatial user interaction*, pp. 66–69, 2014.

[55] F. Steinicke, G. Bruder, and S. Kuhl. Realistic perspective projections for virtual objects and environments. *ACM Transactions on Graphics (TOG)*, 30(5):1–10, 2011.

[56] E. A. Strand. Uncovering the role of gender stereotypes in speech perception. *Journal of language and social psychology*, 18(1):86–100, 1999.

[57] T. M. Takala, L. Malmi, R. Pugliese, and T. Takala. Empowering students to create better virtual reality applications: A longitudinal study of a vr capstone course. *Informatics in Education*, 15(2):287–317, 2016.

[58] I. Tarnanas, W. Schlee, M. Tsolaki, R. Müri, U. Mosimann, and T. Nef. Ecological validity of virtual reality daily living activities screening for early dementia: longitudinal study. *JMIR serious games*, 1(1):e2778, 2013.

[59] V. Venkatesh and P. Johnson. Telecommuting technology implementations: a within-and between-subjects longitudinal field study. *Personnel psychology*, 55(3):661–687, 2002.

[60] C. C. Voeten. Using 'buildmer'to automatically find & compare maximal (mixed) models, 2020.

[61] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009.

[62] A. Winkler-Schwartz, K. Bajunaid, M. A. Mullah, I. Marwa, F. E. Alotaibi, J. Fares, M. Baggiani, H. Azarnoush, G. Al Zharni, S. Christie, et al. Bimanual psychomotor performance in neurosurgical resident applicants assessed using neurotouch, a virtual reality simulator. *Journal of surgical education*, 73(6):942–953, 2016.

[63] B. G. Witmer and P. B. Kline. Judging perceived and traversed distance in virtual environments. *Presence*, 7(2):144–167, 1998. doi: 10.1162/105474698565640

[64] D. Zielasko. Subject 001-a detailed self-report of virtual reality induced sickness. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 165–168. IEEE, 2021.