Matching and Reaching Depth Judgments with Real and Augmented Reality Targets

J. Edward Swan II, *Member, IEEE*, Gurjot Singh, *Member, IEEE*, and Stephen R. Ellis

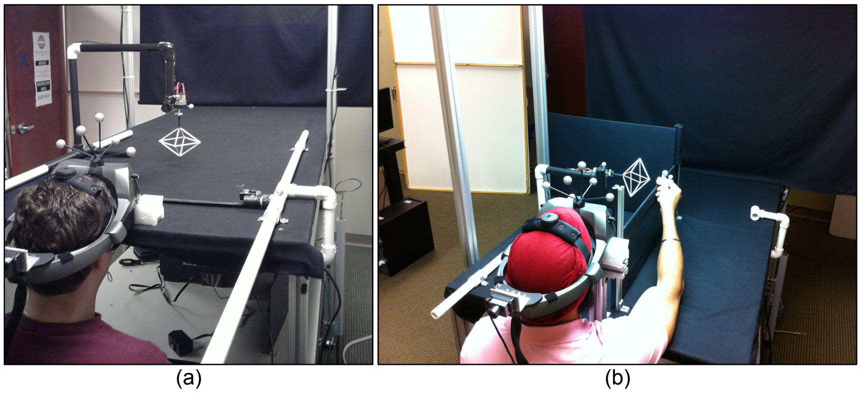

Fig. 1. We conducted two experiments that carefully measured depth judgments, using both matching and reaching tasks, with both real targets—as shown—as well as augmented reality (AR) targets. Experiment I (a) involved matching and reaching by manipulating a slider mounted underneath a table, while Experiment II (b) involved matching and reaching with the tip of the finger. These experiments were motivated by AR applications in medicine, manufacturing, and maintenance, where virtual objects may need depth placement accuracy of 1 mm or less.

Abstract—Many compelling augmented reality (AR) applications require users to correctly perceive the location of virtual objects, some with accuracies as tight as 1 mm. However, measuring the perceived depth of AR objects at these accuracies has not yet been demonstrated. In this paper, we address this challenge by employing two different depth judgment methods, *perceptual matching* and *blind reaching*, in a series of three experiments, where observers judged the depth of real and AR target objects presented at reaching distances. Our experiments found that observers can accurately match the distance of a real target, but when viewing an AR target through collimating optics, their matches systematically overestimate the distance by 0.5 to 4.0 cm. However, these results can be explained by a model where the collimation causes the eyes' vergence angle to rotate outward by a constant angular amount. These findings give error bounds for using collimating AR displays at reaching distances, and suggest that for these applications, AR displays need to provide an adjustable focus. Our experiments further found that observers initially reach ∼4 cm too short, but reaching accuracy improves with both consistent proprioception and corrective visual feedback, and eventually becomes nearly as accurate as matching.

Index Terms—Depth judgment, perceptual matching, blind reaching, accommodation, vergence, augmented reality.

1 Introduction

Many compelling applications of augmented reality (AR), such as image-guided surgery (e.g. Kersten-Oertel, Jannin, and Collins [13]), manufacturing (e.g. Curtis, Mizell, Gruenbaum, and Janin [7]), and maintenance (e.g. Henderson and Feiner [12]), to name a few, require interacting with real and virtual objects at reaching distances. Success depends, among various factors, on how accurately observers can judge the distance to virtual objects, relative to other real objects in the scene. In particular, for AR to be useful for image-guided surgery of the brain, surgeons must be able to match the distance of a real object to a virtual marker within a tolerance of 1 mm or less (Edwards, King, Maurer Jr., de Cunha, Hawkes, Hill, Gaston, Fenlon, Jusczyzck, Strong, Chandler, and Gleeson [9]). This quantitative goal has motived the work reported here: we are investigating methods for accurately measuring the judged depth of AR objects at reaching distances. Our ultimate goal is to use these measurement techniques to engineer methods for positioning AR objects in depth within accuracy and precision requirements, with a particular focus on AR medical applications such as image-guided surgery.

Depth Judgments: In the work reported here, we have studied depth judgments of real and AR targets using two different depth judgment tasks: *perceptual matching*, and *blind reaching*. In the perceptual matching task, an observer indicates a target's distance by pointing at the target with a pointing object, which could be the observer's hand. In the blind reaching task the action is similar, except that the observer cannot see the pointing object. In the general study of depth perception and manual reaching, there is a long history, spanning many decades, of using both methods. In this section we briefly survey this history, and motivate our decision to employ both methods in the current work.

It is generally believed that the human perceptual system is sensitive to various *depth cues* in the environment; for the reaching distances of interest here, Cutting and Vishton [8] list occlusion, binocular disparity, motion perspective, relative size, accommodation, vergence, and relative density as the most important. From these cues, Anderson [2], in his general theory of Functional Measurement, describes distance perception as a three-step process: (1) a *psychophysical transform*, which transforms information from depth cues into distance signals, (2) an *integration process*, which combines all the distance signals

- *J. Edward Swan II is with Mississippi State University. E-mail: swan@acm.org.*
- *Gurjot Singh is with Virginia Tech. E-mail: gurjot@acm.org.*
- *Stephen R. Ellis is with NASA Ames Research Center. E-mail: ellisstephenr3@gmail.com.*

into a single final distance signal, and (3) a *psychomotor transform*, which converts the final distance signal into an action, such as moving the arm. Note that while the first two processes are purely perceptual, the third process involves motor action. This interplay of perception and motor action motivates the widely-used category of depth judgment methods called *perception-action techniques*. These techniques involve performing some motor action task to indicate perceived distance (Bingham, Bradley, Bailey, and Vinner [3]), either in the presence or absence of visual feedback. When visual feedback is present, the task is a visually *closed-loop* task, while when visual feedback is absent, the task is an visually *open-loop* task. The terms *closed-loop* and *open-loop* are from the field of control theory (Levine [14]), and they categorize the behavior of a dynamic system under the effect of feedback. In a closed-loop system, input to the system is adjusted based on a feedback signal, and therefore the output is defined by the input signal as well as the feedback. However, in an open-loop system, there is no feedback, and therefore the output is defined solely by the input signal.

Perceptual Matching: At reaching distances, *perceptual matching* (or, just *matching*) is a closed-loop task for measuring depth judgments, where an observer moves a pointer towards the target object until the observer judges both objects to be at the same distance. Perceptual matching involves two subprocesses: (1) *visual perception*, where the observer perceives a distance difference between the pointer and the target object, and (2) a *matching action*, where the observer moves the pointer to reduce this difference. The distance difference between the pointer and target object is primarily perceived through *binocular disparity*, the stereo vision difference between the pointer and target object. The observer therefore performs the matching action to minimize disparity, making perceptual matching a visually closed-loop task. Along with visual feedback, perceptual matching also involves *proprioceptive feedback*, the observer's sense of the position of their arm, wrist, and fingers. Matching tasks have been widely used for real world depth judgments (Prablanc, Echallier, Komilis, and Jeannerod [24], Bingham et al. [3]), as well as in AR environments (Ellis and Menges [10], McCandless, Ellis, and Adelstein [15], Rolland, Gibson, and Ariely [25], Rolland, Meyer, Arthur, and Rinalducci [26], Edwards et al. [9], Singh, Swan II, Jones, and Ellis [28]).

Although perceptual matching has been widely used, a number of scientists do not consider the task to measure what has been called *definite distance perception* (Bingham and Pagano [4], Prablanc et al. [24]). The reason is that the visual feedback available during matching means the task involves minimizing the disparity between the target and pointer, as described above. However, two objects can have the same disparity between them at many different distances from the observer. Therefore, this technique gives only a *relative* measure of distance perception: the pointer is placed in depth only relative to the depth of the target object, and therefore the task does not require the observer to have an internal representation of the distance to the target object (Bingham and Pagano [4]).

Blind Reaching: This has motivated the development of visually open-loop perception-action tasks, in particular *blind reaching* (or, just *reaching*). With blind reaching, an observer reaches to a target object with their hand, which is hidden from their view. Because the hand is hidden, correcting the reach based on visual feedback is not possible, and therefore the observer must rely on some internal sense of perceived distance to perform the task. Blind reaching is therefore considered to measure *definite distance perception* (Bingham and Pagano [4], Prablanc et al. [24]). The method has been widely used to study depth perception at reaching distances in both real and virtual environments (Bingham, Zaal, Robin, and Shull [5], Mon-Williams and Tresilian [16, 17], Naceri, Chellali, and Hoinville [19], Napieralski, Altenhoff, Bertrand, Long, Babu, Pagano, Kern, and Davis [20], Altenhoff, Napieralski, Long, Bertrand, Pagano, Babu, and Davis [1]).

With blind reaching, no visual feedback is available, and so observers use proprioception to sense when their hand has reached the target. However, proprioception has been found to be pliable and easily susceptible to drift in the absence of corrective feedback (Wann, Rushton, and Mon-Williams [33], Paillard and Brouchon [22]). Based on this concern, Bingham and Pagano [4] advocate using perception-action tasks with some form of feedback when evaluating definite distance perception. In addition, the pliability of proprioception is the likely reason why the accuracy of blind reaching has been found to vary rather widely: while some studies have found very accurate responses for reaching (Mon-Williams and Tresilian [16], Mon-Williams, Wann, Jenkinson, and Rushton [18], van Beers, Sittig, and Denier van der Gon [31], Wann [32]), other studies have found far less accurate responses, with median errors of up to 25 cm (Foley [11]). To the best of our knowledge, to date our group has published the only study that used blind reaching to study depth judgments in an AR environment (Singh et al. [28]).

Despite the fact that perceptual matching only measures relative distance perception, as a task it models the primary interaction for many important AR applications. In particular, for image-guided surgery, as well as manufacturing applications, the primary interaction involves placing a real object, such as a scalpel, at a location indicated by a virtual marker (Kersten-Oertel et al. [13], Edwards et al. [9], Curtis et al. [7]). The engineers and surgeons developing these AR applications are not concerned with whether this task involves relative or definite distance perception. Instead, they are primarily concerned about accuracy and precision limits, and what training it might take to achieve these limits.

Experimental Purpose: Therefore, in the experiments reported here, we studied depth judgments of real and AR objects, using perceptual matching and blind reaching tasks, with the purpose of comparing the results from both tasks. An additional purpose, driven by the accuracy requirements of image-guided surgery (Edwards et al. [9]), was to determine how to build an apparatus that allowed us to measure depth judgments with accuracy and precision limits of at most a few millimeters[1].

2 Experiment I: Real vs. AR, Matching vs. Reaching

In Singh et al. [28], our group reported an experiment that used perceptual matching and blind reaching to study depth judgments of AR targets in the presence and absence of a highly-salient occluding surface, at reaching distances of 34 to 50 cm. We found relatively accurate performance for perceptual matching, $\sim$4 cm of underestimation for blind reaching, and complex effects when an occluding surface was present. However, this experiment only studied AR targets, which do not have a ground-truth location that can be objectively measured in the real world. Therefore, an initial purpose of Experiment I was to replicate Singh et al. [28], while addressing this limitation. Accordingly, in Experiment I we compared AR and real targets, and employed both perceptual matching and blind reaching tasks.

As discussed above, perceptual matching is a visually closed-loop task, while reaching is a visually open-loop task, and therefore we anticipated that perceptual matching judgments would be more accurate than blind reaching results. However, we did not know how much more accurate matching would be over reaching, and determining this was a major purpose of Experiment I. In addition, because—by definition—real targets are seen with accurate depth cues, while all AR displays present virtual objects with degraded depth cues, we anticipated that real target judgments would be more accurate than AR target judgments. However, we did not know the magnitude of this accuracy difference, and measuring this—the accuracy cost for AR versus real targets—was another major purpose of Experiment I.

2.1 Method

2.1.1 Apparatus and Task

We developed a table apparatus for accurately measuring depth judgments. Our table was based on an apparatus first described by Ellis and Menges [10], and then employed by Singh et al. [28], but modified to

[1] Portions of these experiments have been reported in the form of poster abstracts (Singh, Swan II, Jones, and Ellis [29, 30]) and a PhD dissertation (Singh [27]).

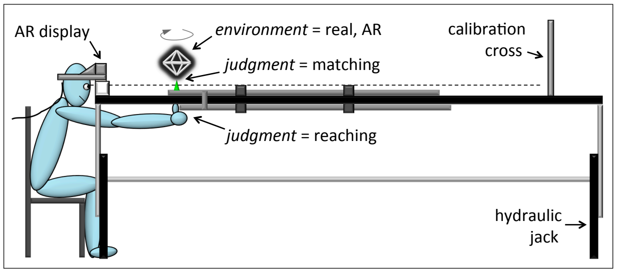

Fig. 2. Experiment I: side-view diagram of the table apparatus.

present either a real or AR target. Fig. 2 shows a side-view diagram of the table and depicts the depth judgment tasks, while Fig. 1a shows an observer performing the experiment. The top of the table was a custom-designed optical breadboard, 244 cm long by 92 cm wide by 3.8 cm thick. As shown in Fig. 1, we covered the table with black cloth, creating a smooth and featureless surface. We created an aluminum frame for holding the table, which had six legs, including two middle legs that extended upwards and held two tracking cameras. We equipped the table with a six-jack hydraulic lift system (Fig. 2), which allowed us to raise and lower the entire apparatus, and position the tabletop between 104 cm and 134 cm above the ground. This allowed the seated observer to comfortably rest the AR display against the edge of the table (Figs. 1a, 2). In this position, the observers' pupils were $\sim$3.5 cm above the table surface. Finally, we hung an opaque, black background curtain 220 cm from the end of the table (Fig. 1a).

Also as shown in Fig. 1a, we mounted plastic pipes, which easily slid through plastic collars, along the left-hand and right-hand sides of the table. On the left-hand pipe we attached a supporting frame that held the real target (Fig. 1a), which allowed us to position the target at different distances from the observer. The real target was a slowly rotating (4 rpm) wireframe octahedron, painted a bright white color, with a 10 cm square base and 10 cm height (Fig. 1a). When we did not want observers to see the real target, we rotated the entire supporting frame out of the observer's field of view.

Our virtual AR target was an exact replica of the real target. We carefully calibrated the AR target so that it precisely matched the size and position of the real target at every tested distance. We also carefully designed the lighting so that the real and AR targets appeared as similar as possible. In particular, the lighting made the real target appear to glow against an otherwise dark background, and we ensured that the real target did not cast any visible shadows or reflections. We visually matched the apparent brightness of the real and AR targets.

Observers performed a depth judgment by reaching under the table with either hand, and grabbing a handle mounted on the end of a pipe attached to the center of the underside of the table (Fig. 2). This pipe was attached to the right-hand pipe on top of the table (Fig. 1a), so as the observer moved their hand in depth, both pipes slid in depth. When the observer made a matching judgment, we attached a bar to the right-hand pipe (attached in Fig. 1a). At the end of this bar was a small pointer, approximately 4 mm in diameter. We calibrated the target's height so that there was an $\sim$1 cm gap between the pointer and the target's bottom tip. When the observer made a reaching judgment, this bar was not attached. For either type of judgment, matching or reaching, when the judgment was correct, the observer's thumb was directly below the tip of the rotating target. In addition, unlike Singh et al. [28], where observers used different gestures to make matching and reaching judgments, here the gesture was identical for either type of judgment.

Observers wore an nVisor ST60 head-mounted AR display, by NVIS, Inc. This is an optical see-through display, with a resolution of 1280 × 1024 per eye, a 60° diagonal field-of-view, and 100% stereo overlap. It supports inter-pupillary distances from 53 to 73 mm, by independently adjusting the horizontal position of left and right monocles. The display's optical elements are collimating, presenting the scene at an infinite focal depth, and are not adjustable. The display is heavy, weighing 1.56 kg. However, because observers rested the display against the edge of the table (Fig. 2), they felt much less weight. In addition, two foam mounting brackets precisely positioned the display (Fig. 1a), and this position, combined with the display's field of view, ensured that observers could not see either of the left- or right-hand sliders, nor their hand when they made a depth judgment.

A TrackPack system by A.R.T. GmbH provided both 3 degree-of-freedom (DOF) and 6 DOF tracking. We measured the accuracy and precision of this tracker to be better than 1 mm. We attached a 6 DOF tracking configuration to the AR display, which allowed us to render the AR target at precise real world locations. We also attached a 3 DOF tracking target to the supporting frame that held the real target (Fig. 1a), which allowed us to precisely position the target in depth. Finally, we attached another 3 DOF tracking target to the right-hand pipe, which allowed us to measure and automatically encode observers' depth judgments.

2.1.2 Experimental Design

Independent Variables: We recruited 40 *observers* from a population of university students and staff. The observers ranged in age from 18 to 27, the mean age was 20.0, and 23 were female and 17 male. We paid 6 observers $12 an hour, and the rest received course credit. Observers performed the experiment in two *environment conditions*, real and AR. Observers performed two kinds of *depth judgments*, matching and reaching. The target object appeared at 5 different *distances* from the observer: 34, 38, 42, 46, and 50 cm. Finally, observers saw 6 *repetitions* of each combination of the other dependent variables.

Dependent Variables: The primary dependent variable was judged distance, which we measured using either the matching or the reaching depth judgment. We also calculated error = judged distance – actual distance. An error = 0 cm indicated an accurately judged distance, an error $>$ 0 cm indicated an overestimated distance, and an error $<$ 0 cm indicated an underestimated distance.

Design: The primary variables were *environment* (real, AR) and *depth judgment* (matching, reaching). We used a 2 × 2 between-subjects design, with four main conditions: (1) matching, real; (2) matching, AR; (3) reaching, real; (4) reaching, AR. There were 10 observers in each condition. We varied the presentation order of the main condition in a round-robin fashion, so each group of four observers covered all conditions. We randomly permuted distance × repetition, with the restriction that the distance changed every trial. Therefore, each observer completed 5 (distance) × 6 (repetition) = 30 trials. We measured the judged distance for every trial, and collected a total of 1200 data points (40 observers × 30 trials).

2.1.3 Procedure

For each observer we used a pupilometer, with the vergence distance set to 40 cm, to measure their inter-pupillary distance. Following this measurement, we described the depth judgment task—matching or reaching—to the observer, and demonstrated the task using the real target. For reaching judgements, we instructed observers to slide the pipe under the table, until their unseen thumb was below the tip of the target. We adjusted the table height for the observer, and they practiced the task—matching or reaching—three times, again using the real target. Observers did not wear the AR display while practicing.

We next fitted the display on the observer's head, and calibrated the display using the techniques described by Singh et al. [27, 29]. After calibration, (1) the observer was looking through the optical center of each of the display's eyepieces, (2) translational tracker errors related to the way the display fits on the observer's head were corrected, and (3) rotational tracker errors also related to the display's fit were corrected. Following calibration, the observer performed the experiment.

2.2 Analysis

To analyze the data, we calculated linear functions that predict judged distance from actual distance, and examined the resulting slopes and intercepts. We used multiple regression methods (Pedhazur [23]) to determine if the slopes and intercepts significantly differed. We found

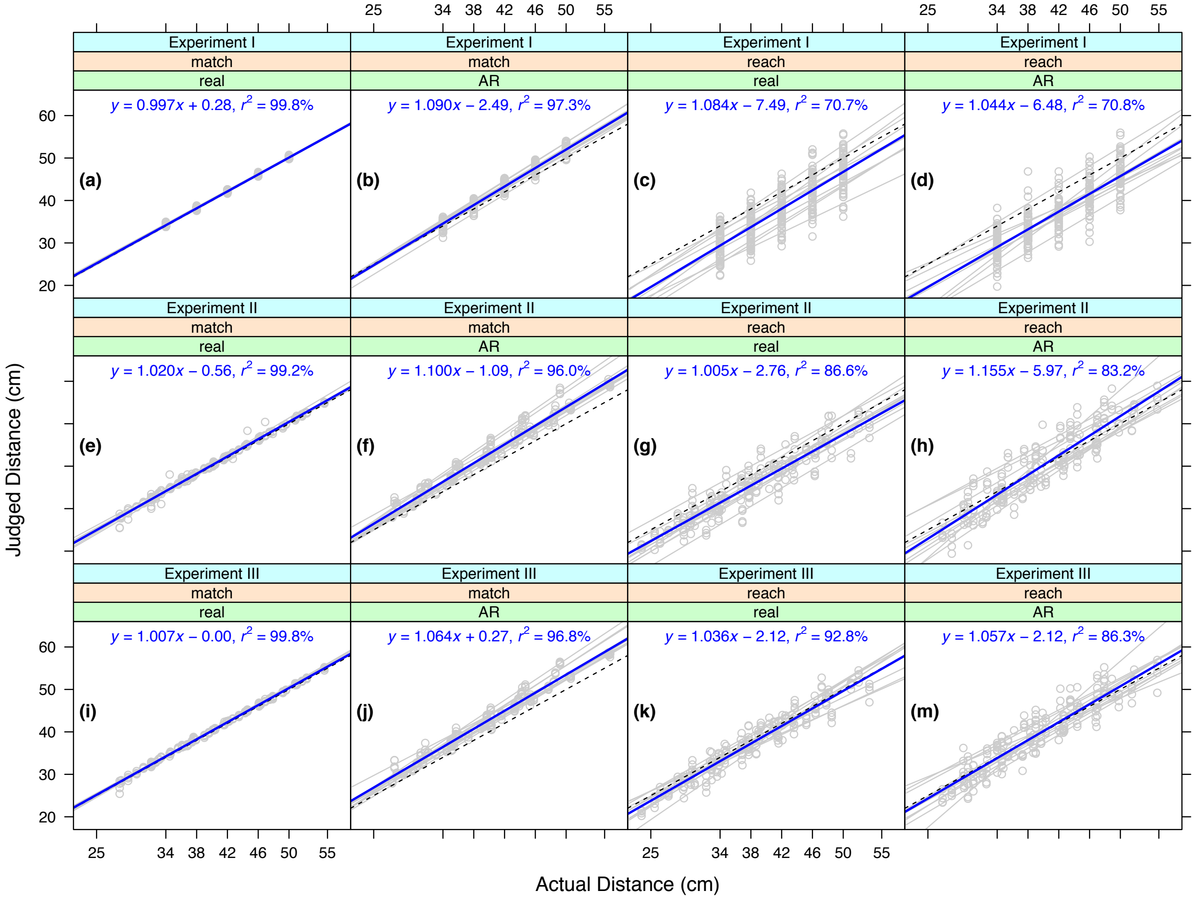

Fig. 3. The results for all three experiments, plotted as a scatterplot of judged against actual distance, with $N = 1200$ (Experiment I) and $N = 800$ (Experiments II and III) ghosted data points, and fit with blue regression lines for each main experimental condition. The thin dashed line in each panel represents veridical performance. The ghosted light grey lines are the linear fits of individual observers.

multiple regression methods preferable to ANOVA analysis, because multiple regression allows us to predict a continuous dependent variable (judged distance) from a continuous independent variable (actual target distance), as well as a categorical independent variable, such as environment × judgment. With ANOVA analysis, we are restricted to only examining categorical independent variables, which results in a significant loss of power when an independent variable is inherently continuous (Pedhazur [23]). Finally, multiple regression yields slopes and intercepts, which as descriptive statistics are more useful than means, because they directly describe functions that predict judged distances from actual target distances[2].

2.3 Results

Figs. 3a–d and 4a–d show the results from Experiment I, plotted as a scatterplot of judged against actual distance (Fig. 3), as well as mean error against distance (Fig. 4). Both figures indicate that the data is very well fit by linear equations; note the r^2 values in Fig. 3. Fig. 5a–d shows the results of multiple regression analysis, which compares the linear fits from Fig. 3a–d against each other. To properly account for repeated measurements, for each observer at each distance, we averaged the responses over the 6 repetitions, reducing the size of the analyzed dataset from 1200 to 200 points—note the reduced density of points in Fig. 5a–d relative to Fig. 3a–d.

Each panel in Fig. 5 compares two linear fits from Fig. 3. Our multiple regression analysis first tests whether the *slopes* of the linear fits significantly differ. If they do, as in Fig. 5a, we report both linear fits from Fig. 3 as the best overall description of the data in the panel. If the slopes of the linear fits do not significantly differ, we next test whether the *intercepts* of the linear fits significantly differ. This test first sets the slopes of the linear fits—which do not differ—to a common value. If the intercepts significantly differ, as in Fig. 5c, we report the two linear fits from Fig. 3, with the slopes adjusted to a common value, as the best overall description of the data in the panel. If neither the slopes nor the intercepts significantly differ, as in Fig. 5b, then we report the simple regression $y = x$, where y is judged distance and x is actual target distance, as the best overall description of the data in the panel. Therefore, this multiple regression analysis yields three possible outcomes, which by chance are illustrated in the first three panels of Fig. 5: (1) the slopes significantly differ (Fig. 5a), (2) the slopes do not differ but the intercepts significantly differ (Fig. 5c), or (3) neither the slopes nor the intercepts significantly differ (Fig. 5b). In each case, the panel also indicates two measures of effect size: (1) the overall R^2 value, the percentage of variation in the panel explained by the linear regressions, and (2) the percentage of variation explained by the change in categorical value. If the variation explained by the change in categorical value is too small, we do not perform hypothesis testing, because any statistical differences would be too small to be meaningful (Pedhazur [23]). Based on the results reported in this paper, we require an effect size of at least 0.1% of variation to justify hypothesis testing. Note that Fig. 5 graphically illustrates each pos-

[2] In addition, multiple regression is a *superset* of ANOVA, and is often the algorithm by which ANOVA results are calculated (e.g., the General Linear Model in SPSS). In addition to Pedhazur [23], multiple regression techniques are discussed in standard textbooks such as Cohen, Cohen, West, and Aiken [6]. These methods have been used to analyze a large number of experiments that have examined depth judgments of real and virtual targets, such as Altenhoff et al. [1], Bingham et al. [3], Bingham and Pagano [4], Napieralski et al. [20], and Pagano and Isenhower [21].

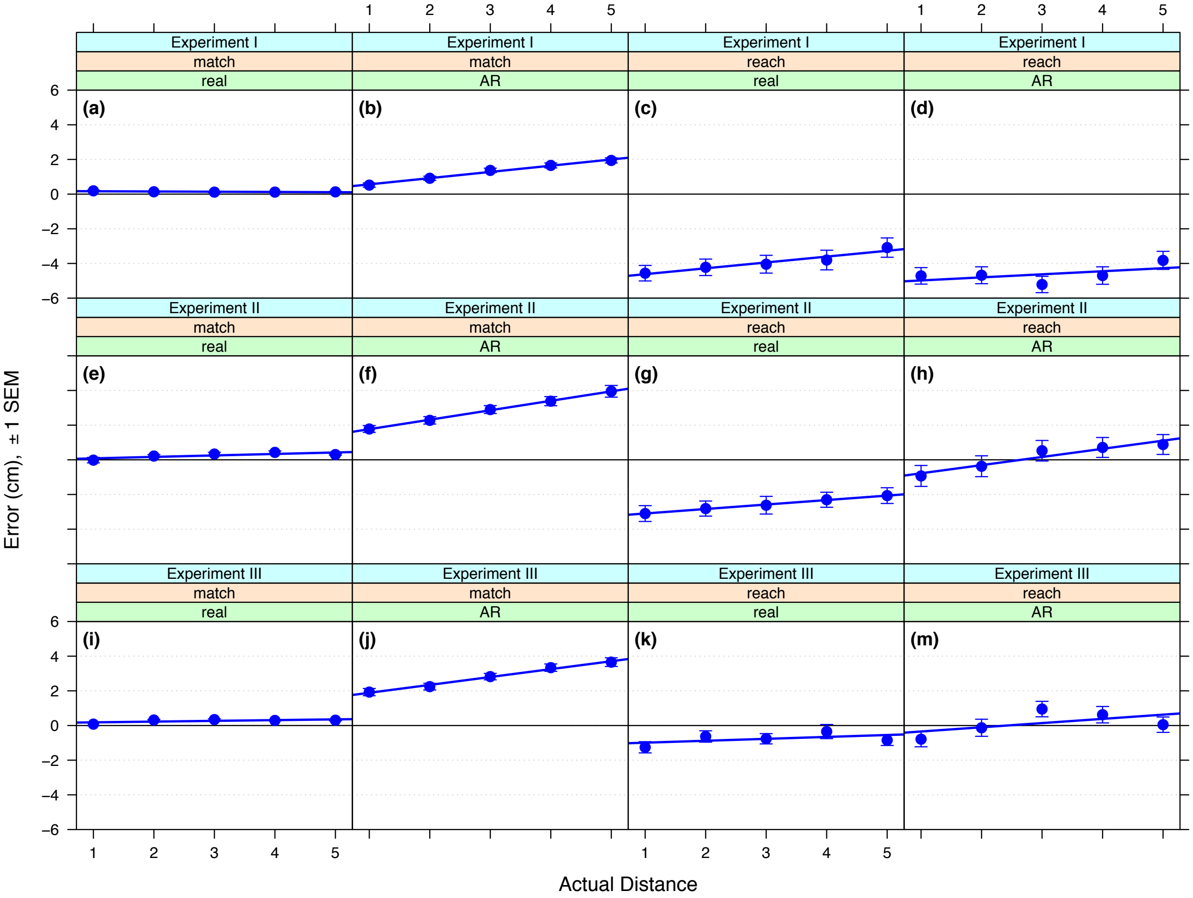

Fig. 4. The results for all three experiments, plotted as mean error against distance, with $N = 1200$ (Experiment I) and $N = 800$ (Experiments II and III). For Experiment I, the actual distances are 34, 38, 42, 46, and 50 cm, while for Experiments II and III the actual distances are 55, 63, 71, 79, and 87% of each observer's maximum reach.

sible statistical outcome. Additional statistical details of the multiple regression calculations are included as supplemental material for this paper (Appendix A.1).

Fig. 5a compares matching of real versus AR targets. The slopes of the linear fits significantly differ across the tested distances ($F_{1,96} = 14.0, p < 0.001$). Matching real targets was very accurate, with an average error of only 1.4 mm (Fig. 4a), which means that observing the targets through the display's optical combiners did not hamper matching. However, matching AR targets was increasingly overestimated, from 0.5 cm at 34 cm to 1.9 cm at 50 cm (Fig. 4b). Fig. 5b compares reaching for real versus AR targets. Here neither the slopes ($F_{1,96} = 0.12, p = 0.73$) nor the intercepts ($F_{1,97} = 1.09, p = 0.30$) significantly differ, and the reaching data is best fit by the single equation $y = 1.064x - 6.99$ cm. Therefore, Experiment I did not find a difference between reaching for real versus AR targets. Fig. 5c compares matching versus reaching of real targets. Here the slopes do not significantly differ ($F_{1,96} = 1.01, p = 0.32$), but the intercepts do ($F_{1,97} = 68.1, p < 0.001$). Matching real targets was very accurate, but reaching for real targets was underestimated by a constant 4.1 cm. Finally, Fig. 5d compares matching versus reaching of AR targets. Here again the slopes do not significantly differ ($F_{1,96} = 0.32, p = 0.57$), but the intercepts do ($F_{1,97} = 170.5, p < 0.001$). Reaching for AR targets resulted in a constant 5.9 cm of underestimation relative to matching AR targets.

Therefore, the overall findings for Experiment I were accurate matching of real targets, increasingly overestimated matches for AR targets, and underestimated reaching for both real and AR targets. In addition, reaching was notably less precise than matching: note the spread of the means in Fig. 3 for reaching versus matching, as well as the size of the error bars in Fig. 4. Also, in Fig. 3, the ghosted grey lines are linear fits for each of the 10 observers in each condition. We see very little variation in per-observer fits for matching, but substantial variation for reaching.

We next consider the AR matching results in detail; the underestimated reaching is further considered in Experiments II and III.

2.4 AR Matching Results

Here we consider the significant differences we found between matching real versus AR targets (Fig. 5a). Our AR display generates collimated images—the light rays are parallel—which means the AR targets were presented at optical infinity. This is the most salient optical difference between the real and AR targets. Observers who saw the real targets focused and verged to the same distance, and very accurately matched the targets. However, observers who saw the AR targets focused at infinity, but verged at the target distance. This accommodative / vergence mismatch is known to drive the resting vergence angle of the eyes outward, and is expected to result in targets being perceived as farther than their actual positions (Mon-Williams and Tresilian [17], Wann et al. [33]).

We therefore developed a model for explaining the accurate matching of real targets versus the increasingly overestimated matching of AR targets. Recall that the AR target was a virtual object, seen with accommodative / vergence mismatch, while the pointer was a real object, seen with consistent accommodative and vergence cues. To perform the matching task, the observer attempted to minimize the binocular disparity between the AR target and the pointer, which means that the observer's gaze was constantly shifting between the two objects. In our model, whenever an observer's gaze is on the AR target, their ver-

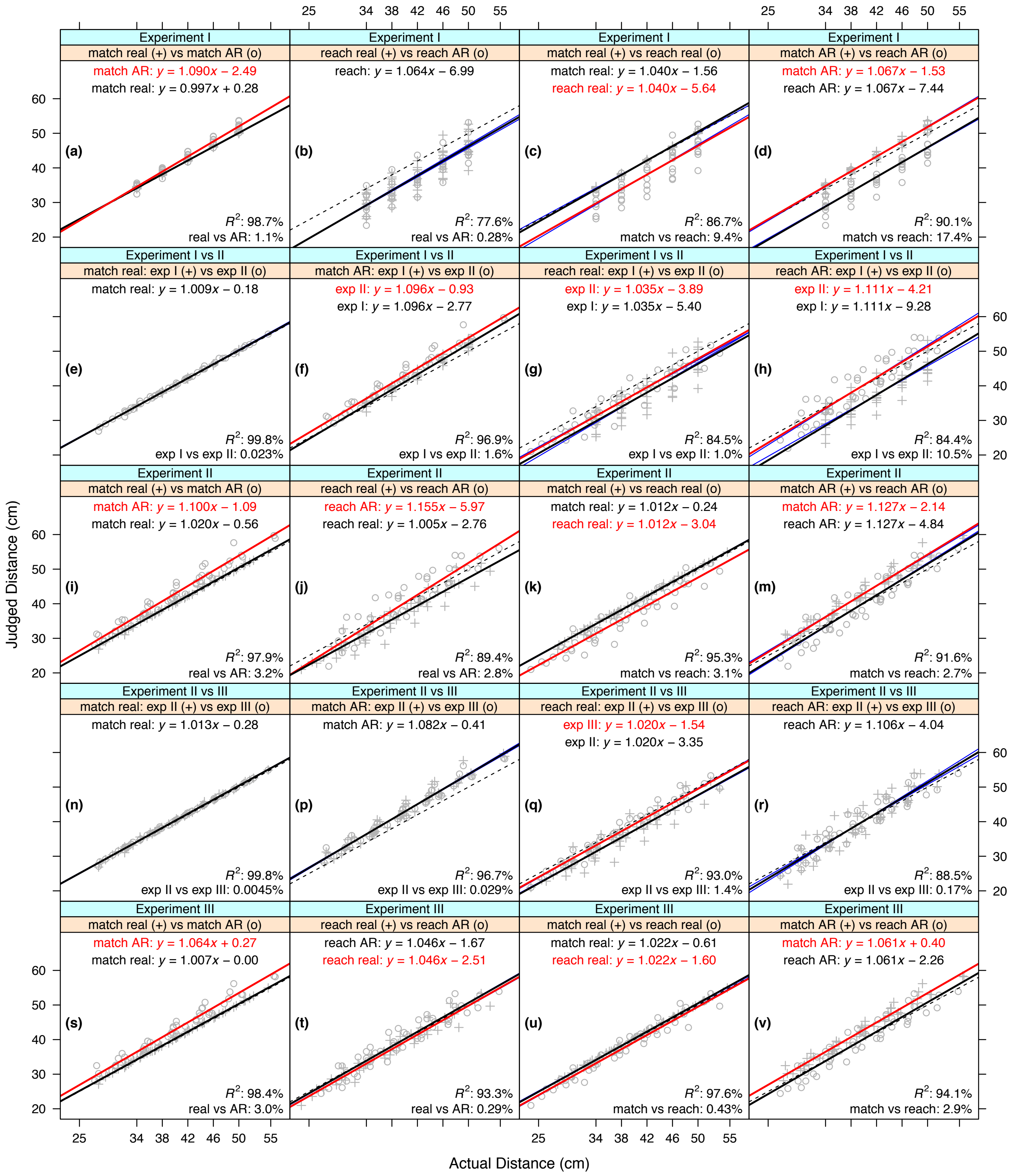

Fig. 5. Multiple regression analysis, plotted as a scatterplot of judged against actual distance, with $N = 200$ (Experiments I, II, and III) and $N = 400$ (Experiments I vs II and II vs III) ghosted data points, marked as $+$ or $\circ$. The thin dashed lines represent veridical performance. Blue lines represent fitted regression lines from Fig. 3. Black and red lines represent the linear regressions shown in each panel. Blue lines are not visible when overlaid by black or red lines; the degree of blue line visibility is a graphical indication of how closely the regressions in each panel agree with the regressions from Fig. 3.

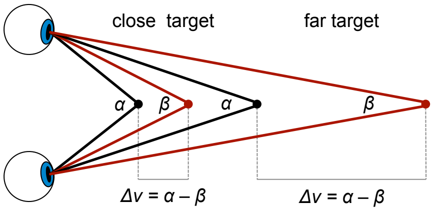

Fig. 6. The effect of a constant change in vergence angle.

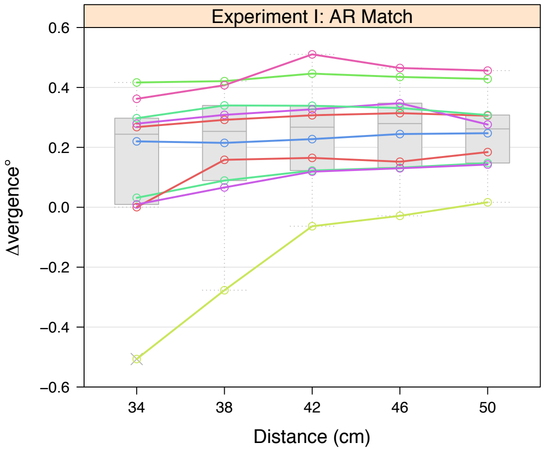

Fig. 7. The computed change in vergence angle with distance for the AR match data.

gence angle is biased outward by a constant amount, relative to their vergence angle when their gaze is on the pointer.

Fig. 6 illustrates this model. Suppose that an observer is gazing at the real target: their angle of binocular parallax is α. Now suppose that they shift their gaze to the AR target: their angle of binocular parallax increases to β. Our model is that $\Delta v = \alpha - \beta$, the change in angle of binocular parallax, is constant at every tested distance. In Fig. 6, the left-hand pair of angles illustrate matching a close target (e.g. 34 cm). Here Δv results in a virtual target β appearing farther than a real target α. The right-hand pair of angles illustrate matching a farther target (e.g. 50 cm). Here the same Δv results in a larger distance between a virtual target β and a real target α. This pattern matches the collected data (Fig. 4b): to minimize the disparity between the AR target and the real pointer, as actual distance increases, observers increasingly place the pointer beyond the correct distance.

In Fig. 7 we calculate α, β, and Δv for the 10 AR matching observers. We calculate $\alpha = 2\arctan(i/2x)$, where i is the observer's inter-pupillary distance and x is the actual target distance. Note that using x in this formula assumes that the observer would match a real object with perfect accuracy, but the very accurate and precise results for matching real objects suggest this assumption is reasonable. We then calculate $\beta = 2\arctan(i/2y)$, where y is the judged distance. Fig. 7 shows the resulting Δv for the 10 AR matching observers at each distance. For 9 of the 10 observers Δv changes less than 0.2° across the 5 distances, and for the outlying observer it changes by 0.52°. The median line seen in the under-printed boxplot changes less than 0.05°. These angular changes are small, and support our model that the collimating optics of our AR display drive a constant outward bias in vergence angle, resulting in increasingly overestimated matches for AR targets.

2.5 Discussion

The purpose of Experiment I was to compare AR and real targets, using both perceptual matching and blind reaching tasks. Because perceptual matching is a visually closed-loop task, while blind reaching is a visually open-loop task, we anticipated that perceptual matching would be both more more accurate and precise than blind reaching. Experiment I confirmed this finding: Fig. 5c, d shows that matching is more accurate than reaching for both real and AR objects, and Fig. 3a–d shows that matching has much less variability than reaching. Overall, reaches were ~4 cm underestimated, which replicates our finding from Singh et al. [28]. However, while Singh et al. [28] only studied AR targets, in Experiment I we replicated this degree of underestimation for both AR and real targets. We also anticipated that real judgments would be more accurate and precise than AR judgments. This hypothesis was confirmed for matching (Fig. 5a), but not for reaching (Fig. 5b).

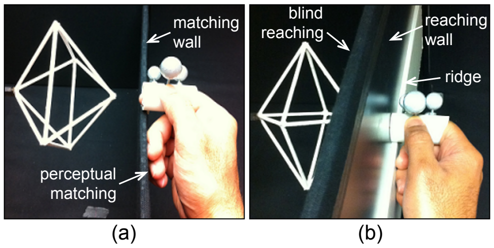

Fig. 8. Experiment II: (a) perceptual matching and (b) blind reaching tasks.

Comparing our reaching results to previous studies involving real target objects, our findings were as accurate as some (van Beers et al. [31], Mon-Williams et al. [18]), but notably less accurate than others (Mon-Williams and Tresilian [16], Wann [32]). In addition, as discussed in Section 1 above, blind reaching uses proprioception, which is pliable and susceptible to drift in the absence of corrective feedback, and this has lead Bingham and Pagano [4] to call for using corrective feedback with blind reaching. However, Mon-Williams and Tresilian [16] achieved very accurate blind reaching of real targets, even in the absence of corrective feedback. Their data yielded the regression $y = 1.08x - 1.35$ cm, which matches Experiment I's reach, real regression $y = 1.084 - 7.49$ cm in slope, while being considerably more accurate in intercept (Fig. 3c). Furthermore, their experimental apparatus and task is generally similar to our own.

We therefore conducted two additional experiments to see if reaching judgments could be improved. In Experiments II and III, we first modified our apparatus and task to closely model the apparatus and task of Mon-Williams and Tresilian [16]. In addition, we implemented an idea explored by Pagano and colleagues [1, 20, 21], where proprioceptive feedback is normalized between observers by setting the target distances to constant percentages of each observer's maximum arm reach. In this way, even though observers' arms are different lengths, they use the same arm gesture when reaching to each distance. Experiment II did not use corrective feedback, while Experiment III did use corrective feedback.

3 Experiment II: Apparatus and Task

The purpose of Experiment II was to replicate Experiment I, but very closely implement the blind reaching apparatus and task of Mon-Williams and Tresilian [16]. In addition, in Experiment II distances were constant percentages of each observer's arm length. Before running the experiment, we anticipated that these modifications would increase the accuracy of reaching results, relative to Experiment I. We also anticipated that matching results would be similar to Experiment I.

3.1 Method

Experiment II used the same method and procedures as Experiment I, except for what is noted here.

3.1.1 Apparatus and Task

Fig. 1b illustrates the primary difference in Experiment II's task: observers used their finger to point at the target. To support this direct pointing gesture, we mounted the target sideways. The tip of the target was at eye level, 25 cm above the tabletop, and in the line of sight of the observer's right eye. Observers viewed the target against a background wall positioned 72 cm away. As in Experiment I, we carefully calibrated the AR target to be the same size as the real target at every tested distance, and both targets appeared visually similar; the real target did not cast any visible shadows or reflections.

For the matching task (Fig. 8a), observers judged the distance to the tip of the target by resting their right index finger on top of the *matching wall*. Observers wore a pointer on their index finger—a short section of pipe embedded with tracking fiducials. In addition, a small screw extended out of the bottom of the pointer, which made it easy for observers to place the tip of their finger on the matching wall without accidentally reaching over and touching the target. The matching wall was just tall enough so that the tip of the observer's finger was the same height as the tip of the target. We asked observers to match the tip of the nail of their index finger to the tip of the target, which provided a specific point to match to another specific point.

For the reaching task (Fig. 8b), we replaced the matching wall with a *reaching wall*, which was tall enough to hide the observer's view of their hand. During a reaching gesture, observers rested the tip of their index finger on a ridge mounted the same height as the matching wall. These two designs resulted in exactly the same biomechanical movement for matching and reaching.

3.1.2 Experimental Design

We recruited 40 new *observers* from a population of university students; no observers from Experiment I participated. The observers ranged in age from 19 to 28; the mean age was 20.7, and 27 were male and 15 female. The maximum reach of the observers ranged from 43.3 cm to 64 cm; the mean maximum reach was 55.1 cm. Observers received course credit for their participation. Observers again performed the experiment in two *environment conditions* and used two kinds of *depth judgments*. The target object appeared at 5 different *distances* from the observer: 55, 63, 71, 79, and 87% of the observer's maximum reach, ranging from 23.8 cm to 55.7 cm. Finally, observers saw 4 *repetitions* of each combination of the other dependent variables.

As with Experiment I, the primary variables were *environment* (real, AR) and *depth judgment* (matching, reaching). We used a 2 × 2 between-subjects design, with four main conditions: (1) matching, real; (2) matching, AR; (3) reaching, real; (4) reaching, AR. There were 10 observers in each condition, and each observer completed 5 (distance) × 4 (repetition) = 20 trials. We collected a total of 800 data points (40 observers × 20 trials).

3.2 Results

Figs. 3e–h and 4e–h show the results from Experiment II. While Fig. 3e–h shows all 800 data points, for the multiple regression analysis, for each observer we averaged the responses over the 4 repetitions, reducing the size of the dataset to 200 points.

Fig. 5e–h compares each condition of Experiment II with the corresponding condition of Experiment I. As shown in Fig. 5e, for matching real targets the effect size of the difference between Experiments I and II is 0.023% of the variation, which is too small for any statistical differences to be meaningful. Therefore, the matching real data is best fit by the single equation $y = 1.009x - 0.18$ cm, indicating very accurate matching over both experiments. Fig. 5f compares matching in AR. Here the slopes do not significantly differ ($F_{1,96} = 0.07, p = 0.80$), but the intercepts do ($F_{1,97} = 48.7, p < 0.001$). In Experiment II observers additionally overestimated AR matches by a constant 1.8 cm. Fig. 5g compares reaching to real targets. Here again the slopes do not significantly differ ($F_{1,96} = 0.70, p = 0.41$), but the intercepts do ($F_{1,97} = 5.72, p = 0.019$). Reaching for real targets improved in Experiment II by a constant 1.5 cm. Finally, Fig. 5h compares reaching to AR targets. Again the slopes do not significantly differ ($F_{1,96} = 1.24, p = 0.27$), but the intercepts do ($F_{1,97} = 64.3, p < 0.001$). Reaching for AR targets improved in Experiment II by a constant 5.1 cm.

From these results we conclude that the modified apparatus and task in Experiment II did improve reaching results relative to Experiment I. Furthermore, this improvement was constant, with only intercept changes, for both real (1.5 cm) and AR (5.1 cm) targets; compare Fig. 4c to 4g and 4d to 4h. The improvement for AR targets was quite substantial. However, we also anticipated that matching results would be unchanged in Experiment II. Although this was the case for real targets, where observers remained very accurate, for AR targets matching was additionally overestimated by 1.8 cm (compare Fig. 4b to 4f).

Fig. 5i–m compares the main conditions in Experiment II against each other. Fig. 5i compares matching of real and AR targets. The slopes of the linear fits significantly differ across the tested distances ($F_{1,96} = 6.38, p = 0.013$); while matching real targets was very accurate, matching AR targets was increasingly overestimated. Fig. 5j compares reaching for real and AR targets. Here, unlike Experiment I, where we found no difference, the slopes show a trend of significantly differing ($F_{1,96} = 3.62, p = 0.060$), and if we were to consider this non-significant, then the intercepts strongly differ ($F_{1,97} = 20.9, p < 0.001$). Either way, the interpretation is that in Experiment II there was a significant difference between reaching for real and AR targets, and we have judged that two fits with different slopes is the best overall description of this difference. Fig. 5k compares matching and reaching for real targets. Here the slopes do not significantly differ ($F_{1,96} = 0.10, p = 0.76$), but the intercepts do ($F_{1,97} = 63.9, p < 0.001$). Matching real targets was very accurate, but reaching for real targets was underestimated by a constant 2.8 cm. Finally, Fig. 5m compares matching and reaching of AR targets. Here again the slopes do not significantly differ ($F_{1,96} = 0.61, p = 0.44$), but the intercepts do ($F_{1,97} = 30.1, p < 0.001$). Reaching for AR targets was relatively accurate on average, although with a slope $\gg 1$, but matching AR targets was overestimated by a constant 2.7 cm.

Overall, the pattern of findings for Experiment II was similar to Experiment I, with the exception that reaching became more accurate on average—compare the first two rows of Fig. 4. In addition, also like Experiment I, AR matching remained increasingly overestimated with distance. The model of the collimating display biasing the observer's vergence angle outward by a constant amount also explains this finding in Experiment II. When we calculate Δv for the 10 AR matching observers and draw a plot like Fig. 7, the plot qualitatively looks very similar. For 9 of the 10 observers Δv changes less than 0.2° across the 5 distances, while for the outlying observer it changes 0.76°. The median Δv changes less than 0.07° across the 5 distances.

3.3 Discussion

The purpose of Experiment II was to replicate Experiment I, but determine if two modifications would improve the reaching results: (1) closely implementing the blind reaching apparatus and task of Mon-Williams and Tresilian [16], who reported very accurate reaching results, and (2) setting tested distances to constant percentages of arm length, to equalize the reaching gestures between observers. The modifications were successful, in that reaching become more accurate, by 1.5 cm for real targets and 5.1 cm for AR targets. However, AR matching became 1.8 cm less accurate.

Therefore, Experiment II did improve the reaching data, resulting in a linear fit of $y = 1.005x - 2.76$ cm when reaching for real targets (Fig. 3g). However, this is still less accurate than Mon-Williams and Tresilian's [16] linear fit of $y = 1.08x - 1.35$ cm. But, additional improvement is likely possible by employing corrective feedback to calibrate the proprioception involved in reaching, as called for by Bingham and Pagano [4]. We employed this feedback in Experiment III.

After collecting the data for Experiment II, the same observers received a round of corrective feedback, and then repeated the same tasks. Therefore, Experiment II's observers participated in a *pretest*, *intervention*, *posttest* design, where the *pretest* data is reported as Experiment II, the *intervention* involved interaction with corrective feedback for the reaching observers, and the *posttest* data is next reported as Experiment III. We anticipated that reaching observers would improve during the posttest. In addition, during the intervention phase matching observers also continued to interact with the apparatus, but did not receive additional corrective feedback. We did this to equalize the amount of time observers spent during the intervention. Because matching observers already obtained corrective feedback during closed-loop matching, we anticipated that their posttest matching would not differ from their pretest matching.

4 Experiment III: Corrective Feedback

The purpose of Experiment III was to test whether an intervention phase of closed-loop matching, which provides corrective feedback, would further improve the reaching results from Experiment II. In addition, after the intervention phase for both reaching and matching, observers had spent ∼30 to ∼40 minutes performing the tasks and interacting with the apparatus. Therefore, a secondary purpose of Experiment III was to collect data from these now practiced observers.

4.1 Method

Experiments II and III were part of one larger experiment, which we judged was more clearly reported as two separate experiments in this paper. For the depth judgment task, observers either *reached*, *matched*, *reached* (RMR) or *matched*, *reached*, *matched* (MRM). Therefore, RMR observers reached during Experiment II, then performed a round of matching, and then reached again during Experiment III. MRM observers matched during Experiment II, then performed a round of reaching, and then matched again during Experiment III. Therefore, the larger experiment used a 2×2 between-subjects design, with four main conditions: (1) MRM, real; (2) MRM, AR; (3) RMR, real; (4) RMR, AR. During the pretest phase (Experiment II), each observer completed 5 (distance) × 4 (repetition) = 20 trials. During the intervention phase each observer completed an additional 5 (distance) × 4 (repetition) = 20 trials, and then during the posttest phase (Experiment III) each observer completed a final 5 (distance) × 4 (repetition) = 20 trials. All methods and procedures remained the same between the pretest, intervention, and posttest phases. For Experiment III we collected a total of 800 data points (40 observes × 20 trials).

4.2 Results and Discussion

Figs. 3i–m and 4i–m show the results from Experiment III. While Fig. 3i–m shows all 800 data points, for the multiple regression analysis, for each observer we averaged the responses over the 4 repetitions, reducing the size of the dataset to 200 points.

Fig. 5n–r compares each condition of Experiment III with the corresponding condition of Experiment II. Fig. 5n compares matching real targets between the two experiments. Here the effect size is very small: 0.0045% of the variation, which is much too small for any statistical differences to be meaningful. Therefore, the real data for both experiments is best fit with the single equation $y = 1.013x - 0.28$ cm, indicating very accurate matching of real targets. Fig. 5p compares matching in AR. Here again the effect size of 0.029% is too small for any statistical differences to be meaningful, and the AR matching data for both experiments is best fit with the single equation $y = 1.082x - 0.41$ cm. Therefore, AR matching continued to be increasingly overestimated with increasing distance. Fig. 5q compares reaching for real targets. Here the slopes do not significantly differ ($F_{1,96} = 0.29, p = 0.59$), but the intercepts do ($F_{1,97} = 18.9, p < 0.001$). In Experiment III observers improved their reaches by a constant 1.8 cm; also compare Fig. 4g to 4k. Finally, Fig. 5r compares reaching for AR targets. Here neither the slopes ($F_{1,96} = 1.49, p = 0.23$) nor the intercepts ($F_{1,97} = 0.0018, p = 0.97$) significantly differ, indicating that AR reaching did not improve in Experiment III.

From these results we conclude that observers did become better at reaching for real objects after an intervention phase of closed-loop matching. While observers did not become better at reaching for AR objects, they were already relatively accurate in Experiment II; again compare Fig. 4h to 4m. Also, as anticipated, matching did not change between Experiments II and III.

A secondary purpose of Experiment III was to examine data from practiced observers, and Fig. 5s–v compares the main conditions against each other. Fig. 5s compares matching of real and AR targets. As with Experiments I and II, the slopes of the linear fits significantly differ across the tested distances ($F_{1,96} = 4.38, p = 0.039$); while matching real targets was very accurate, matching AR targets remained increasingly overestimated. Fig. 5t compares reaching for real and AR targets. Here the slopes do not significantly differ ($F_{1,96} = 0.14, p = 0.71$), but the intercepts do ($F_{1,97} = 4.07, p = 0.047$). Reaching for AR targets was slightly more accurate than reaching for real targets, by a constant 0.83 cm. Fig. 5u compares matching and reaching for real targets. Here the slopes do not significantly differ ($F_{1,96} = 0.74, p = 0.39$), but the intercepts do ($F_{1,97} = 16.3, p < 0.001$). Matching real targets was very accurate, but reaching for real targets was underestimated by a constant 0.99 cm. Finally, Fig. 5v compares matching versus reaching of AR targets. Here again the slopes do not significantly differ ($F_{1,96} = 0.016, p = 0.90$), but the intercepts do ($F_{1,97} = 47.8, p < 0.001$). Reaching for AR targets was relatively accurate, while matching AR targets resulted in a constant 2.7 cm of overestimation.

Overall, the pattern of findings from Experiment III was similar to Experiment II. However, several differences between conditions did became smaller with practice: The difference between reaching for AR versus real objects shrunk from an effect size of 2.8% (Fig. 5j) to 0.29% (Fig. 5t). And, the difference between matching versus reaching for real objects shrunk from 3.1% (Fig. 5k) to 0.43% (Fig. 5u). For both effects, note that the linear fits have moved noticeably closer together in Experiment III. On the other hand, comparing Fig. 5i to 5s and 5m to 5v shows that the remaining differences did not change with practice: matching real versus AR targets, 3.2% (Fig. 5i) versus 3.0% (Fig. 5s), and matching versus reaching for AR targets, 2.7% (Fig. 5m) versus 2.9% (Fig. 5v). Both of these differences involve matching AR targets, which remain increasingly overestimated with increasing distance. The model of the collimating display biasing the observer's vergence angle outward by a constant amount continues to explain this finding in Experiment III. When we calculate Δv for the 10 AR matching observers, who are the same observers from Experiment II, and draw a plot like Fig. 7, the plot again qualitatively looks very similar. For 9 of the 10 observers Δv changes less than 0.22° across the 5 distances, while for the outlying observer it changes 1.16°. The median Δv changes less than 0.08° across the 5 distances.

5 General Discussion and Future Work

The general purpose of these experiments was to compare depth judgments of real and AR objects, using both perceptual matching and blind reaching depth judgment tasks. An additional purpose was to determine how to build an apparatus that allowed us to measure depth judgments with accuracy and precision limits of at most a few millimeters. The work was motivated by AR applications that involve interacting with real and virtual objects at reaching distances.

Overall all three experiments, observers very accurately matched the distance to real targets: the average error was 1.4, 2.5, and 2.7 mm across the three experiments (Fig. 4a, e, i). This is close to the accuracy limit of 1 mm or less required for AR to be useful for image-guided surgery of the brain (Edwards et al. [9]). By definition, a real target is presented with completely consistent depth cues, and therefore these results suggest that as the AR community continues to improve the accuracy of the depth cues presented by AR displays, virtual targets could eventually be matched with similar accuracy. In addition, note that in all three experiments the pointing object was at least several millimeters wide, which limited accuracy. Future experiments should determine if a pointing object that comes to a smaller point results in even more accurate matches.

Also over all three experiments, observers systematically overestimated the matches of AR targets, ranging from 0.5 cm at near distances to 4.0 cm at far distances (Fig. 4b, f, j). In all cases, these results are fit with a model that suggests the collimating optics of the AR display cause the eyes' vergence angle to rotate outward by a constant amount. For AR practitioners, when a collimating AR display is used for applications in reaching space, these results give accuracy limits for AR interaction methods that involve matching-type tasks. These results also strongly suggest using an AR display with an adjustable focus, or at least an AR display with an accommodative demand that more closely matches reaching space distances. The results more generally suggest that an AR display's optical quality and ability to present correct optical depth cues limit the kind of AR applications that can use the display. However, note that this model needs to be verified in an experiment where the AR display's accommodative demand can be adjusted independently from the display's vergence demand[3].

As discussed in Section 1, blind reaching requires an internal sense of perceived distance, and therefore measures definite distance perception (Bingham and Pagano [4]). Observers initially underestimated reaches to real targets by an average of -3.9 cm (Fig. 4c), but improved to -2.6 cm (Fig. 4g) when they could directly point with their finger. They further improved to -0.8 cm (Fig. 4k) after an intervention of matching, which provided corrective visual feedback.

The corresponding accuracies when reaching for AR targets were -4.6 cm (Fig. 4d), $+1.6$ cm (Fig. 4h), and $+1.4$ cm (Fig. 4m). However, the systematic overestimation of AR matches complicates the interpretation of the AR reaching results, since these reaches were presumably made when the vergence angle of the eyes was rotated outward farther than it was when reaching for real targets. This calls for verifying the results in a subsequent experiment using consistent accommodative and vergence depth cues for AR targets.

Finally, Bingham et al. [5] demonstrated that haptic feedback could effectively calibrate blind reaching. When reaching for AR objects, this suggests a proprioceptive calibration method where the observer feels a haptic bump when pointing to the correct depth location. If verified by future experiments, this could prove to be an effective method for calibrating reaches in AR applications.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under awards IIS-0713609, IIS-1018413, and IIS-1320909, to J. E. Swan II.

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 *Proc. of the ACM Symp. on Applied Perception (SAP)*, pages 71–78, July 2012.

[2] N. H. Anderson. *Foundations of Information Integration Theory*. Academic Press, 1981.

[3] 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. *J. of Exp. Psychology: Human Perception and Performance*, 27(6):1314–1334, 2001.

[4] G. P. Bingham and C. C. Pagano. The necessity of a perception-action approach to definite distance perception: Monocular distance perception to guide reaching. *J. of Exp. Psychology: Human Perception and Performance*, 24(1):145–168, 1998.

[5] G. P. Bingham, F. Zaal, D. Robin, and J. A. Shull. Distortions in definite distance and shape perception as measured by reaching without and with haptic feedback. *J. of Exp. Psychology: Human Perception and Performance*, 26(4):1436–1460, 2000.

[6] J. Cohen, P. Cohen, S. G. West, and L. S. Aiken. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 3rd edition, 2003.

[7] D. Curtis, D. Mizell, P. Gruenbaum, and A. Janin. Several devils in the details: Making an AR application work in the airplane factory. In *Proc. of Intern. Workshop on Augmented Reality (IWAR)*, pages 47–60, 1998.

[8] J. E. Cutting and P. M. Vishton. Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In W. Epstein and S. Rogers, editors, *Handbook of Perception and Cognition: Perception of Space and Motion*, volume 5, pages 69–117. Academic Press, 1995.

[9] P. J. Edwards, A. P. King, C. R. Maurer Jr., D. A. de Cunha, D. J. Hawkes, D. L. G. Hill, R. P. Gaston, M. R. Fenlon, A. Jusczyzck, A. J. Strong, C. L. Chandler, and M. J. Gleeson. Design and evaluation of a system for microscope-assisted guided interventions (MAGI). *IEEE Trans. on Medical Imaging*, 19(11):1082–1093, 2000.

[10] S. R. Ellis and B. M. Menges. Localization of virtual objects in the near visual field. *Human Factors*, 40(3):415–431, 1998.

[11] J. M. Foley. Effect of distance information and range on two indices of visually perceived distance. *Perception*, 6:449–460, 1977.

[12] S. J. Henderson and S. Feiner. Evaluating the benefits of augmented reality for task localization in maintenance of an armored personnel carrier turret. In *Intern. Symp. on Mixed and Augmented Reality (ISMAR)*, pages 135–144, 2009.

[13] M. Kersten-Oertel, P. Jannin, and D. L. Collins. The state of the art of visualization in mixed reality image guided surgery. *Comput. Medical Imaging and Graphics*, 37(2):98–112, 2013.

[14] W. S. Levine. *The Control Handbook*. CRC-Press, 1st edition, 1996.

[15] J. W. McCandless, S. R. Ellis, and B. D. Adelstein. Localization of a time-delayed, monocular virtual object superimposed on a real environment. *Presence: Teleoperators and Virtual Environments*, 9(1):15–24, 2000.

[16] M. Mon-Williams and J. R. Tresilian. Some recent studies on the extraretinal contribution to distance perception. *Perception*, pages 167–181, 1999.

[17] M. Mon-Williams and J. R. Tresilian. Ordinal depth information from accommodation? *Ergonomics*, 43(3):391–404, 2000.

[18] M. Mon-Williams, J. P. Wann, M. Jenkinson, and K. Rushton. Synaesthesia in the normal limb. *Proc. of the Royal Society, Series B*, 264:1007–1010, Jul 1997.

[19] A. Naceri, R. Chellali, and T. Hoinville. Depth perception within peripersonal space using head mounted display. *Presence: Teleoperators and Virtual Environments*, 20(3):254–272, 2011.

[20] 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 Trans. on Applied Perception*, 8(3):1–19, 2011.

[21] C. C. Pagano and R. W. Isenhower. Expectation affects verbal judgments but not reaches to visually perceived egocentric distances. *Psychonomic Bulletin & Review*, 15(2):437–442, 2008.

[22] J. Paillard and M. Brouchon. Active and passive movements in the calibration of position sense. In S. J. Freedman, editor, *The Neuropsychology of Spatially Oriented Behavior*, pages 37–55. Dorsey Press, 1968.

[23] E. J. Pedhazur. *Multiple Regression in Behavioral Research*. Holt, Rinehart and Winston, 2nd edition, 1982.

[24] C. Prablanc, J. F. Echallier, E. Komilis, and M. Jeannerod. Optimal response of eye and hand motor systems in pointing at a visual target. *Biological Cybernetics*, 35(2):113–124, 1979.

[25] J. P. Rolland, W. Gibson, and D. Ariely. Towards quantifying depth and size perception in virtual environments. *Presence: Teleoperators and Virtual Environments*, 4(3):24–49, 1995.

[26] J. P. Rolland, C. Meyer, K. Arthur, and E. Rinalducci. Method of adjustment versus method of constant stimuli in the quantification of accuracy and precision of rendered depth in helmet-mounted displays. *Presence: Teleoperators and Virtual Environments*, 11(6):610–625, 2002.

[27] G. Singh. *Near-Field Depth Perception in Optical See-Through Augmented Reality*. PhD thesis, Mississippi State University, Aug. 2013.

[28] G. Singh, J. E. Swan II, J. A. Jones, and S. R. Ellis. Depth judgment measures and occluding surfaces in near-field augmented reality. In *ACM Symp. on Applied Perception in Graphics and Visualization (APGV)*, pages 149–156, 2010.

[29] G. Singh, J. E. Swan II, J. A. Jones, and S. R. Ellis. Depth judgment tasks and environments in near-field augmented reality. In *Proc. of IEEE Virtual Reality (VR)*, pages 241–242, March 2011.

[30] G. Singh, J. E. Swan II, J. A. Jones, and S. R. Ellis. Depth judgments by reaching and matching in near-field augmented reality. In *Proc. of IEEE Virtual Reality (VR)*, pages 165–166, March 2012.

[31] R. J. van Beers, A. C. Sittig, and J. J. Denier van der Gon. How humans combine simultaneous proprioceptive and visual position information. *Experimental Brain Research*, 111(2):253–61, 1996.

[32] J. P. Wann. The integrity of visual-proprioceptive mapping in cerebral palsy. *Neuropsychologia*, 29:1095–1106, 1991.

[33] J. P. Wann, S. Rushton, and M. Mon-Williams. Natural problems for stereoscopic depth perception in virtual environments. *Vision Research*, 35(19):2731–2736, 1995.

[3] In Singh's dissertation [27], the model has been verified in such an experiment.

APPENDIX

A.1 Multiple Regression Analysis

Tables 1 and 2 give complete statistical details for each multiple regression performed in this paper. These statistical procedures are described in detail by Pedhazur [23], chapter 12.

Each row of the table gives the details for a single multiple regression procedure. The *Panel* column indicates the panel in Fig. 5 that the row describes. Within the row, the *Tested Fits* indicates which two linear fits are statistically compared. *Calculated Regressions* list three multiple regression equations, where judged distance (y) is regressed on actual target distance (x) as well as one of three categorical variables: environment (e: real, AR), judgment (j: match, reach), or experiment (p: I, II, III). The categorical variable is always coded as ±1 to represent the two different levels. These regressions yield the values $R^2{}_1, R^2{}_2$, and $R^2{}_3$, which give the percentage of the variation of the points in the panel described by each regression equation.

Next, we calculate the *Slope Test*, which indicates whether the slopes of the tested linear fits significantly differ. Specifically, this test determines if $R^2{}_1 - R^2{}_2$ is significant, or if $R^2{}_1$ describes a significantly greater amount of variation than $R^2{}_2$. Note that the regression equations differ in the interaction term ex, jx, or px, the interaction between the categorical variable and actual target distance. If this test indicates a significant difference, no further testing is done: the two fits are reported as the best overall description of the panel, and $R^2{}_1$ is reported as the overall R^2 for the panel.

If the slope test is not significant, we next calculate the *Intercept Test*. This test sets the slopes of the two tested fits to a common slope—because they do not differ—and then determines if the intercepts significantly differ. Specifically, this test determines if $R^2{}_2 - R^2{}_3$ is significant, or if $R^2{}_2$ describes a significantly greater amount of variation than $R^2{}_3$. Note that the regression equations differ in the categorical term e, j, or p. If this test indicates a significant difference, the two fits with the common slope are reported as the best overall description of the panel, and $R^2{}_2$ is reported as the overall R^2 for the panel.

If the intercept test is not significant, then the best overall description of the panel is the simple regression $y = x$. This regression equation is reported, and $R^2{}_3$ is reported as the overall R^2 for the panel.

Finally, in all cases $R^2{}_1 - R^2{}_3$ is reported as the overall effect size for the panel—the percentage of variation explained by including the categorical variable in the regression.

Table 1. Multiple Regression Analysis

Panel	Tested Fits	Calculated Regressions		Slope Test	Intercept Test
Experiment I					
Fig. 5a	match real: $y = 0.997x + 0.28$ match AR: $y = 1.090x - 2.49$	$y = x + e + ex$ $y = x + e$ $y = x$	$R^2{}_1 = 98.7\%$ $R^2{}_2 = 98.5\%$ $R^2{}_3 = 97.6\%$	$F_{1,96} = 14.0, p < 0.001^{***}$	
Fig. 5b	reach real: $y = 1.084x - 7.49$ reach AR: $y = 1.044x - 6.48$	$y = x + e + ex$ $y = x + e$ $y = x$	$R^2{}_1 = 77.9\%$ $R^2{}_2 = 77.8\%$ $R^2{}_3 = 77.6\%$	$F_{1,96} = 0.12, p = 0.73$	$F_{1,97} = 1.09, p = 0.30$
Fig. 5c	match real: $y = 0.997x + 0.28$ reach real: $y = 1.084x - 7.49$	$y = x + j + jx$ $y = x + j$ $y = x$	$R^2{}_1 = 86.9\%$ $R^2{}_2 = 86.7\%$ $R^2{}_3 = 77.4\%$	$F_{1,96} = 1.01, p = 0.32$	$F_{1,97} = 68.1, p < 0.001^{***}$
Fig. 5d	match AR: $y = 1.090x - 2.49$ reach AR: $y = 1.044x - 6.48$	$y = x + j + jx$ $y = x + j$ $y = x$	$R^2{}_1 = 90.1\%$ $R^2{}_2 = 90.1\%$ $R^2{}_3 = 72.7\%$	$F_{1,96} = 0.32, p = 0.57$	$F_{1,97} = 170.5, p < 0.001^{***}$
Experiment I vs II					
Fig. 5e	match real Ex I: $y = 0.997x + 0.28$ match real Ex II: $y = 1.020x - 0.56$	$y = x + p + px$ $y = x + p$ $y = x$	$R^2{}_1 = 99.8\%$ $R^2{}_2 = 99.8\%$ $R^2{}_3 = 99.8\%$	Note 1	
Fig. 5f	match AR Ex I: $y = 1.090x - 2.49$ match AR Ex II: $y = 1.100x - 1.09$	$y = x + p + px$ $y = x + p$ $y = x$	$R^2{}_1 = 96.9\%$ $R^2{}_2 = 96.9\%$ $R^2{}_3 = 95.4\%$	$F_{1,96} = 0.07, p = 0.80$	$F_{1,97} = 48.7, p < 0.001^{***}$
Fig. 5g	reach real Ex I: $y = 1.084x - 7.49$ reach real Ex II: $y = 1.005x - 2.76$	$y = x + p + px$ $y = x + p$ $y = x$	$R^2{}_1 = 84.6\%$ $R^2{}_2 = 84.5\%$ $R^2{}_3 = 83.5\%$	$F_{1,96} = 0.70, p = 0.41$	$F_{1,97} = 5.72, p = 0.019^{*}$
Fig. 5h	reach AR Ex I: $y = 1.044x + 6.48$ reach AR Ex II: $y = 1.155x - 5.97$	$y = x + p + px$ $y = x + p$ $y = x$	$R^2{}_1 = 84.6\%$ $R^2{}_2 = 84.4\%$ $R^2{}_3 = 74.1\%$	$F_{1,96} = 1.24, p = 0.27$	$F_{1,97} = 64.3, p < 0.001^{***}$

Note 1: The overall effect size of $R^2{}_1 - R^2{}_3 = 0.023\%$ is too small for any statistical differences to be meaningful.

Table 2. Multiple Regression Analysis (continued)

Panel	Tested Fits	Calculated Regressions		Slope Test	Intercept Test
Experiment II					
Fig. 5i	match real: $y = 1.020x - 0.56$	$y = x + e + ex$	${R^2}_1 = 97.9\%$	$F_{1,96} = 6.38, p = 0.013^{**}$	
	match AR: $y = 1.100x - 1.09$	$y = x + e$	${R^2}_2 = 97.8\%$		
		$y = x$	${R^2}_3 = 94.8\%$		
Fig. 5j	reach real: $y = 1.005x - 2.76$	$y = x + e + ex$	${R^2}_1 = 89.4\%$	$F_{1,96} = 3.62, p = 0.060^{*}$	
	reach AR: $y = 1.155x - 5.97$	$y = x + e$	${R^2}_2 = 89.0\%$		
		$y = x$	${R^2}_3 = 86.6\%$		
Fig. 5k	match real: $y = 1.020x - 0.56$	$y = x + j + jx$	${R^2}_1 = 95.3\%$	$F_{1,96} = 0.10, p = 0.76$	$F_{1,97} = 63.9, p < 0.001^{***}$
	reach real: $y = 1.005x - 2.76$	$y = x + j$	${R^2}_2 = 95.3\%$		
		$y = x$	${R^2}_3 = 92.2\%$		
Fig. 5m	match AR: $y = 1.100x - 1.09$	$y = x + j + jx$	${R^2}_1 = 91.7\%$	$F_{1,96} = 0.61, p = 0.44$	$F_{1,97} = 30.1, p < 0.001^{***}$
	reach AR: $y = 1.155x - 5.97$	$y = x + j$	${R^2}_2 = 91.6\%$		
		$y = x$	${R^2}_3 = 89.0\%$		
Experiment II vs III					
Fig. 5n	match real Ex II: $y = 1.020x - 0.56$	$y = x + p + px$	${R^2}_1 = 99.8\%$	Note 2	
	match real Ex III: $y = 1.007x - 0.00$	$y = x + p$	${R^2}_2 = 99.8\%$		
		$y = x$	${R^2}_3 = 99.8\%$		
Fig. 5p	match AR Ex II: $y = 1.100x - 1.09$	$y = x + p + px$	${R^2}_1 = 96.8\%$	Note 3	
	match AR Ex III: $y = 1.064x + 0.27$	$y = x + p$	${R^2}_2 = 96.7\%$		
		$y = x$	${R^2}_3 = 96.7\%$		
Fig. 5q	reach real Ex II: $y = 1.005x - 2.76$	$y = x + p + px$	${R^2}_1 = 93.0\%$	$F_{1,96} = 0.29, p = 0.59$	$F_{1,97} = 18.9, p < 0.001^{***}$
	reach real Ex III: $y = 1.036x - 2.12$	$y = x + p$	${R^2}_2 = 93.0\%$		
		$y = x$	${R^2}_3 = 91.7\%$		
Fig. 5r	reach AR Ex II: $y = 1.155x - 5.97$	$y = x + p + px$	${R^2}_1 = 88.7\%$	$F_{1,96} = 1.49, p = 0.23$	$F_{1,97} = 0.0018, p = 0.97$
	reach AR Ex III: $y = 1.057x - 2.12$	$y = x + p$	${R^2}_2 = 88.5\%$		
		$y = x$	${R^2}_3 = 88.5\%$		
Experiment III					
Fig. 5s	match real: $y = 1.007x - 0.00$	$y = x + e + ex$	${R^2}_1 = 98.4\%$	$F_{1,96} = 4.38, p = 0.039^{*}$	
	match AR: $y = 1.064x + 0.27$	$y = x + e$	${R^2}_2 = 98.3\%$		
		$y = x$	${R^2}_3 = 95.3\%$		
Fig. 5t	reach real: $y = 1.036x - 2.12$	$y = x + e + ex$	${R^2}_1 = 93.3\%$	$F_{1,96} = 0.14, p = 0.71$	$F_{1,97} = 4.07, p = 0.047^{*}$
	reach AR: $y = 1.057x - 2.12$	$y = x + e$	${R^2}_2 = 93.3\%$		
		$y = x$	${R^2}_3 = 93.0\%$		
Fig. 5u	match real: $y = 1.007x - 0.00$	$y = x + j + jx$	${R^2}_1 = 97.6\%$	$F_{1,96} = 0.74, p = 0.39$	$F_{1,97} = 16.3, p < 0.001^{***}$
	reach real: $y = 1.036x - 2.12$	$y = x + j$	${R^2}_2 = 97.6\%$		
		$y = x$	${R^2}_3 = 97.1\%$		
Fig. 5v	match AR: $y = 1.064x + 0.27$	$y = x + j + jx$	${R^2}_1 = 94.1\%$	$F_{1,96} = 0.016, p = 0.90$	$F_{1,97} = 47.8, p < 0.001^{***}$
	reach AR: $y = 1.057x - 2.12$	$y = x + j$	${R^2}_2 = 94.1\%$		
		$y = x$	${R^2}_3 = 91.2\%$		

Note 2: The overall effect size of ${R^2}_1 - {R^2}_3 = 0.0045\%$ is too small for any statistical differences to be meaningful.

Note 3: The overall effect size of ${R^2}_1 - {R^2}_3 = 0.029\%$ is too small for any statistical differences to be meaningful.