Understanding the Impact of Visual and Kinematic Information on the Perception of Physicality Errors

GOKSU YAMAC, Meta Reality Labs Research, USA; Trinity College Dublin, Ireland
CAROL O'SULLIVAN, Trinity College Dublin, Ireland
MICHAEL NEFF, Meta Reality Labs Research, USA; University of California, Davis, USA

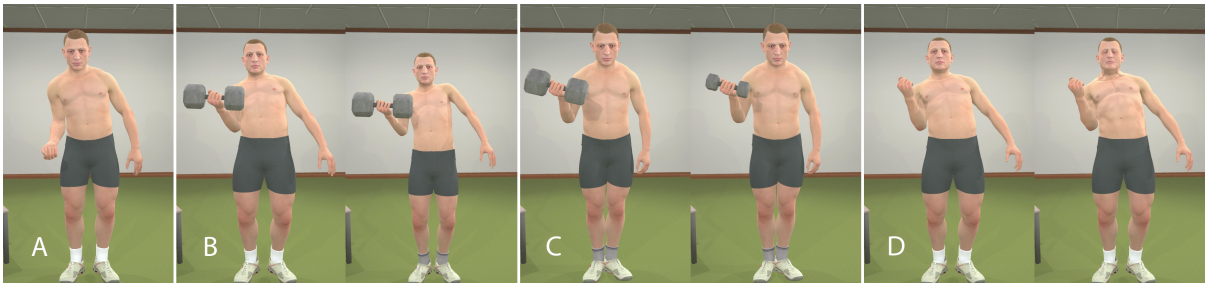

Fig. 1. This study used dumbbell lifts to explore the impact of A) kinematics (Exps. 1 and 5), B) body shape (Exp. 2), C) object size (Exp. 3), and D) muscle flexion (Exps. 4 and 5) on perceived effort, object weight and motion realism.

Errors that arise due to a mismatch in the dynamics of a person's motion and the visualized movements of their avatar in virtual reality are termed 'physicality errors' to distinguish them from simple physical errors, such as footskate. Physicality errors involve plausible motions, but with dynamic inconsistencies. Even with perfect tracking and ideal virtual worlds, such errors are inevitable in virtual reality whenever a person adopts an avatar that does not match their own proportions or lifts a virtual object that appears heavier than the movement of their hand. This study investigates people's sensitivity to physicality errors in order to understand when they are likely to be noticeable and need to be mitigated. It uses a simple, well-understood exercise of a dumbbell lift to explore the impact of motion kinematics and varied sources of visual information, including changing body size, changing the size of manipulated objects, and displaying muscular strain. Results suggest that kinematic (motion) information has a dominant impact on perception of effort, but visual information, particularly the visual size of the lifted object, has a strong impact on perceived weight. This can lead to perceptual mismatches which reduce perceived naturalness. Small errors may not be noticeable, but large errors reduce naturalness. Further results are discussed, which inform the requirements for animation algorithms.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: VR, motion perception

Authors' addresses: Goksu Yamac, Meta Reality Labs Research, Sausalito, USA; and Trinity College Dublin, Dublin, Ireland; Carol O'Sullivan, Trinity College Dublin, Dublin, Ireland; Michael Neff, Meta Reality Labs Research, Sausalito, USA; and University of California, Davis, USA.

1544-3558/2023/1-ART1 $15.00
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Goksu Yamac, Carol O'Sullivan, and Michael Neff. 2023. Understanding the Impact of Visual and Kinematic Information on the Perception of Physicality Errors. *ACM Trans. Appl. Percept.* 1, 1, Article 1 (January 2023), 31 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

You can be anyone or anything you want in virtual reality (VR), or so the story goes. But what if the proportions of an avatar do not match the user? Or if someone is visualized moving larger objects than the controllers they are moving in the real world? Given that the dominant approach to drive people's avatars in VR is to track their movements in the real world, mismatches between people's dynamics and those of their avatars will result in perceptible errors. Such 'physicality errors' are an inevitable part of many embodied virtual reality [43, 44] or metaverse scenarios as people wish to use their motion to naturally drive a range of avatars, without restriction to match their own appearance, and to do so across a spectrum of activities. This is not only the case for their own self-avatars, but also for the avatars with whom they share a virtual environment. It is, perhaps, even more important that the avatars of virtual companions are believable for a sense of co-presence [16] to be perceived, as their bodies and motions will be fully visible to the user. It is thus critical to understand whether these errors are noticeable and likely to impact user experience and what types of ameliorations may be necessary.

In this paper, we investigate the impact of errors in physicality on people observing the avatars of others. We employ the scenario of lifting dumbbells, a simple physical activity that is well understood by the general public and can easily be performed with different levels of resistance. Through five experiments, we seek answers to the following questions about the perceptual impact of both the kinematic signal (i.e., the motion) and visual signals including the size of the avatar, the size of the lifted object and the presence of muscle deformations:

(1) Baseline: How accurately can people perceive effort and infer weight from motion kinematics on a meshed character in VR?
(2) Body Shape: If people are visualized with avatars that have different proportions and mass than their own, how sensitive are observers to the resulting errors in their motion dynamics?
(3) Dumbbell Size: How sensitive are people to visualizations that show an avatar moving a different mass than what they actually moved?
(4) Strain Deformations: Can the display of muscle deformations, including facial strain, shift people's perception of effort and inference of weight in lifts?
(5) Discrimination: Can people distinguish between a zero-weight lift, as would be performed by a VR user in standard use cases, and an accurate lift for various weight dumbbells? Can displays of muscular sensitivity reduce sensitivity to these mismatches?

In the experiments, participants watch a series of animations in virtual reality (VR) of avatars lifting dumbbells and either estimate the effort and weight of the action, sometimes including naturalness ratings, or try to determine which of the two animations accurately reflects a visualized lift. Four different avatars are driven by the recorded motion of two average-strength male lifters and two strong male lifters. Avatar bodies are matched to the lifters, with the strong lifters being taller and heavier. This allows multiple combinations of body type, motion kinematics, and displayed weight to be shown. A blendshape model is added to support realistic muscle deformations. Blendshape models average different versions of the character mesh that are sculpted to show particular features, in this work used for showing muscle bulges.

Results suggest that: (1) when presented with kinematic data – an avatar lifting an unseen weight – people make largely consistent judgments of effort and weight, modulo modest scaling based on the size of the avatar body. This scaling is less than the actual range of strength variation seen in our lifters; (2) when visual information on the weight of the lift is introduced – a dumbbell is shown – weight and effort estimates diverge. Effort estimates

are still largely driven by kinematics, but weight estimates are changed substantially to be consistent with the new visual information; (3) changes in body size and in dumbbell size also have a larger impact on weight estimates than effort estimates. This can create discordant signals where effort and weight estimates no longer match when the visualized motion varies from the actual motion; (4) adding displays of muscle strain impacts both effort and weight estimates in the absence of visualized dumbbells; (5) when dumbbells are present, the strain signal can make incorrect motion kinematics less noticeable when the strain is appropriate for the visualized dumbbell weight, but the clips with muscular deformation become more distinguishable when the muscular deformation is excessive for the visualized dumbbell. In general, naturalness ratings tend to decline for large mismatches. Within a narrower range of lower-effort motions, however, people are less sensitive to inconsistencies and there may be room for variation without negative consequences. Taken together, these results shed light on how people interpret animated motion and help illuminate the animation requirements that will be necessary to support envisioned metaverse scenarios.

2 BACKGROUND

The perception of weight and effort has been studied in fields such as Computer Graphics, VR, perception and psychophysics. In this section, we provide an overview of the most relevant literature.

2.1 Character Animation and Virtual Reality

The perception of animated character motion has been studied in the field of Computer Graphics, e.g., the perception of child vs. adult motion [25] and the perception of sex from walking motion [32]. It has been shown that anomalies in facial motion were more disturbing than body motion [21] and that motion attractiveness affects users' comfort level in proximity to avatars [48]. Sensitivity to limb length changes also varies based on factors such as motion speed and attention [20], and viewers were found to be accurate at detecting force errors when one animated character pushes another [22]. Concerning interactions in VR, it has been shown that, although performance was better when an animated hand was accurately tracked and allowed to penetrate a grasped object, users actually preferred it when the hand motion was adjusted to avoid interpenetrations [12].

Related work has examined how motion frameworks such as Laban Movement Analysis (LMA) can be used to synthesize [14, 45] or perceive [30] variation in motion, including force as manifested in the Weight quality of LMA Effort. In our work, we asked the actors to use the force required for the functional motion of lifting the weight and not to intentionally manipulate their effort beyond this. Expressive animation tools may prove useful in addressing the perceptual errors that surfaced in our studies.

Perception of weight in VR has also been extensively studied and still remains an open challenge [31]. It has been shown that people underestimated weight by 10-20% when they were asked to adjust an avatar to match their body proportions [47]. As haptic hardware is often inaccessible to incorporate in VR applications, pseudo-haptics cues, i.e., visual information that is associated with haptic sensations, can be used to trick users by manipulating the visual information provided in the Virtual Environment (VE) [34]. These approaches include the manipulation of self-avatar animation [26], control-display ratio [39], and tracking offsets [35].

Most relevant to our work, Kenny et al. [28, 29] measured sensitivity to mismatches between avatar body size and motion for pushing, lifting, and throwing actions. Two weight groups of male actors performed each action and participants viewed them on avatars that matched or mismatched the actor's size. Although the detection rate of these mismatches was low and the ratings of naturalness remained unchanged, they did change the interpretation of the physical activity, e.g., heavier avatars were perceived as lifting heavier objects and light avatars with a heavy actor's motion were perceived as pushing lighter than vice versa. Notably, the experiment did not visualize the objects that the avatars were interacting with, which allowed observers to interpret the

change in stimuli as reflecting a change in object properties (e.g., a heavier imagined object). As such freedom will not be the case in most practical scenarios, our study includes experiments that visualize the manipulated object.

2.2 Perception of Human Motion Dynamics

Much of the research on the perception of dynamics from human motion has employed point-light displays, i.e., videos that show only points, normally at joint centers. Participants were consistently able to estimate the weight of a lifted box [8, 36, 40–42] or dumbbell [7]. However, the accuracy of estimates varied widely, from very high correlation between actual and perceived weight [36, 42], to much lower accuracy [41, 42]. It was found accuracy was lower for weights below 30lbs [8] and that light weights were overestimated, whereas heavy ones were underestimated [41]. The lift phase of a lift and carry motion was found to be sufficient to make a judgment [40]. Variations of the study design that improved estimates include: showing a reference lift of a specified weight [7, 36, 42] (although this may have introduced systematic errors [8]), not telling the actors the weight of the box [36], participants performing their own max lift to gain haptic experience [7, 42], knowing the size of the lifter [8], rating a single lifter at a time [41], and using average strength actors [42]. When compared with point-light displays, showing the full video of the actor resulted in the highest accuracy [18], while in another study with virtual and real conditions [23], the additional visual data available for the real lifts improved performance, which was most likely due to muscle contractions that were lacking in the virtual condition. These results suggest that both motion and human form are important for judging actions [9]. We also explore this question by adding muscle strain cues.

Kinematic changes that correlate with weight have been observed, in particular object velocity decreases [7, 18, 40, 41], although not for all actors [41]. Dwell time at the start of the lift, hip angle [40], and max trunk velocity [41] also vary. Manipulating kinematic patterns can change weight perceptions [40, 42], potentially because the consequences of gravitational attraction on objects has been incorporated into mental representations [24]. However, it was insufficient to show only the motion of the box [18, 42] or one degree of freedom movement of the elbow for dumbbell lifts [7]. Neck strain in a filmed pilot study was mentioned as a cue by participants [36], which we simulate in ours. Finally, there is a response in the motor cortex consistent with force cues from kinematics and hand contraction state, which manifests visually in the color and deformation of the hand [3]. It has also been suggested that the motor cortex may be active in motion perception [19].

Providing information about the size of the box did not confound the perception of lifted weight for point-light displays [18], but a small box was judged to weigh less with a video presentation. Although people scaled their motor functions appropriately when lifting a small box [17], they later reported that the weight of the smaller box was heavier. This suggests that Visual cues are integrated into the programming of manipulative forces during precision grip. When participants were asked viewers to estimate both effort and weight [41], they were more accurate at estimating the effort of lifters. However, if they knew a lifter's size and weight beforehand, their estimates of weight and effort were comparable. An unexpected result showed participants judged the weight of a heavier object lifted by a stronger and larger actor as being lighter than a weight lifted by a normal, weaker lifter. Since ordinal judgments were correct, they suggest observers may be more attuned to effort than weight [42].

The field of psychophysics involves the study of how people perceive physical phenomena. In general, there is not a linear but rather a power law relationship between the perception of a phenomenon and the underlying physical quantity [10], e.g., when asked to judge the heaviness of objects, people may feel that an object with 70% of the actual mass is half as heavy, rather than one with 50% of the mass. Such judgments are affected by factors beyond the mass alone. The well-studied size-weight illusion (SWI) shows that when people are asked to lift and then estimate the heaviness of two objects that are actually the same weight, but differ in size, they will estimate the smaller object as heavier [11, 15, 27, 46]. Other factors like the object's material [11, 38], color [27] and shape [37] will also impact the perceived heaviness, although size has a larger effect [38]. The SWI is

Table 1. Summary of the Experiments

Experiment #	Name	Deformations	Body	Dumbbell	Medium
1	Baseline	No	Lifter Matched	Not Shown	VR
2	Body Shape	No	All Four Lifters	Motion Matched	VR
3	Dumbbell Size	No	Lifter Matched	All weights for lifter	VR
4	Muscle Strain	Yes	Lifter Matched	Not Shown	VR
5	Discrimination w/Strain	Yes	Lifter Matched	Matched and mismatched	VR

still not fully understood, but there is evidence that both bottom-up physical factors, such as the object's inertia tensor [4] or perceived density [46], and top-down conceptual priors play a role [11, 15, 38].

3 EXPERIMENTAL METHODS

The study was organized into a sequence of experiments (Table 1) that explored various signals of physicality (kinematics, avatar representation, object representation, and muscle deformations, respectively). All experiments followed the same basic design, which will be described here.

3.1 Experimental Design

During each experiment, participants were recruited for a single session in which they completed a survey in VR. After an orientation on using a VR headset and the controls employed in the survey, they were given brief instructions on the experiment in the headset. They were then shown a range of clips that indicated the type of variation they might see in the experiment. This enables them to start the experiment knowing the stimuli range. After this, they saw a randomized sequence of short motion clips that showed a single person lifting a dumbbell in a simple room. After each clip, participants were asked to rate the clip using a floating 2-D slider in terms of the perceived effort of the lifter, from 0 to 100% where 100% represented their maximum effort; their estimation of the weight that was lifted, from 0 to 100 lbs (they were told that weights may not span the entire scale) and, for Exps. 2-4, the naturalness of the motion by rating the statement "The motion in this clip appears natural" on a 7-point Likert scale ranging from Strongly Disagree to Strongly Agree. At the end of the experiment, they participated in a brief exit interview and debrief.

Apparatus. The experiment environment was developed in Unity and presented on an Oculus Quest 2 headset. This headset has a resolution of 1832×1920 pixels per eye with around 90° horizontal and vertical Field of View. The virtual environment (VE) was a standard room with a door, several lights, and several power outlets on the walls. It contained a chair that acted as a scale marker. An X was placed on the floor to make sure that every participant observed the stimuli from a same distance of about 1.5m (Figure 2). Participants interacted with the VE using an Oculus Touch controller. A visible ray coming out from the virtual representation of the controller was used to control interface widgets.

Stimuli. The lifters that provided the source motion for the study were recruited through online advertisements to a pool of actors. The actors submitted their maximum dumbbell curl and a short video of themselves performing a lift. The selection was based on establishing two sets of lifters, two "strong" lifters and two "average" that had clearly differentiated maximum lifts. This allowed us to examine the impact of both body size and lifter strength on observations of dynamics. Details on the selected lifters are summarized in Table 2.

To establish their actual maximum lift, all lifters came to the studio at least a day before the actual motion recording. They started by lifting what they estimated would be their maximum lift. The weight was gradually increased as long as they could complete three repetitions of the lift. They were given time to rest between each

Fig. 2. Participants stood on an X on the floor in front of the lifter throughout the experiment

Lifter	Max Dumbbell Weight	75%	50%	25%	Age	Weight (lbs)	Height	Dominant Hand
Avg. Male 1 (AA1)	27 lbs	20.25	13.5	6.75	24	139	5ft. 9in.	R
Avg. Male 2 (AA2)	35 lbs	26.25	17.5	8.75	38	191	5ft. 9in.	R
Strong Male 1 (SA1)	60 lbs	45	30	15	30	237	6ft. 1in.	R
Strong Male 2 (SA2)	60 lbs	45	30	15	27	231	6ft. 2in.	R

Table 2. Performers in weight lifting task.

set. The process stopped when they decided they had reached their maximum, which was recorded for use during the motion capture session.

The motion of each lifter was recorded using a Vicon, marker-based optical motion capture system featuring with forty 16MP cameras. A standard marker set was used consisting of 67 markers. Each lifter performed lifts at 0 (holding nothing), 25, 50, 75, and 100% of their maximum lift. They performed a single set of three lifts at a given weight before resting. The order of weights was randomized for each lifter to avoid any fatigue patterns. Two sets of lifts were recorded for each weight for each lifter. Lifters were instructed to do the lift as they would normally to meet the functional requirements of the action, with no effort to accentuate or minimize the appearance of applied force. The motion capture data were solved such that the captured marker data were used to fit a skeleton that matches the actor's body. This was performed using a custom solver that minimized the root mean square error over the marker set. In the preparation of the set of virtual dumbbells used in the experiments, we first created a virtual replica of a real 10-lb dumbbell. The other dumbbells were created by scaling the volume of the dumbbell ends, assuming fixed density.

Model and Deformations. The avatar model needed to support two research goals: allow variation in body shape and provide plausible muscle deformations. It was beyond the scope of the project to create and validate a full simulation model of human muscle. Instead, we employed an artist-driven approach whereby an artist with twenty years of experience in the visual effects industry generated a model and set of blend shapes to control deformations. The muscle deformations used in Exps. 4 and 5 were designed to show a high level of strain, rather than being tuned to each lift. This allowed us to investigate if strain cues are impactful, but further work would be

Fig. 3. The models for each of the four performers, AA1, AA2, SA1, SA2.

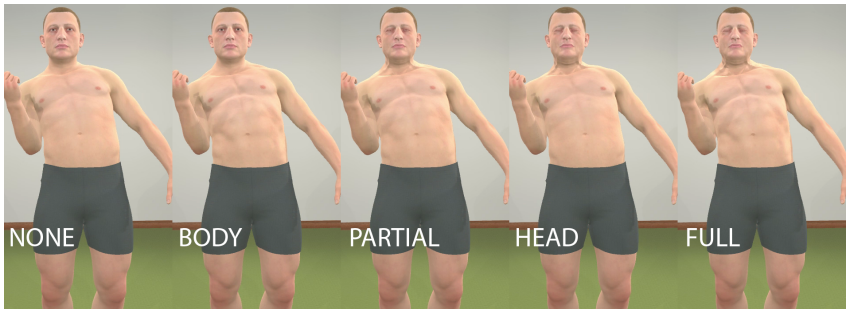

Fig. 4. Muscle deformations shown on avatar SA1 for an intense moment in a lift with the five different deformation levels: A) no deformation, B) BODY, C) PARTIAL, D) HEAD, and E) FULL. See the video for examples of the deformations animated.

required to tune deformations to arbitrary lifts. A full discussion of the model is contained in the supplementary material. The effectiveness of the model deformations for conveying strain was validated (see Appendix).

3.2 Demographics

The number of participants, mean age (SD), and gender data for each experiment are shown in Table 3. Other data were similar across participant pools. The ethnicity was: 81.3% White/Caucasian, 6.7% Black/African/African American, 6% Asian/Asian American, 4.4% Latin/Hispanic, and 3.1% preferred not to say. Experience with VR was some (48.2%), none (35.2%), or more extensive (16.6%), while 62.0% had some experience exercising or weightlifting or did it regularly (31.3%). Most had some experience seeing others lift weights (62.6%), and some regularly (24.2%). Participants were non-overlapping, but Exp. 3 and 5 shared a pool. The stimuli differed, but some learning may have occurred across Exp. 3 and 5.

3.3 Analysis

Analysis was generally performed using linear mixed effect models [5, 6]. These offer a more general approach than ANOVA as they include fixed and random effects, but similarly predict the dependence of a response term (e.g., perceived effort) on one or more factors (e.g., the size of the avatar's body). The participant ID was treated as a random effect since the participant pool is merely a sample of the more general population. Type II Wald chisquare tests were used to evaluate significance within the models. Post-hoc analysis was conducted by calculating pairwise comparisons using estimated marginal means with Tukey correction. Naturalness ratings

Exp.	N	Age	Gender
1	30	37.7 (12.7)	F 50%, M 46.7%, O 3.3%
2	35	35.7 (9.7)	F 48.6%, M 48.6%, O 2.9%
3 & 5	35	33.9 (9.5)	F 45.7%, M 45.7%, O 8.6%
4	35	34.5 (8.3)	F 45.7%, M 51.4%, O 2.9%

Table 3. Participant demographics. Gender category O is a composite of Non-binary/third gender and Other.

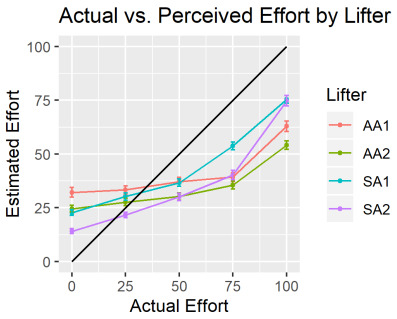

Fig. 5. Exp. 1: Perceived effort by lifter compared to actual effort. The black line indicates perfect performance.

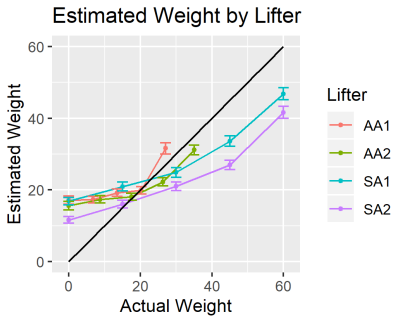

Fig. 6. Exp. 1: Inferred weight by lifter.

were instead fit with a Cumulative Link Model [13], which treats the Likert scores as ordinal data. Exceptions to this analysis scheme will be noted in the relevant sections. The error bars in all plots show standard error. The statistical results are shown in tables or the Appendix.

4 EXP. 1, BASELINE: PERCEPTION OF EFFORT AND INFERRED WEIGHT

The first experiment was designed to provide a baseline measurement of how well people can perceive effort and infer weight from motion kinematics on a meshed character in VR. For this reason, the lifted dumbbells were not shown. This also makes the work more directly comparable with previous work on point-light displays. The displayed body was approximately matched to that of the lifter, as described in Sec. 3.1. Forty clips were used in this experiment (4 lifters x 5 different weights x 2 repetitions). Weights were evenly spaced at 0, 25, 50, 75, and 100% of each person's maximum lift, i.e., they differed per lifter (Table 2). Since previous research [41] has shown that people are more accurate when making estimates for a single lifter at a time, clips were grouped by lifter and randomized within lifter. Participants rated their estimates of both effort and weight for each clip.

Figure 5 shows perceived effort as a function of the actual effort made by the lifter, considering their maximum lift to be 100% effort. Participants' estimates of the weights for the various lifts by lifter are shown in Figure 6. For every lifter, there is a highly significant correlation between estimated effort/weight and actual effort/weight ("highly significant" is used for p-values < .001, where Pearson's product-moment correlation had all p-values less than 1e-14). The correlations are strong for the stronger lifters and medium for the average lifters on both measures: (Effort Pearson's r: $AA1 = .41, AA2 = .48, SA1 = .73, SA2 = .73$; Weight Pearson's r: $AA1 = .36, AA2 = .41, SA1 = .61, SA2 = .64$). These findings suggest that at least to some degree, people are able to infer both weight and effort from lifters' kinematics.

It can also be observed that correlations are somewhat stronger for effort. Using Fisher's Z transform to compare correlations shows that these differences are not significant for the average strength lifters (AA1: $Z = .72, p = .47; AA2 : Z = 1.06, p = .29$), but are significant for the strong lifters (SA1: $Z = 2.68, p = .0074; SA2 :$

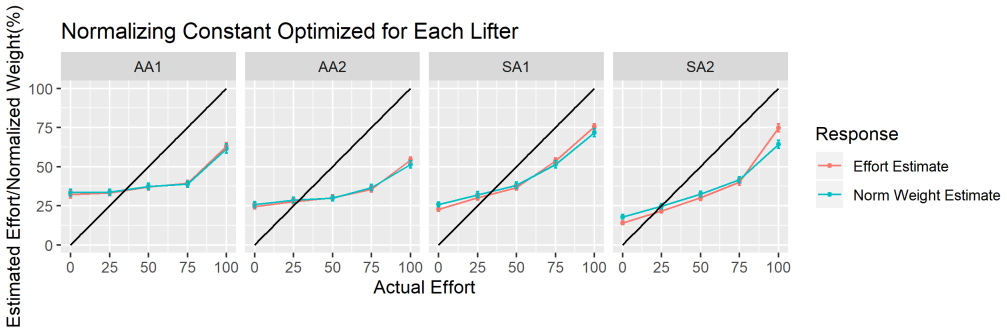

Fig. 7. Exp. 1: Perceived effort and normalized weight where the normalizing constant is optimized for each lifter (AA1: 51.3 lbs, AA2: 60.6 lbs, SA1: 65.3 lbs, SA2: 64.8 lbs).

$Z = 2.08, p = .038$). This suggests that people may be better able to observe effort than weight for the strong lifters, perhaps because effort can be more directly observed and weight needs to be inferred.

Observation of Figures 5 and 6 shows a curvilinear relationship between the actual and estimated values such that the curve is flatter for lower effort/weight and steeper towards the maximum. Mirroring previous analyses (e.g., [41]), a linear mixed effects model was fitted to each lifter for each Effort and Weight. Since the Weight levels differ across lifters, we chose to fit an individual model to each lifter, rather than treating Lifter as an additional factor. In all cases, the independent variable (effort or weight) had a highly significant impact on the estimates (p<.001, see Appendix), reflecting that people adjust their judgments based on both the actual effort and actual weight. Pairwise comparisons were done for each model. These show that the 100% lifts differed significantly from the 75% lifts for both Effort and Weight for all lifters, but lower levels were generally not significantly different for the average lifters (see Appendix). This reflects the more moderate slope in this part of the response curve and implies that observers may be less sensitive to these more subtle variations in kinematics. As with other psychophysical tasks [10], the relationship between the actual weight lifted and the participants' weight estimates based on visual perception can be related by a power function. However, the exponent found here is not consistent across lifters, and variance in ratings is high.

Relationship between Effort and Weight Estimates. In order to better understand the relationship between effort and weight estimates, we can normalize the estimated weight values (i.e. express them as a percentage of some max lift) so that effort and weight can be plotted on the same scale. If we optimize for a normalizing constant for each lifter by minimizing the difference between the normalized weight and estimated effort, we produce the chart in Figure 7 with weights: AA1: 51.3 lbs, AA2: 60.6 lbs, SA1: 65.3 lbs, SA2: 64.8 lbs. There is clearly a strong correspondence between the curves for effort and weight estimates (Pearson's correlation $r = 0.74$), so it may be that people were making a single judgment based on the motion and then scaling that to estimate the other quantity. We will revisit this in the discussion (Sec. 9) when there is more evidence that effort is the quantity estimated. Note that these optimized max weights are more similar than the lifters' actual max lifts.

Discussion

To answer *"How accurately can people perceive effort and infer weight from motion kinematics on a meshed character in VR?"*, while estimates are not perfect, Exp. 1 provides evidence that people can gauge both effort and weight from kinematic signals for characters in virtual reality. This confirms earlier work with point-light displays. The overall correlation between actual and perceived values is consistent with those reported in [41], but below the highest estimates in the literature (e.g., [36]). We did not provide a reference lift with a specified weight in

Table 4. Exp. 2: Effort Ratings (l), Lifted weight ratings (r). "Body" is short for "Body shape."
(Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ')

Motion		χ^2	dof	$p(>\chi^2)$
AA1	Effort	533.2	3	< .0001 ***
	Body	3.97	3	.26
	Effort:Body	14.55	9	.10
AA2	Effort	253.46	3	< .0001 ***
	Body	1.08	3	.78
	Effort:Body	12.64	9	.18
SA1	Effort	915.8	3	< .0001 ***
	Body	63.72	3	< .0001 ***
	Effort:Body	26.10	9	.0020 *
SA2	Effort	723.9	3	< .0001 ***
	Body	6.26	3	.10
	Effort:Body	13.27	9	.15

Lifter		χ^2	dof	$p(>\chi^2)$
AA1	Weight	813.9	3	< .0001 ***
	Body	22.34	3	< .0001 ***
	Weight:Body	14.34	9	.11
AA2	Weight	711.6	3	< .0001 ***
	Body	46.87	3	< .0001 ***
	Weight:Body	11.61	9	.24
SA1	Weight	730.1	3	< .0001 ***
	Body	62.35	3	< .0001 ***
	Weight:Body	11.72	9	.23
SA2	Weight	696.7	3	< .0001 ***
	Body	42.8	3	< .0001 ***
	Weight:Body	6.87	9	.65

any of our experiments. While this has been shown to improve the accuracy of predictions [7, 36, 42], it is not a likely scenario in real VR applications. This may account for why our correlations are somewhat lower than the highest values reported in the literature, although we did always include an indication of lifter size, which is also a useful leveling cue [8].

Notably, people are less sensitive to differences at lower weight or effort levels. The Discrimination experiment (Exp. 5) showed a similar finding. People were not sensitive to the differences between 0 and 25% effort lifts, but they were sensitive to differences between higher effort lifts (somewhere between 50 and 100%, depending on the lifter), or lifts over about 30 lbs. Designers of VR experiences may be able to mitigate much of the potential negative impact of potential discrepancies by staying under these thresholds in the object manipulations they display. We also found people tended to overestimate lower weight lifts and underestimate higher weight lifts [8, 41]. As with previous work [41], we found in Exp. 1 that people made less error estimating effort than estimating weight. We found larger, stronger lifters were accurately perceived as lifting heavier weights when they did so, unlike previous findings where this was unexpectedly reversed [42].

5 EXP. 2, BODY SHAPE

The goal of Exp. 2 was to understand how the size and shape of the avatar body impacted weight and effort estimates. This is relevant for situations where people are observing someone's avatar that may not match the observed person's actual body proportions. For a more realistic VR use case, the dumbbells were visualized with the avatar. In all cases, the size of the dumbbell was matched to the lift motion used (e.g., a 30-lb dumbbell if the lifter had lifted 30 lbs). Each motion was displayed on the avatar models of each of the four lifters (Body shape factor). Thus the motion and visualized dumbbell provided consistent signals, but the body shape could be inconsistent (i.e., matched to the original lifter in some clips, and not in others). One matched animation was included for each mismatched, so there were six clips for each lifter-weight combination (3 on the matched lifter's avatar and 1 on each of the unmatched ones). In total, there were 96 clips (6 body shape-motion combinations x 4 lifters x 4 weights). Note that the zero-weight lifts were not used. The presentation was fully randomized. After each clip, participants were asked to rate weight and effort as before, and also to rate the naturalness of the clip on a 7-point Likert scale. Since we are combining different lifter body shapes and lifter motions, we add 'm' after a lifter ID to specify motion and 'b' body shape, e.g., AA1m and AA1b.

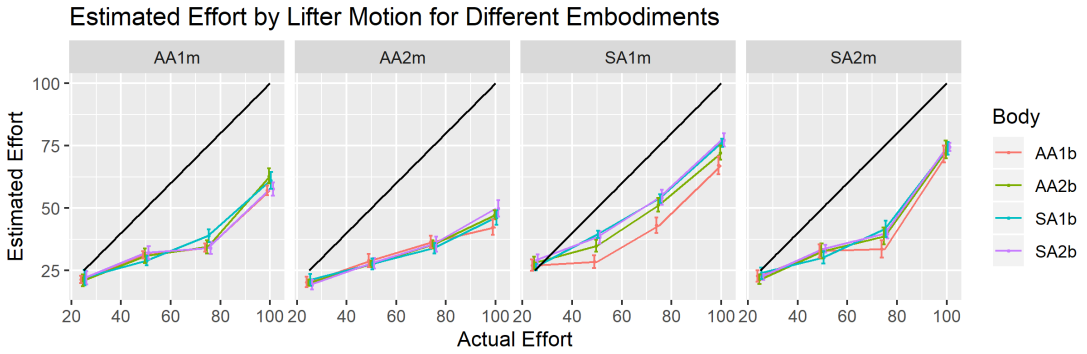

Fig. 8. Exp. 2: Perceived effort as body shape is changed. Facets show motion from different lifters, colors code different avatar bodies.

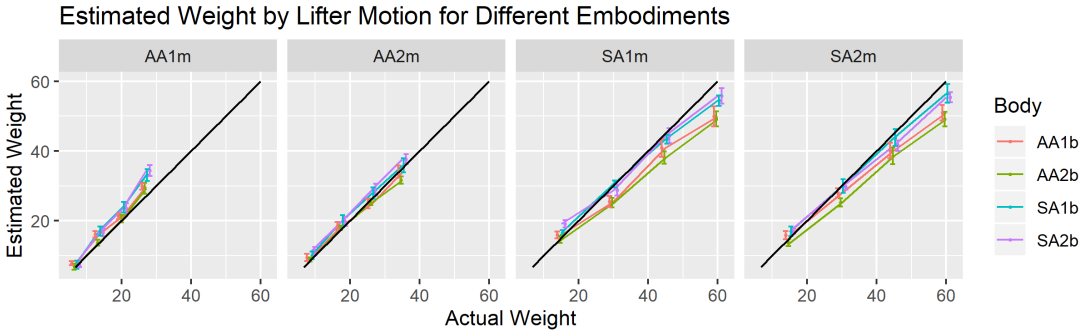

Fig. 9. Exp 2.: Estimated dumbbell weight as body shape is changed. Facets show motion from different lifters, colors code different avatar bodies.

Does body shape impact the perception of effort? Effort ratings for the motions of each lifter displayed on each avatar model are shown in Figure 8. It is clear that effort ratings are largely consistent across these variations in body shape. Linear mixed effect models fit to each lifter motion show that body shape did not have a significant impact on Effort ratings for AA1m, AA2m, or SA2m (Table 4(l)). However, there was a significant impact of Body shape on Effort ratings for the motion of SA1m. Post-hoc analysis shows that the only significant differences related to the AA1 body. Perceived effort on the AA1 avatar was significantly less than SA1 and SA2 bodies at 50 ($p < .0001, p = .0005$), 75 ($p < .0001, p < .0001$), and 100% ($p = .0001, p = .0003$) actual effort. It was also significantly less than AA2 at 25% ($p = .005$) and almost at 50% ($p = .0504$). It is not clear what is causing this difference.

Does body shape impact the inference of lifted weight? The effect of body shape on weight estimates is shown in Figure 9. Linear mixed effect models fitted to the data for each lifter all show significant main impacts of Body shape and actual Weight on inferred weight, but no interaction (Table 4(r)). Post-hoc analysis shows that the weight estimates for the average avatars were always significantly lower than those for the strong avatars (larger bodies), though the avatars of AA1 vs. SA1 fell slightly below significance for the motion of AA2 ($t = -2.411, p = 0.076$). This suggests that given the same perception of effort, people assume the larger avatars are lifting more weight.

Table 5. Exps. 1 and 2: Effort Ratings (l), Lifted weight ratings (r). "Dumbbell" is short for "Dumbbell Visibility". (Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ')

Lifter		χ^2	dof	$p(>\chi^2)$
AA1	Effort	622.7	3	< .0001 ***
	Dumbbell	3.92	1	.047 *
	Effort:Dumbbell	7.63	3	.054 .
AA2	Effort	421.2	3	< .0001 ***
	Dumbbell	1.92	1	.17
	Effort:Dumbbell	9.65	3	.023 *
SA1	Effort	1243.7	3	< .0001 ***
	Dumbbell	.0020	1	.96
	Effort:Dumbbell	5.68	3	.13
SA2	Effort	1052.9	3	< .0001 ***
	Dumbbell	0.83	1	.77
	Effort:Dumbbell	1.79	3	.62

Lifter		χ^2	dof	$p(>\chi^2)$
AA1	Weight	613.6	3	< .0001 ***
	Dumbbell	3.04	1	.081 .
	Weight:Dumbbell	52.62	3	< .0001 ***
AA2	Weight	659.2	3	< .0001 ***
	Dumbbell	.514	1	.47
	Weight:Dumbbell	63.88	3	< .0001 ***
SA1	Weight	976.2	3	< .0001 ***
	Dumbbell	3.61	1	.057 .
	Weight:Dumbbell	48.60	3	< .0001 ***
SA2	Weight	870.6	3	< .0001 ***
	Dumbbell	17.74	1	< .0001 ***
	Weight:Dumbbell	44.08	3	< .0001 ***

How much does body shape impact the inference of lifted weight? Since the effect of body shape on weight estimates was significant at the class level, Average vs. Strong, differences were calculated by comparing the means of the ratings for these classes. In 15 of the 16 cases, the estimates were heavier for the Strong class. The one outlier is the lightest lift performed by AA1 (lift was 6.75 lbs, mean Average estimate 7.56 lbs, Strong estimate 7.36 lbs). For the remaining classes, the mean Strong estimate was between 1.5 and 5.7 lbs heavier and generally increased for larger weights. As a percentage of the actual lift, the Strong estimates averaged 11% higher.

Does changing the body size impact the naturalness of the motion? A Cumulative Link Model was fit to the data with response variable Naturalness ratings and factors Effort, lifter Motion, and Body shape (character model displayed). There were significant main effects for Body shape and lifter Motion, all 2-way interactions were significant, but the 3-way was not. Both interactions, Effort:Body shape and lifter Motion:Body shape are shown in the Appendix. The significant differences from post-hoc analysis are marked. Most drops in Naturalness are almost exclusively related to the AA1 model. When AA1b is used on motion from larger lifters (which also feature larger dumbbells) and at higher efforts (heavier lifts), it looks less natural.

Impact of Showing Dumbbells. Between Exps. 1 and 2, we have data for the same motions with and without the display of dumbbells. This allows us to consider whether the visual appearance of dumbbells influences the perception of effort. Plots of effort with and without visualized dumbbells for the same motion and body shape from Exp. 1 and Exp. 2 are shown in Figure 10. Results are mixed (Table 5). There is no significant difference for SA1 and SA2. For AA1, there is a significant main effect for Dumbbell, but no significant interaction between Dumbbell and Actual Effort. Effort estimates with dumbbells present are lower. For AA2, there is a significant interaction between Dumbbell and Actual Effort, with post-hoc analysis showing the clips with dumbbells rate significantly lower ($t = 2.10, p = 0.038$) at 25% effort and no significant differences at other effort levels (100% effort is tendential ($t = 1.906, p = 0.059$)). The appearance of the dumbbells lowered the perceived effort for the Average lifters, likely because the participants believed these lifters could lift heavier weights than they could (cf. Exp. 1).

Does the visualization of dumbbells impact the inference of weight? Data from Exps. 1 and 2 for the same lifts, with and without dumbbells, are shown in Figure 11. The patterns are clearly different, with much-improved estimates when dumbbells are present. This difference was confirmed by again fitting a linear mixed effects model to the motion of each lifter. In all cases, there is a significant interaction between the lifted weight and the visual presence of dumbbells (Table 5). The lift magnitude where the presence of the dumbbell led to significant changes

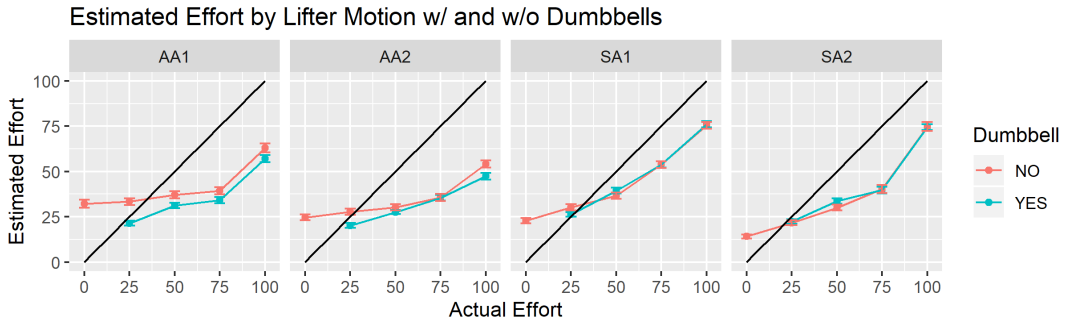

Fig. 10. Perceived effort with and without displayed dumbbells. Facets show motion from different lifters.

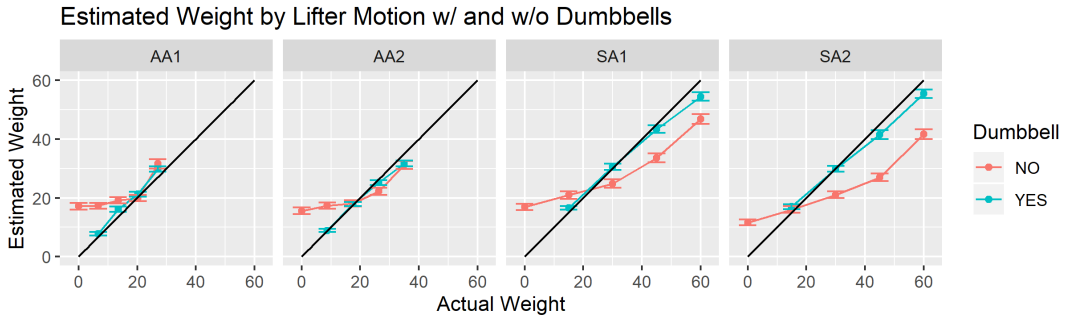

Fig. 11. Estimated weight with and without displayed dumbbells. Facets show motion from different lifters.

in ratings are the ones that appear visually different in Figure 11 (6.75 lbs for AA1; 8.75 lbs for AA2; 30, 45 and 60lbs for both SA1 and SA2) . The presence of weights seems to have led to the correction of judgment errors made in their absence. Correlations between real and inferred weight substantially improved when dumbbells were shown, and in all cases, the improvement is statistically significant using Fisher's Z transform to compare correlations (AA1: $r = .69$ vs. $.36$, $z = 6.2$, $p < .00001$; AA2: $r = .73$ vs. $.41$, $z = 6.49$, $p < .00001$; AA2: $r = .77$ vs. $.61$, $z = 4.1$, $p < .00001$; SA2: $r = .76$ vs. $.64$, $z = 3.14$, $p < .00001$) , suggesting that visualizing the dumbbells had a large impact on the ability to accurately estimate weight. Interestingly, if we fit a power function to these, the exponent is much closer to 1, also indicating a more linear relationship.

Discussion

In answering the question *"If people are visualized with avatars that have different proportions and mass than their own, how sensitive are observers to the resulting errors in their motion dynamics?"*, we saw that body shape had a limited impact on perceived effort. However, we saw that body shape changes do impact judgment of weight, with the larger "strong" avatars being judged as lifting heavier weights. This is consistent with Kenny et al. [28, 29], although their animations did not include a visualization of the lifted object, so that may not be necessary for the difference. This study adds a potential causal mechanism to that finding: since the impression of effort is largely constant, the same effort combined with a larger body appears to produce a larger estimated

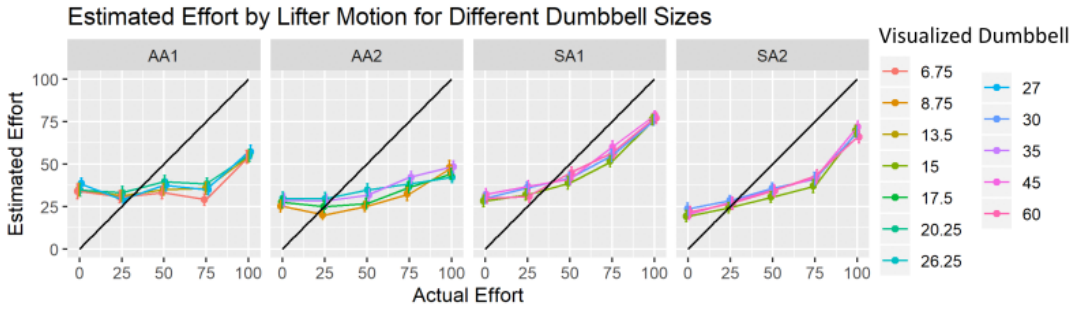

Fig. 12. Exp. 3: Perceived effort as a function of actual effort for different dumbbell sizes (indicated by line color). Facets show motion from different lifters. Note that estimated effort is largely a function of actual effort with limited variation from the different dumbbell sizes.

weight. We also found degradations in naturalness not observed in Kenny et al. that may result from including a visualization of the lifted object. This grounds the task and no longer allows people to justify mismatches by imagining that the weight of the object is different from what it is. The findings for weight may be connected to the Shape-Weight Illusion, where there seems to be the same underlying assumption that larger should be heavier.

6 EXP. 3, DUMBBELL SIZE

Next, we explored how changing the dumbbell size impacted weight and effort estimates, i.e., when a user's avatar is shown moving objects of different mass to what they are actually moving. The varied parameter was dumbbell size. Motions for each weight lift were used from each lifter, and displayed on the corresponding body, but each motion was shown with all possible dumbbell sizes for that lifter. In total, 80 clips were prepared (5 motions x 4 dumbbell sizes x 4 lifters). The zero lift motion was used, but only dumbbells with mass > 0 were displayed. As with Exp. 1, lifters were grouped and clips were randomized within each lifter. Participants rated weight, effort, and naturalness. Figure 12 shows the relation between actual effort and perceived effort for each lifter when dumbbell size is changed across clips. The clear takeaway is that estimated effort is largely a function of actual effort, with only a small impact from the visualized dumbbell as can be seen by the minimal spacing between the different weight lines on each plot.

Again, linear mixed-effect models were fitted to the data for each lifter. Statistical analysis indicates that in all cases, there was a significant main effect for dumbbell size (size indicates a particular weight), and for AA2 there is also a significant interaction between dumbbell size and effort (Table 6). The majority of the variation results from Effort, not Dumbbell size, however, and post-hoc analysis shows that the impact of dumbbell size is almost exclusively limited to the lightest dumbbell for each lifter being perceived as less effort to lift than some of the heavier three (For AA1, 6.75 lbs < 20.25 ($p = .0038$) and 27 ($p = .0280$); AA2: 8.75 lbs < 26.25 at 50% Effort ($p = .0003$) and (25% Effort ($p = .0002$); 8.75 lbs < 35 at 75% Effort ($p = .0002$), 50% Effort ($p = .027$) and 25% Effort ($p = .0021$). In the one case involving a dumbbell other than the lightest, 17.5 lbs < 26.25 at 50% Effort ($p = .0040$). For SA1, 15 lbs < 45 ($p = .0027$) and for SA2, 15 lbs < 30 ($p = .018$) and 45 ($p = .046$). This rather limited impact is consistent with the visually quite consistent ratings across dumbbell sizes shown in Figure 12.

The impact of dumbbell size on the inference of weight is shown in Figure 13. The visual size of the dumbbell has a very strong impact on the inferred weight. Fitting linear mixed effect models to each lifter showed a significant main effect for both actual Weight and displayed Dumbbell size for all lifters, and no significant

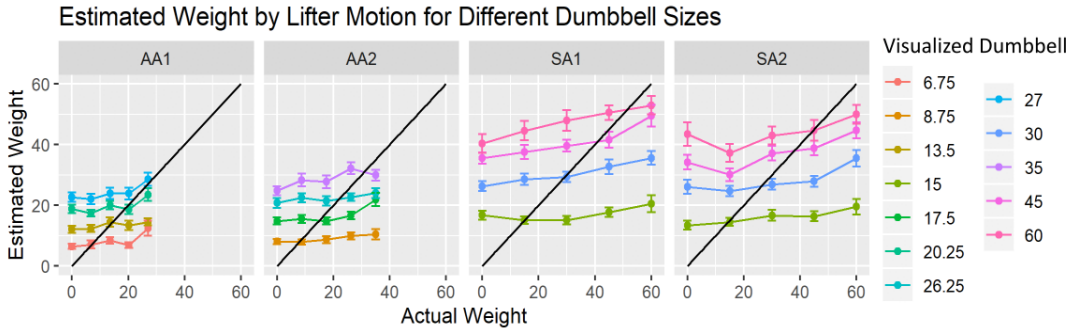

Fig. 13. Exp. 3: Estimated weight as a function of actual lift weight for different dumbbell sizes. Facets show motion from different lifters. Note that visualized dumbbell size has a clear impact on inferred weight.

Table 6. Exp. 3 Effort ratings (l), Lifted weight ratings (r). "Dumbbell" is short for "Dumbbell size".
(Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ')

Lifter		χ^2	dof	$p(>\chi^2)$
AA1	Effort	303.5	3	< .0001 ***
	Dumbbell	13.75	3	.0032 **
	Effort:Dumbbell	15.19	9	.23
AA2	Effort	176.2	3	< .0001 ***
	Dumbbell	39.67	3	< .0001 ***
	Effort:Dumbbell	35.93	9	.0033 **
SA1	Effort	649.9	3	< .0001 ***
	Dumbbell	12.78	3	.0052 **
	Effort:Dumbbell	11.48	9	.49
SA2	Effort	603.3	3	< .0001 ***
	Dumbbell	10.66	3	.014 *
	Effort:Dumbbell	10.98	9	.53

Lifter		χ^2	dof	$p(>\chi^2)$
AA1	Weight	38.3	3	< .0001 ***
	Dumbbell	427.1	3	< .0001 ***
	Weight:Dumbbell	6.24	9	.90
AA2	Weight	25.23	3	< .0001 ***
	Dumbbell	628.5	3	< .0001 ***
	Weight:Dumbbell	15.6	9	.211
SA1	Weight	65.30	3	< .0001 ***
	Dumbbell	794.3	3	< .0001 ***
	Weight:Dumbbell	14.36	9	.28
SA2	Weight	59.66	3	< .0001 ***
	Dumbbell	592.7	3	< .0001 ***
	Weight:Dumbbell	10.92	9	.53

interactions (Table 6). Unlike for effort, the majority of the variance was due to visualized Dumbbell size, not the actual weight lifted. Post-hoc analysis showed that every dumbbell size led to a significantly different estimated weight than every other dumbbell size for all lifters (all $p < .0001$). The impact of motion kinematics is more modest, but does exist, and is discussed in the supplemental material.

Naturalness ratings were used to evaluate if people were sensitive to mismatches between the lift motion and the displayed dumbbell. A single Cumulative Link Model was fit to the full data set with response variable Naturalness ratings and factors lifter Effort, Dumbbell size (as % of largest dumbbell used by lifter) and motion Lifter, along with all interactions. There were significant main effects for all three factors and significant interactions for Effort:Dumbbell size. Post-hoc analysis on the Effort:Dumbbell size interaction is shown inFigure 14 with significant differences marked. It can be seen that Naturalness ratings decrease for the more extreme combinations. For 0% effort lifts, the largest dumbbell was perceived as less natural.

Discussion

To answer the question *"How sensitive are people to visualizations that show an avatar moving a different mass than what they actually moved?"*, we conclude that dumbbell size has a major impact on estimated weight, but only a minor impact on perceived effort. As the discrepancy between visualized weight and effort reflected in the kinematics increases, these discordant signals are found to be less natural. This occurs when people lift nothing,

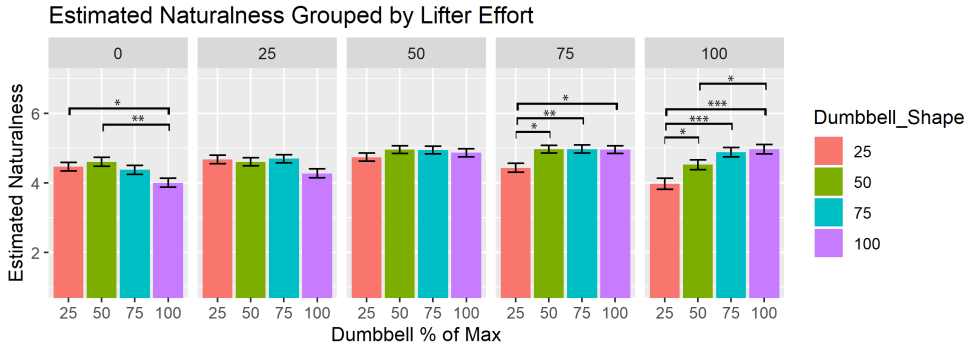

Fig. 14. Exp. 3: Naturalness ratings as a function of different dumbbell sizes. Facets show different effort levels.

but are shown with a heavy weight. This is the most likely problem case in VR, when a person just moves their hands or a light controller but is shown lifting a heavy object. At the other end of the spectrum, the smallest dumbbell size was seen as unnatural at both 75 and 100% lifts. At the 100% lift, the 50% dumbbell was also seen as less natural. This is an unlikely practical scenario in VR, but something similar might manifest in cases of fatigue or a particularly weak or sick user. While variation in object shape has been shown to impact weight judgments [37], we tried to minimize this impact by only scaling the object, so the general shape was fixed.

7 EXP. 4, STRAIN DEFORMATIONS

The goal of this experiment was to understand the impact of adding muscle deformations on the inference of weight and effort. In all cases, the motion clip was matched to the body model of the original performer. The dumbbells were not shown to allow a direct investigation of the relative impact of motion kinematics and visualized muscle strain. Each lift was shown with one of five deformation strain levels: NONE (the base clips used before), FULL (flexion of the body, face and neck in correspondence with the lifted motion), FACE (the face and neck flexion from FULL), BODY (the body and arm flexion from FULL) and PARTIAL (a reduced magnitude version of FULL), as shown in Figure 4 and the accompanying video. See Sec. 3.1 for details of how deformations were modeled. As this was a preliminary investigation into the impact of muscle strain, no attempt was made to tune the strain to the particular weight lifted, which would be a significant endeavor. Rather, we test the impact of a strong display of strain relative to motion kinematics. In total, there were 100 clips (5 strain levels x 5 motions x 4 lifters). Clips were randomized within each lifter group. A manipulation check of the muscle strain displays was successful and is described in the supplementary material.

Figure 15 shows the impact of deformation conditions on perceived effort. Linear mixed effect models were fit to the data for each lifter. The general ordering in terms of increasing perceived Effort is: NONE, BODY, PARTIAL, HEAD, and FULL. The differences largely relate to which of these are statistically separated (Table 7). For AA1 and SA1, there is a significant main effect of Deformation and no interaction. In both cases, HEAD and FULL are not significantly different, but each of the other levels are. For AA2 and SA2, there is a significant interaction between Effort and Deformation. For AA2, the differences that are not significant are: BODY - NONE and FULL - HEAD at all effort levels, HEAD - PARTIAL 0, 50 and 100% effort and FULL - PARTIAL 50 and 75% effort. For SA2, HEAD - PARTIAL are not significantly different at 0% effort and BODY - NONE, FULL - HEAD, FULL - PARTIAL and HEAD - PARTIAL are not significant at 50 or 75% effort. At 100% effort, there are two separated groups: FULL, HEAD, PARTIAL and NONE, BODY. Overall, the deformations have a clear impact on effort, especially those involving the head and neck (FULL, HEAD, PARTIAL). The impact of BODY on its own is more limited. Results are similar for the inference of weight and detailed in the Appendix.

Table 7. Exp. 4 Effort ratings (l), Naturalness ratings (r).
(Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ')

Lifter		χ^2	dof	$p(>\chi^2)$
AA1	Effort	126.2	4	< .0001 ***
	Deformation	549.0	4	< .0001 ***
	Effort:Deformation	17.1	16	.38
AA2	Effort	136.5	4	< .0001 ***
	Deformation	809.1	4	< .0001 ***
	Effort:Deformation	26.48	16	.048
SA1	Effort	454.3	4	< .0001 ***
	Deformation	482.7	4	< .0001 ***
	Effort:Deformation	24.77	16	.074 .
SA2	Effort	328.9	4	< .0001 ***
	Deformation	385.7	4	< .0001 ***
	Effort:Deformation	30.10	16	.017

	χ^2	dof	$p(>\chi^2)$
Effort	29.539	4	< .0001 ***
Deformation	18.077	4	.0011 **
Effort:Deformation	37.320	16	.0019 **

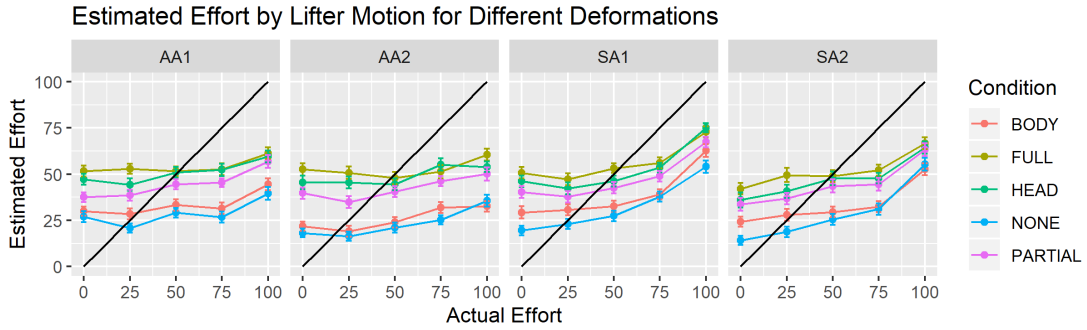

Fig. 15. Exp. 4: Perceived effort as a function of actual effort for different deformations. Facets show motion from different lifters.

Figures 16 and 17 quantify the impact of the strain deformations on effort and weight, respectively, by plotting the mean difference from the examples with no deformation. The general trends across deformation conditions reflect those of the previous analysis.

To analyze any impact on naturalness, we fit a Cumulative Link Model to the full set of data with Naturalness ratings as the response variable and factors lift Effort and Deformation, along with their interaction. It showed significant main effects for Effort and Deformation, and a significant interaction (Table 7). The data for the interaction is plotted in Figure 18, with significant differences marked based on a post-hoc, pairwise comparison. It can be seen that Naturalness ratings drop at either end when the strain deformations do not match the motion. For 0-effort lifts, the strain on the head and neck was seen as significantly less natural than no deformations or body-only deformations, which is consistent with the strain being a mismatch with the lift. Interestingly, it was also seen as less natural than FULL, so the combined body and facial strain cues were still plausible. At the maximum, 100% Effort, the no deformation condition was seen as less natural than PARTIAL, HEAD, and FULL, and BODY-only deformation was seen as less natural than HEAD and FULL.

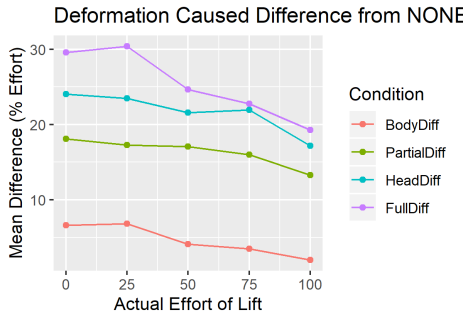

Fig. 16. Exp. 4: Perceived effort as a function of actual effort with various levels of muscle deformation.

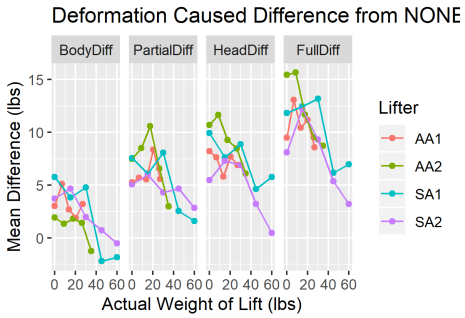

Fig. 17. Exp. 4: Estimated weight as a function of actual weight with various levels of muscle deformation.

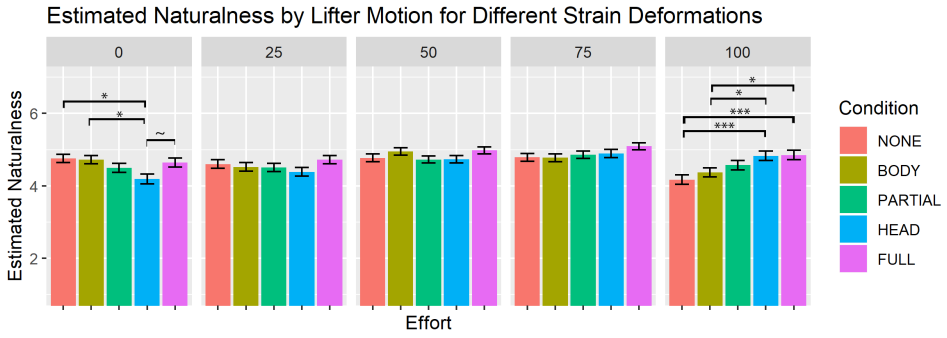

Fig. 18. Naturalness ratings with changes in deformation. Facets correspond to different naturalness levels.

Discussion

The answer to the question *"Can the display of muscle deformations, including facial strain, shift people's perception of effort and inference of weight in lifts?"* is a clear yes. The impact of the deformations can be substantial: up to a 30% increase of perceived effort for the FULL deformations and up to a 10-15lb increase in estimated weight. Results indicate that the impact of deformations is largest on the lightest lifts and reduces as the lifts become heavier. This suggests that the deformation and kinematic signals are acting in concert and when there is limited evidence of effort on the kinematic channel, the deformation channel can have more impact.

8 EXP. 5, DISCRIMINATION BETWEEN CORRECT AND FAKE MOTIONS, WITH AND WITHOUT MUSCLE DEFORMATIONS

This experiment had twin goals. The first was to understand when people could detect a zero-weight lift as being a fake lift, as compared to a lift that matched the visualized dumbbell size, and the second was to investigate if adding muscle deformations made it more difficult to detect zero-weight lifts. Zero-effort lifts were chosen as the comparison point because because they correspond to the motion a person would perform if they were interacting in VR using hand tracking. Unlike the previous experiments, this experiment was run as an Interval Forced-Choice experiment in which participants were shown two clips in sequence and had to decide which clip was a correct visualization of the lift. In all cases, one lift was a zero weight lift and the other lift was performed

Fig. 19. Exp. 5: Detection rates for zero lifts with and without deformations, compared to actual weight lifts. Asterisks indicate when the actual lift without deformations is significantly more likely to be correctly identified than chance. Braces indicate when the proportion of correct identifications is significantly different with muscle deformations than without.

with one of the weights greater than zero. Before each case, participants were told the weight of the target lift, and a dumbbell of that size was used in both clips. The order was randomized. Two types of pairs were run. One had no muscle deformation on either clip (NONE). The second had FULL deformation on the zero-weight clip and no deformation on the actual lift. Each participant saw 64 pairs (4 lifters x 4 weights x 4 pairs that randomly showed FULL or NONE as the comparison).

Psycophysics experiments are traditionally run with a high repetition count and a low number of participants. Instead, this was run with a relatively high participant count (35), and a low number of repetitions, in part because it is not clear that intersubject variation would not effect these judgments. It is also not the goal of the work to fit a psychometric function to the data nor establish precise detection thresholds. However, this does not provide enough repetitions to calculate a per-participant average. Each bar in Figure 19 shows the result of about 140 samples across the participant pool.

The teal bars in Figure 19 show the proportion of times people can detect the correct weight lift compared to a zero lift when both show the same dumbbell size. Those significantly above chance based on a one-tail exact binomial test are marked with an asterisk. Bonferroni correction was used for all the statistical tests in this section and the numeric test data is contained in the supplemental material. Light lifts of 25% effort are detected roughly at chance level, meaning people had difficulty distinguishing between these and zero lifts. The remainder are significantly above chance and for both average and strong lifters, the heavier lifts exceed the 75% threshold traditionally used in discrimination tasks.

The orange bars in Figure 19 show the proportion of people able to identify the correct lift, shown with no muscle deformation, when the opposing lift is a zero effort lift with the FULL strain deformation applied.The success proportions for the deformation and no-deformation zero lifts are compared with prop.test in R, which does a 2-sample test for equality of proportions with continuity correction. For both average and strong lifters,

people are more likely to detect the fake lift at 25% with full deformations, but less likely to do so for higher efforts (100 and 75%, respectively).

Discussion

The answer to our first question, *"Can people distinguish between a zero-weight lift and an accurate lift for various weight dumbbells?"*, is yes for heavier weights. It is helpful for VR designers to know that our lightest lifts, 25% of a person's maximum, were basically indistinguishable from zero-weight lifts for all lifters. This suggests that amelioration is not required in cases where avatars are shown lifting only relatively light objects. However, in all cases for the lifters' max lift, and sometimes as low as 50% of their max, people were able to detect the fake lift over 75% of the time. In these cases, some adjustment to the VR experience is likely required. Exact thresholds should be developed through follow up experiments, with more densely sampled weight increments and higher repetitions.

In all but one case, if the dumbbell lifted had been heavier than 25 lbs, the difference with a 0 weight lift was detectable at 75% or above. Kinematic aspects of the lift such as a compensatory weight shift will depend both on the amount lifted and the size of the lifter. It may be that heavier lifts created more kinematic signal, even at lower effort, so were easier to distinguish.

For the second question, *"Can displays of muscular sensitivity reduce sensitivity to these mismatches?"*, the answer is again yes, in some cases. For strong and average lifters, there was at least one weight at which participants were significantly less likely to detect the 0 weight lift if deformations were present. Large improvements generally occurred for weights between 13 and 45 lbs. In these cases, the addition of muscle deformation could be a useful technique for obscuring the fact that users were unencumbered when their avatars were lifting objects. For the max lift of the strong lifters, adding strain to the zero animation did make it less detectable, but this difference was no longer significant. It may be that at these extreme lifts, the kinematic signal was so strong that muscle deformations alone did not provide enough counter information to override it.

Interestingly, in all cases for 25% effort lifts, adding muscle deformations to the zero-lift made it easier to detect the correct lift. This is likely because the FULL deformation simply showed too much strain for the displayed dumbbell. The general trend is for the detection of the "zero lift with strain" to be easy at light lifts and become more difficult at heavy lifts. This is reasonable as the strain level used in these clips was quite intense, so only appropriate for heavier lifts. It speaks to the need to tune the strain level to the desired exertion of the character.

9 GENERAL DISCUSSION

Perception of effort appears to be largely driven by motion kinematics and, if present, displays of muscular strain. Visual size indicators of either the avatar or lifted object look to have a limited impact. The avatar size had no impact on effort for motion from three of our lifters (Exp. 2). For the fourth, SA1's motion, the effort estimates were lower when shown on the smallest avatar, AA1, for 25, 50, and 100% lifts, but did not otherwise significantly differ. When comparing the same motions with and without displaying dumbbells from Exps. 2 and 1, there was no significant impact on effort estimates for the strong lifters, but effort estimates were lower for the average lifters when dumbbells were present. A possible explanation is that, once dumbbells are shown, people corrected a faulty baseline assumption that the average lifters were stronger than they actually are. Varying dumbbell size had a limited impact on effort, largely constrained to the smallest of the four dumbbells being seen as lower effort than the remainder at particular effort levels for each lifter. When only kinematic information and these visual size indicators are present, it appears that kinematic information dominates the perception of effort. However, when muscle deformations are added to the animation, these have a clear impact on effort (Exp. 4), particularly those involving the head and neck (FULL, HEAD, PARTIAL). The impact of BODY flexion alone is more limited.

Inference of weight seems to be dominated by visual size indicators, especially of the lifted object, while kinematics still contribute. There was a consistent impact of body shape, where the smaller avatars were

estimated to lift about 11% less on average. Comparing motions with and without dumbbells from Exps. 2 and 1 established that displaying the dumbbells leads to improvements in weight estimates with significantly higher correlations between actual and estimated weight. Exp. 3 showed that visualization of the dumbbells had a large, and likely dominant, impact on the inference of weight with every dumbbell size seen as a significantly different weight than every other dumbbell for all lifters (e.g., the 60-lb dumbbell shown for SA1's 15-lb lift was estimated to be 44.6 lbs on average, whereas the 15-lb dumbbell shown with a 60-lb lift was only estimated to be about 20.5 lbs). Adding muscle deformations also has a clear impact, although the interaction of muscle deformation and dumbbell size remains to be explored. For AA1, every level of deformation was significantly different from the others except PARTIAL and HEAD. For the remaining lifters, there were three groupings: {NONE, BODY}, {PARTIAL, HEAD}, and {FULL}, from least to most impact. The strong impact of visual information on the inference of weight can be compared with that of visual information in the Size-Weight Illusion, where size impacts the perceived tactile heaviness of an object (non-visual size indicators can also invoke this effect).

Visualizing the size of the lifted object had a strong effect on people's estimates of weight, unlike previous findings for point-light displays [18]. This may in part stem from dumbbells providing a clearer indication of weight than boxes, as boxes may be filled with vastly different density material. Given the clear correspondence between effort and normalized weight estimates in Figure 7 when dumbbells were not shown (strongly correlated, with a Pearson's $r = .74$), it may be that people were using a single estimate of performance based largely on motion kinematics to estimate both weight and effort. When the visual dumbbell information was introduced, they relied heavily on that channel to estimate weight, but effort estimates were still largely based on kinematics. This led to divergence in the two estimates when information on the control channels was not congruent (e.g., mismatched dumbbells). It is possible that participants had fairly similar mental estimates of what a person could lift across the body variations shown and these were not consistent with the lifters' actual strength range. When they were given additional information, they appeared to revise these estimates (e.g., when shown small dumbbell size for the average lifters, they reduced effort estimates).

Muscle deformations added an additional signal that influenced the estimation of both effort and weight. In Exp. 4, the impact of the strain deformations is greatest at the lowest effort and weight levels and attenuates as these increase, which suggests that the deformation and kinematic signals are acting in concert. With limited visual evidence of effort on the kinematic channel, the deformation channel can have more impact. Exp. 5 showed that adding deformations to zero lifts with heavier dumbbells can make it harder to notice the difference between these animations and the correct lifts without muscle deformation. For light dumbbells, adding FULL deformations to the zero lifts makes it more noticeable that these are incorrect, presumably because this is an unrealistic amount of strain for the dumbbell shown. This result highlights the need for VR systems to carefully tune deformations to the desired weight/effort perception.

Strain animations could be used to mitigate the impact of mismatches between user kinematics and visualized motion. It is interesting that the HEAD deformation seems to carry much of the impact of FULL deformation for effort, and to a lesser degree for weight. This implies that VR applications could use only face and neck deformation on clothed characters, if appropriate, and still achieve most of the impact, if tuned to the desired effect. Such an approach would introduce an artificial strain signal to replace a signal missing in the motion kinematics, raising an interesting issue of how to balance verisimilitude with actual faithfulness to the person's behavior.

Naturalness ratings are lower for mismatching body types when the motion and dumbbell size of larger avatars are displayed on smaller avatars (motion and dumbbell sizes from all other lifters played on AA1). The AA1 motion also looked less natural at higher effort/weights, perhaps because the larger dumbbells looked less plausible for smaller bodies. In Exp. 3, naturalness fell when the size of the dumbbell was a poor match for the actual lift (either large dumbbells with low effort or small dumbbells with high effort). The cases where dumbbell size looked less natural in Figure 14 correspond to cases where the effort was out of line with weight estimates. Effort

perception was largely constant across the different displayed dumbbells for a particular weight lift, but weight varied heavily based on the displayed dumbbell. Finally, the HEAD deformation was less natural at 0% effort, as this shows a high level of strain that would mismatch with the kinematic signal. However, there is no similar drop for FULL. NONE and BODY were seen as less natural at 100% effort, due to a mismatch of a kinematic signal indicating high effort, but muscle deformations that do not reflect this effort. These findings emphasize the importance of calibrating all signals – kinematics, visual size indicators, and muscle deformations – to avoid degrading the user experience.

Limitations of our work include the use of male-only avatars that were not racially diverse. As a first study, this allowed us to easily have quite muscular avatars and display them shirtless, while also roughly matching the avatar to the lifter pool. It is important to explore if any of the findings here might change as the gender or race of the avatar varies. Stereotypes may come into play and this may also vary with the participant pool. A much larger study would be required to explore this. It would also be worthwhile to look at nonhuman avatars and the full range of beings that people may wish to embody in VR. Finally, looking at other physical quantities, such as the velocity of the motion, may be informative.

10 CONCLUSION

This paper describes a series of experiments that explore how people understand the dynamic properties of actions based on motion kinematics, the avatar's body, the size of manipulated objects, and muscle strain. By looking separately at people's perception of effort and weight, we were able to show that their judgments of these quantities are impacted by different signals. While effort is influenced by all control channels, motion kinematics appears to have a dominant role, especially when muscle flexion is not shown. On the other hand, visual indicators of size, particularly of the lifted object, have a strong influence on the inference of weight. If kinematics and visualization are not matched, this can produce incongruent information, where people's estimates of effort and weight are inconsistent and can lead to degraded naturalness. It may even be that such mismatched signals are one of the contributors to the Uncanny Valley effect [33]. While this paper indicates there is an operating range of moderate weights where such discrepancies are not likely noticed, going beyond that creates errors that must be avoided or mitigated, motivating the need for new animation algorithms.

ACKNOWLEDGMENTS

We thank Aaron Ferguson for his work on the deformation models and many insightful discussions. We also thank Vivian Lo and Hannah McDonald for assistance with logistics and experiment execution.

REFERENCES

[1] [n. d.]. Human Generator V2. https://www.humgen3d.com/ or https://blendermarket.com/products/humgen3d. [Tool for character modeling in Blender; Online; Accessed April 2021].
[2] [n. d.]. Wrap3D. https://www.russian3dscanner.com/. [Tool for working with textures; Accessed December 2021].
[3] Kaat Alaerts, Stephan P Swinnen, and Nicole Wenderoth. 2010. Observing how others lift light or heavy objects: which visual cues mediate the encoding of muscular force in the primary motor cortex? *Neuropsychologia* 48, 7 (2010), 2082–2090.
[4] Eric L Amazeen and Michael T Turvey. 1996. Weight perception and the haptic size–weight illusion are functions of the inertia tensor. *Journal of Experimental Psychology: Human perception and performance* 22, 1 (1996), 213–232.
[5] Douglas Bates et al. 2005. Fitting linear mixed models in R. *R News* 5, 1 (2005), 27–30.
[6] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2014. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823* (2014).
[7] Geoffrey P Bingham. 1987. Kinematic form and scaling: Further investigations on the visual perception of lifted weight. *Journal of Experimental Psychology: Human Perception and Performance* 13, 2 (1987), 155–177.
[8] Geoffrey P Bingham. 1993. Scaling judgments of lifted weight: Lifter size and the role of the standard. *Ecological Psychology* 5, 1 (1993), 31–64.
[9] Randolph Blake and Maggie Shiffrar. 2007. Perception of human motion. *Annual Review of Psychology* 58 (2007).

[10] Gunnar Borg. 1990. Psychophysical scaling with applications in physical work and the perception of exertion. *Scandinavian journal of work, environment & health* (1990), 55–58.
[11] Gavin Buckingham. 2014. Getting a grip on heaviness perception: A review of weight illusions and their probable causes. *Experimental brain research* 232 (2014), 1623–1629.
[12] Ryan Canales, Aline Normoyle, Yu Sun, Yuting Ye, Massimiliano Di Luca, and Sophie Jörg. 2019. Virtual grasping feedback and virtual hand ownership. In *ACM Symposium on Applied Perception 2019*. 1–9.
[13] Rune Haubo B Christensen. 2018. Cumulative link models for ordinal regression with the R package ordinal. *Submitted in J. Stat. Software* (2018).
[14] Funda Durupinar, Mubbasir Kapadia, Susan Deutsch, Michael Neff, and Norman I Badler. 2016. Perform: Perceptual approach for adding ocean personality to human motion using laban movement analysis. *ACM Transactions on Graphics (TOG)* 36, 1 (2016), 1–16.
[15] J Randall Flanagan, Jennifer P Bittner, and Roland S Johansson. 2008. Experience can change distinct size-weight priors engaged in lifting objects and judging their weights. *Current Biology* 18, 22 (2008), 1742–1747.
[16] Jann Philipp Freiwald, Julius Schenke, Nale Lehmann-Willenbrock, and Frank Steinicke. 2021. Effects of Avatar Appearance and Locomotion on Co-Presence in Virtual Reality Collaborations. In *Proceedings of Mensch Und Computer 2021 (MuC '21)*. 393–401.
[17] AM Gordon, H Forssberg, RS Johansson, and G Westling. 1991. Visual size cues in the programming of manipulative forces during precision grip. *Experimental Brain Research* 83, 3 (1991), 477–482.
[18] Lawrence EM Grierson, Simran Ohson, and James Lyons. 2013. The relative influences of movement kinematics and extrinsic object characteristics on the perception of lifted weight. *Attention, Perception, & Psychophysics* 75, 8 (2013), 1906–1913.
[19] Antonia Hamilton, Daniel Wolpert, and Uta Frith. 2004. Your own action influences how you perceive another person's action. *Current biology* 14, 6 (2004), 493–498.
[20] Jason Harrison, Ronald A Rensink, and Michiel Van De Panne. 2004. Obscuring length changes during animated motion. *ACM Transactions on Graphics (TOG)* 23, 3 (2004), 569–573.
[21] Jessica Hodgins, Sophie Jörg, Carol O'Sullivan, Sang Il Park, and Moshe Mahler. 2010. The saliency of anomalies in animated human characters. *ACM Transactions on Applied Perception (TAP)* 7, 4 (2010), 1–14.
[22] Ludovic Hoyet, Rachel McDonnell, and Carol O'Sullivan. 2012. Push it real: Perceiving causality in virtual interactions. *ACM Transactions on Graphics* 31, 4 (2012), 1–9.
[23] Ludovic Hoyet, Franck Multon, Anatole Lecuyer, and Taku Komura. 2010. Can we distinguish biological motions of virtual humans? perceptual study with captured motions of weight lifting. In *Proceedings of the 17th ACM Symposium on Virtual Reality Software and Technology*. 87–90.
[24] Timothy L Hubbard. 2020. Representational gravity: Empirical findings and theoretical implications. *Psychonomic bulletin & review* 27 (2020), 36–55.
[25] Eakta Jain, Lisa Anthony, Aishat Aloba, Amanda Castonguay, Isabella Cuba, Alex Shaw, and Julia Woodward. 2016. Is the Motion of a Child Perceivably Different from the Motion of an Adult? *ACM Transactions on Applied Perception (TAP)* 13, 4 (2016), 1–17.
[26] David Antonio Gomez Jauregui, Ferran Argelaguet, Anne-Hélène Olivier, Maud Marchal, Franck Multon, and Anatole Lecuyer. 2014. Toward" pseudo-haptic avatars": Modifying the visual animation of self-avatar can simulate the perception of weight lifting. *IEEE transactions on visualization and computer graphics* 20, 4 (2014), 654–661.
[27] Lynette A Jones. 1986. Perception of force and weight: theory and research. *Psychological bulletin* 100, 1 (1986), 29–42.
[28] Sophie Kenny, Naureen Mahmood, Claire Honda, Michael J Black, and Nikolaus F Troje. 2017. Effects of animation retargeting on perceived action outcomes. In *Proceedings of the ACM Symposium on Applied Perception*. 1–7.
[29] Sophie Kenny, Naureen Mahmood, Claire Honda, Michael J Black, and Nikolaus F Troje. 2019. Perceptual effects of inconsistency in human animations. *ACM Transactions on Applied Perception (TAP)* 16, 1 (2019), 1–18.
[30] Hye Ji Kim, Michael Neff, and Sung-Hee Lee. 2022. The perceptual consistency and association of the LMA effort elements. *ACM Transactions on Applied Perception (TAP)* 19, 1 (2022), 1–17.
[31] Woan Ning Lim, Kian Meng Yap, Yunli Lee, Chyanna Wee, and Ching Chiuan Yen. 2021. A systematic review of weight perception in virtual reality: techniques, challenges, and road ahead. *IEEE Access* 9 (2021), 163253–163283.
[32] Rachel McDonnell, Sophie Jörg, Jessica K Hodgins, Fiona Newell, and Carol O'sullivan. 2009. Evaluating the effect of motion and body shape on the perceived sex of virtual characters. *ACM Transactions on Applied Perception* 5, 4 (2009), 1–14.
[33] Masahiro Mori, Karl F MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & automation magazine* 19, 2 (2012), 98–100.
[34] Andreas Pusch and Anatole Lécuyer. 2011. Pseudo-Haptics: From the Theoretical Foundations to Practical System Design Guidelines. In *Proceedings of the 13th International Conference on Multimodal Interfaces (ICMI '11)*. 57–64.
[35] Michael Rietzler, Florian Geiselhart, Jan Gugenheimer, and Enrico Rukzio. 2018. Breaking the tracking: Enabling weight perception using perceivable tracking offsets. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.
[36] Sverker Runeson and Gunilla Frykholm. 1981. Visual perception of lifted weight. *Journal of Experimental Psychology: Human Perception and Performance* 7, 4 (1981), 733–740.

[37] Taebeum Ryu, Jaehyun Park, and Olga Vl Bitkina. 2022. Effect on Perceived Weight of Object Shapes. *International Journal of Environmental Research and Public Health* 19, 16 (2022), 9877.
[38] Elizabeth J Saccone, Oriane Landry, and Philippe A Chouinard. 2019. A meta-analysis of the size-weight and material-weight illusions. *Psychonomic bulletin & review* 26 (2019), 1195–1212.
[39] Majed Samad, Elia Gatti, Anne Hermes, Hrvoje Benko, and Cesare Parise. 2019. Pseudo-haptic weight: Changing the perceived weight of virtual objects by manipulating control-display ratio. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.
[40] Jaeho Shim and Les G Carlton. 1997. Perception of kinematic characteristics in the motion of lifted weight. *Journal of motor behavior* 29, 2 (1997), 131–146.
[41] Jaeho Shim, Les G Carlton, and Jitae Kim. 2004. Estimation of lifted weight and produced effort through perception of point-light display. *Perception* 33, 3 (2004), 277–291.
[42] Jaeho Shim, Heiko Hecht, Jung-Eun Lee, Dong-Won Yook, and Ji-Tae Kim. 2009. The limits of visual mass perception. *Quarterly Journal of Experimental Psychology* 62, 11 (2009), 2210–2221.
[43] Mel Slater and Martin Usoh. 1993. The influence of a virtual body on presence in immersive virtual environments. In *VR 93, Virtual Reality International, Proceedings of the Third Annual Conference on Virtual Reality*. 34–42.
[44] Harrison Jesse Smith and Michael Neff. 2018. Communication behavior in embodied Virtual Reality. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 289.
[45] Sinan Sonlu, Uğur Güdükbay, and Funda Durupinar. 2021. A conversational agent framework with multi-modal personality expression. *ACM Transactions on Graphics (TOG)* 40, 1 (2021), 1–16.
[46] Joseph C Stevens and Lee L Rubin. 1970. Psychophysical scales of apparent heaviness and the size-weight illusion. *Perception & Psychophysics* 8, 4 (1970), 225–230.
[47] Anne Thaler, Sergi Pujades, Jeanine K Stefanucci, Sarah H Creem-Regehr, Joachim Tesch, Michael J Black, and Betty J Mohler. 2019. The influence of visual perspective on body size estimation in immersive virtual Reality. In *ACM Symposium on Applied Perception 2019*. 1–12.
[48] Katja Zibrek, Benjamin Niay, Anne-Hélène Olivier, Ludovic Hoyet, Julien Pettré, and Rachel Mcdonnell. 2020. The effect of gender and attractiveness of motion on proximity in virtual reality. *ACM Transactions on Applied Perception* 17, 4 (2020), 1–15.

Appendix

A MODEL AND DEFORMATIONS

As discussed in the main paper, the character model and muscle deformations were created by an experienced artist. The base avatar model was generated with Human Generator V2, a tool for creating fairly realistic human models with varied body shapes in Blender[1]. Three types of blend shapes were used on the model (Table 8).

Body Type Shapes were used to create a variety of body shapes to account for various amounts of muscle mass and belly fat. These included a thin model, a very muscular model, and a "bellyOut" shape that indicated a large amount of fat around the midriff (used for AA2, who was somewhat stockier). The motion capture solver automatically scales the skeleton limb lengths based on a range of motion recording. The model was further fit to each lifter by adjusting these blendshapes, using both the markers and lifter footage as reference (Figure 3).

Corrective Shapes were created to preserve volume and improve anatomical detail as the model moved through a range of motion. These shapes were driven by the joint angle of a related skeletal joint with the one exception of the "latIn" shapes that were used to prevent penetration of the arm muscles with the latissimus dorsi. These 'latIn" shapes were driven by the distance between the elbow and the side of the body and provided a simulation of the interaction between these surfaces.

Tense Shapes were a collection of shapes built to emulate muscle strain and physical exertion (Figure 4). The limited range of motion in the study and the similarity of the animation cycles allowed these shapes to be grouped into regions. This reduced dimensionality made for easier retargeting to the animation clips.

These clips were animated with shape activations offset from one another in time. For example, shapes indicating great effort (such as the clenching of the mouth or squinting of eyes) were dialed in during "mid curl" where the effort expended is greatest. The video of the mocap session proved uninformative about actual deformation as the lifters were wearing black mocap suits. Instead, reference of weight lifters and videos taken by the artist helped inform the creation and animation of these shapes. Wrap3D [2] was used as a basis for many of the facial shapes.

The muscle deformations were animated manually for one reference clip and then retargeted to all clips. The retargeting process took as input the start-end frame of each up or downswing in both the reference clip and the target clip. It then scaled the timing of the keys from the reference clip to the target clip based on these landmarks in the timeline.

B MODEL VALIDATION

To confirm that the strain animations read as desired, we performed a manipulation check as an online experiment on Amazon Mechanical Turk using Mephisto library[1]. We have set qualifications such that only people who have already completed over 1000 tasks with above 95% approval rating could participate in the experiment. The duration of the task was 30 minutes and the paid amount was $7.5. Fifty participants viewed a sequence of clips and after each clip rated the prompt: "How much strain do you think the person in the video is exhibiting? (0 - no strain, 100 - maximum strain)". The videos contained 5 strain levels x 4 lifters x 2 samples for a total of 40 clips. All strain animations were done on a character in a static A-pose to avoid any impact from motion. A linear mixed effect model showed a significant main effect for Deformation [2]. The data averaged across lifters is plotted in Figure 20 and the pattern was similar for each lifter. Post-hoc analysis shows that the difference between every strain deformation was highly significant (p<.001). This confirms that the strain deformations are read as intended.

[1]https://github.com/facebookresearch/Mephisto

[2]Lifter $\chi^2(3) = 6.51$, $p = .089$, Deformation $\chi^2(4) = 1724.4$, $p < .0001$, Lifter:Deformation $\chi^2(12) = 6.29$, $p = .90$

Table 8. Blendshapes in the Avatar Model

Body Type:	
Slim	Reduces overall muscle mass
BellyOut	Increases torso fat
Corrective shapes:	
Leg	Leg lifting corrective
forearm	Corrective for arm bending at elbow
WristDown	Flexion of the wrist
WristUp	Extension of the wrist
LatIN	Prevents arm from penetrating the side of the torso
Tense Shapes:	
Torso Tense	Abdominal, obliques and pectoralis strain (abs and oblique deltas reduced dramatically for BellyOut variation)
NasilFold	Nasolabial fold (crease at the inside edge of cheek)
PlatFront	Platysmal sheet, the broad sheet of muscle fibers extending from the collarbone to the edge of the jaw
Plat	Platysmal sheet lateral (indicating extra strain/effort)
SternoMastoid	Large muscle from the corner of jaw/head to the start of collarbone
JawClench	Clench jaw
FaceClose	A "wince" comprised of closing of eyes, cheek raiser and tightening, raising of lips
LegFlex	All major muscles around knee
ArmFlex	Flex the bicep/tricep muscles and also added some forearm definition
EyeClose	Used in sync with face close to offset timing

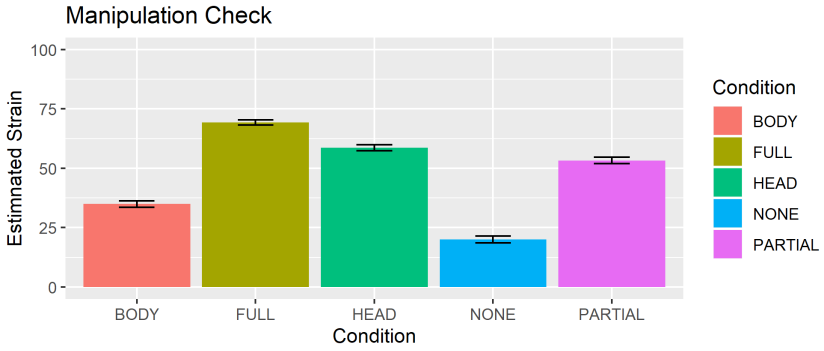

Fig. 20. Manipulation check ratings for the level of strain conveyed by each deformation condition.

C ADDITIONAL VISUAL RESULTS

The figures here complement those shown in the main paper.

Exp. 1: Visual inspection of Figure 5 shows similar slopes within the Average Strength and Strong groups, but differences between them, a trend made more clear in Figure 21. Lifter group does significantly affect observations

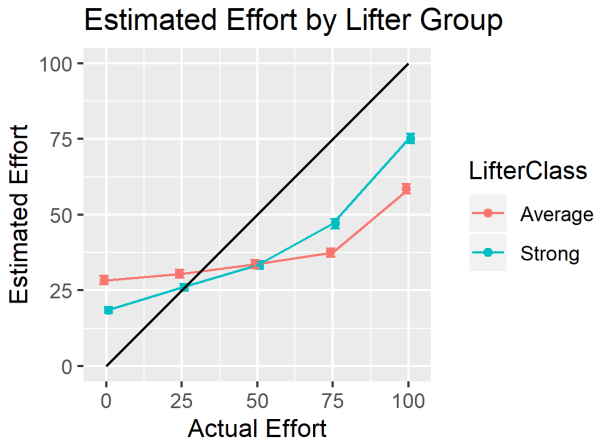

Fig. 21. Exp. 1: Perceived effort, grouped by lifter strength.

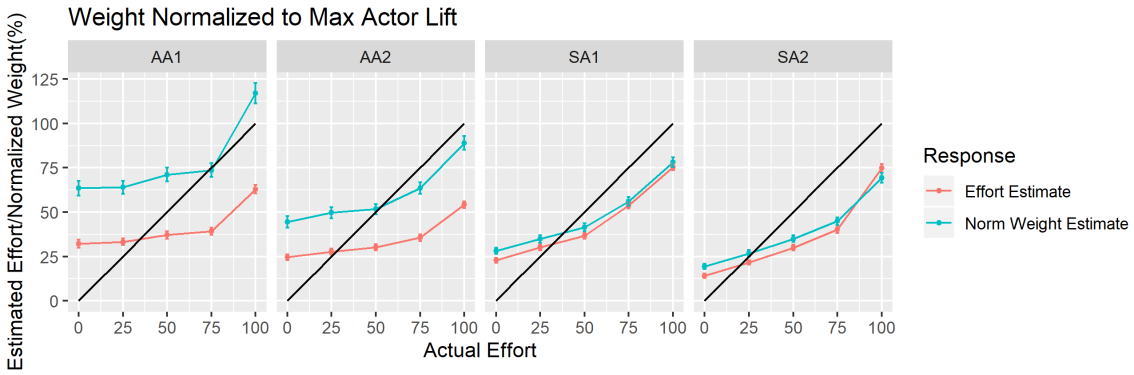

Fig. 22. Exp. 1: Perceived effort and inferred weight for weights that are normalized based on the actual max lift for each lifter.

based on fitting a linear mixed effects model ($\chi^2(1) = 15.645$, $p < .001$). People were more accurate in their effort perceptions for the strong group.

The paper offered a comparison between participants perception of the lifters' effort and estimates of the weight they lifted, with that weight normalized. Normalizing with the actual max lift of each performer produces the plots in Figure 22, which shows the curves have similar shapes, but are not aligned for the average lifters.

Exp. 2 collected naturalness ratings on a Likert scale of 1-7. The results are shown in Figures 23 and 24.

Exp. 3 looked at the impact of visualizing different dumbbell weights. The displayed dumbbell size had a strong impact on people's inference of weight, but motion kinematics still made a modest contribution. For all lifters, the heaviest lift was still seen as heavier than all others, except for the second heaviest lift for AA2. The additional significant differences were: AA2, {0 < 26.25-lb lift}; SA2, {15 < 30, 45-lb lifts}. Plotting estimated weight as a function of dumbbell size (Figure 25) illustrates that much of the weight estimate is based on the dumbbell size, but some clear variation comes from the kinematics of the motion.

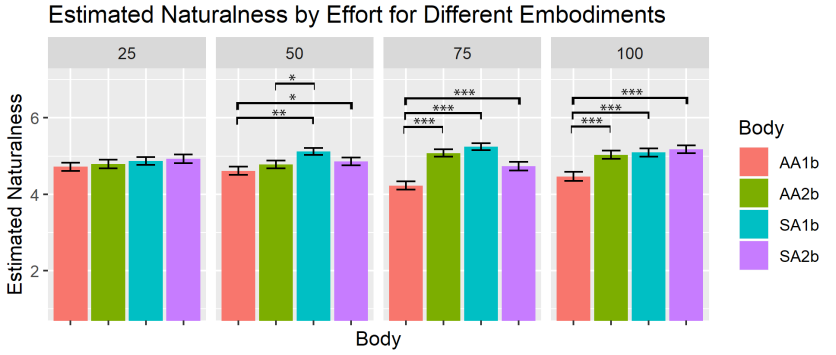

Fig. 23. Exp 2.: Naturalness for different body shapes. Facets show different effort lifts.

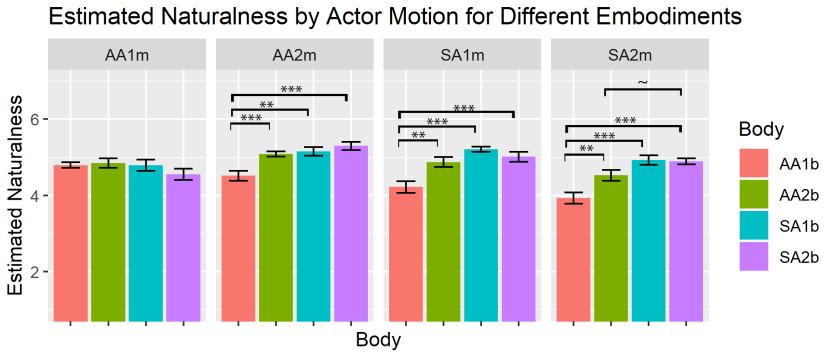

Fig. 24. Exp 2.: Naturalness for different body shapes. Facets show motion from different lifters.

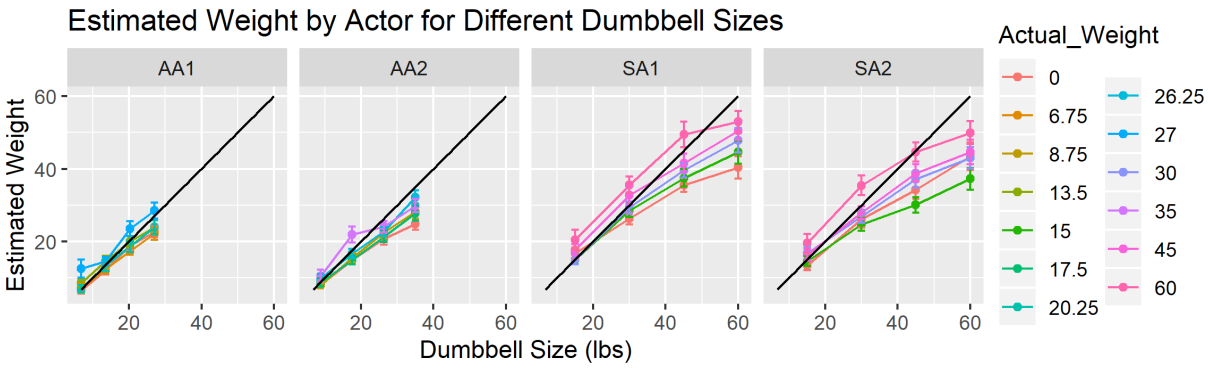

Fig. 25. Exp. 3: Estimated weight as a function of dumbbell size for different actual weight lifts. Facets show motion from different lifters, colors indicate different actual weights.

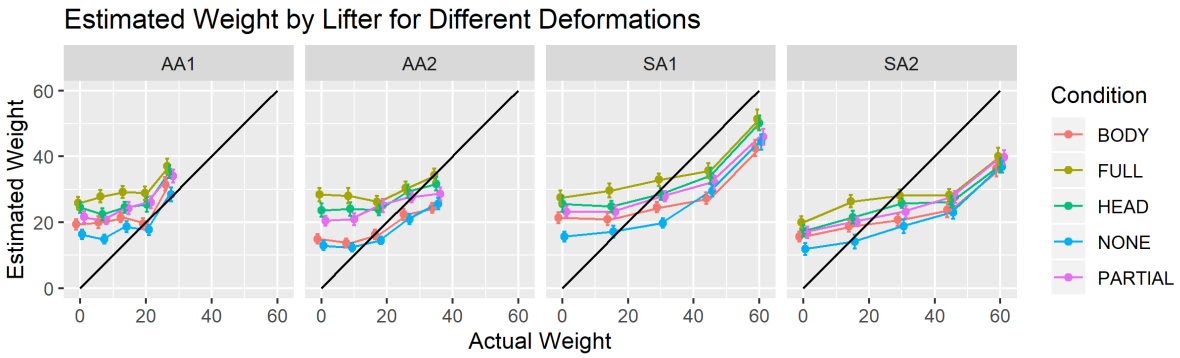

Fig. 26. Exp. 4: Estimated weight as a function of actual weight lifts for different deformations. Facets show motion from different lifters.

Exp. 4 examined the impact of visualized muscle deformations on the perception of effort and inference of weight. Effort perception is discussed in the main paper. Figure 26 shows the impact of muscle deformations on the inference of weight. A linear mixed effects model was fit to the data for each lifter. In each case, there was a main effect of Deformation, but not an interaction between Deformation and Weight (Table 9). The overall pattern is the same as with Effort, with the levels ordered NONE, BODY, PARTIAL, HEAD, and FULL and variation on whether they are statistically separated. For AA2, SA1 and SA2, there are three groupings: NONE, BODY, PARTIAL, HEAD and FULL. For AA1, all levels are significantly different except PARTIAL and HEAD. As with Effort, the deformations have a clear impact on the estimation of lifted Weight.

Table 9. Lifted weight ratings for Exp. 4
(Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ')

Lifter		χ^2	dof	$p(>\chi^2)$
AA1	Weight	174.0	4	< .0001 ***
	Deformation	250.9	4	< .0001 ***
	Weight:Deformation	15.0	16	.51
AA2	Weight	128.7	4	< .0001 ***
	Deformation	347.7	4	< .0001 ***
	Weight:Deformation	25.51	16	.061 .
SA1	Weight	499.2	4	< .0001 ***
	Deformation	205.6	4	< .0001 ***
	Weight:Deformation	25.1	16	.068 .
SA2	Weight	370/6	4	< .0001 ***
	Deformation	101.0	4	< .0001 ***
	Weight:Deformation	21.74	16	.15

Pairwise comparisons were performed for each level of effort and weight. In the ideal case, each would be significantly separated from its predecessor. Instead of just looking at the significance cutoff of $\alpha = .05$, which can overly simplify relationships, Table 10 shows the distribution of p-values for the ten pairwise comparisons for

p-value	[1, .5)	[.5, .1)	[.1, 0.05)	[.05, .01)	[.01, .001)	< .001
Effort						
AA1	0-25 50-75	25-50	0-50	25-75	0-75	rest
AA2	0-25 25-50		50-75	0-50		rest
SA1				25-50	0-25	rest
SA2				0-25 25-50	50-75	rest
Weight						
AA1	0-6.75 13.5-20.25	0-13.5 6.75-13.5 6.75-20.25	0-22.5			rest
AA2	0-8.75 8.75-17.5	0-17.5			17.5-26.25	rest
SA1			15-30	0-15		rest
SA2				0-15	15-30	rest

Table 10. Tukey-adjusted P-values for pairwise comparisons of the different stimuli levels. There are ten comparisons for each lifter. Pairs with significance values higher than .001 are indicated and the remainder are marked "rest". Grey cells are above the alpha = 0.05 test line (not significant).

each model. In every case, the largest effort/weight is clearly separated from all others (p<.001). However, for the average strength lifters, both effort and weight are generally not significantly different at the lower levels, given the sample size. This reflects the more moderate slope in this part of the response curve and a more moderate slope for the average lifters than the strong.

Table 11. Naturalness ratings for Exp. 2 using Analysis of Deviance: Type I Wald chisquare tests
*(Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ')*

Body Shape			
	χ^2	dof	$p(>\chi^2)$
Effort	7.562	3	.056
Body	64.897	3	< .0001 ***
Motion	9.624	3	< .0001 ***
Effort:Body	33.922	9	< .0001 ***
Effort:Motion	68.504	9	< .0001 ***
Body:Motion	47.592	9	< .0001 ***
Effort:Body:Motion	15.260	27	.97

Table 12. Naturalness ratings for Exp. 3 using Analysis of Deviance: Type I Wald chisquare tests
*(Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ')*

Dumbbell Size			
	χ^2	dof	$p(>\chi^2)$
Effort	43.016	4	< .0001 ***
Dumbbell	17.607	3	.00053 ***
Lifter	10.731	3	.0133 *
Effort:Dumbbell	53.709	12	< .0001 ***
Effort:Lifter	10.762	12	.55
Dumbbell:Lifter	10.493	9	.31
Effort:Dumbbell:Lifter	27.502	36	.84

Table 13. Exp. 5: Exact Binomial Tests evaluating the likelihood of detecting a zero weight lift vs. an actual weight lift

Zero lift detection		
Lifter Class	Effort	p
Average	25	.14
Average	50	.00076 *
Average	75	< .0001 *
Average	100	< .0001 *
Strong	25	.75
Strong	50	< .0001 *
Strong	75	< .0001 *
Strong	100	< .0001 *
α with Bonferroni correction is 0.00625		

Table 14. Exp. 5: 2-Sample Test for Equality of Proportions comparing the likelihood of detecting a zero lift, with and without deformations.

Zero lift detection comparison			
Lifter Class	Effort	χ^2	p
Average	25	29.72	< .0001 *
Average	50	6.96	.0084
Average	75	2.12	.15
Average	100	19.87	< .0001 *
Strong	25	42.21	< .0001 *
Strong	50	.662	.42
Strong	75	38.10	< .0001 *
Strong	100	1.86	.17
α with Bonferroni correction is 0.00625			