ECOLOGICAL PSYCHOLOGY, 5(1), 31-64

Scaling Judgments of Lifted Weight: Lifter Size and the Role of the Standard

Geoffrey P. Bingham
Indiana University

Runeson and Frykholm (1981, 1983) discovered that observers could judge the amount of weight lifted by another person when only the motions were visible in patch-light displays. Gilden and Proffitt (1989) suggested that this might have been experimental artifact. A standard had been included and might have been used to scale perceived motions to amounts of lifted weight. Were standards essential to the results? We performed experiments to investigate the role of the displays, of a standard, of haptic experience of a known weight, and of information about lifter size. The results demonstrated that haptic experience of a known weight and standards were equivalent in reducing random errors, but that standards, unlike haptic information, produced a contraction effect that increased systematic errors. Results without a standard were comparable to those of Runeson and Frykholm.

Observers also judged displays of three different-size lifters, each lifting maximum weights equal to a third of their body size. Static information for lifter size was controlled and no standard was used. Results demonstrate that the lifted-weight result is not experimental artifact.

Runeson and Frykholm (1981, 1983) discovered that observers were able to judge the amount of weight lifted by another person when only the lifting and carrying motions were visible in patch-light displays of the lifts. This result was remarkable because the magnitudes judged were mass related although the observed displays contained only kinematic information (i.e., properties of motion described in terms of length and time, but not mass). How did observers obtain information about mass from mere motions?

The design of Runeson and Frykholm's experiments left open the possibility that judgments of lifted weight were not based only on kinematic information in

Requests for reprints should be sent to Geoffrey P. Bingham, Department of Psychology, Indiana University, Bloomington, IN 47405.

the displays. The judgments could have depended on exposure to a standard that was presented prior to the remaining trials. Before judging lifts for a given lifter, participants observed a display of a midrange lift and were informed of the amount of weight lifted. Presumably, this led Gilden and Proffitt (1989) to suggest that the lifted-weight result might be an experimental artifact. They worried whether the ability to judge lifted weight from patch-light displays "is a general ability or one that is specific to the experimental designs that have been developed" (Gilden & Proffitt, 1989, p. 373). There are two separate respects in which the use of a standard may have enabled observers to provide reasonable estimations of lifted weight when they might otherwise not have been able to do so. The first is a question of identification and the second a question of scaling. The identification problem is to discover the properties of events that observers are able to recognize and the information for those properties (Bingham, 1987a, 1991). Are observers truly able to recognize "lifted weight" via particular forms of human movement? The scaling problem is to understand how observers determine magnitudes associated with identifiable perceptible properties? Do particular variations in lifting kinematics correspond uniquely to amounts of lifted weight?

First, is lifted weight truly perceptible? Runeson and Frykholm (1983) hypothesized that kinematic properties of lifting events captured in patch-light displays provided visual information about the perceptible dynamic property, "lifted weight." Kinematic properties enabling observers to recognize lifted weight would determine the form and orientation (i.e., positive or negative slope) of judgment curves. However, observers might have been judging only variations in kinematic properties that had no intrinsic significance concerning lifted weights. If those kinematic properties happened to covary with lifted weight, then conceivably, by virtue of the standard, observers might have been able to express variations in perceptible kinematic magnitudes in numbers descriptive of actual lifted weights.

For instance, consider the extent to which lifters lean to preserve balance when lifted weight has been added in front of the body. Assuming that observers could detect amounts of lean, the problem in using lean to estimate lifted weight would be in scaling angles of lean to amounts of lifted weight. Suppose for purposes of illustration that angle of lean had no intrinsic significance concerning the amount of lifted weight. (That is, it would be like the angle at which a lifter happened to be wearing a baseball cap. Perhaps the cap is adjusted at a large angle tilted upward on the back of the head with large weights and kept straight on with light weights. Perhaps the reverse.) Then, given the constraints of the experimental task, observers might have used the amount of lean observed in standard displays to relate angles and weight values in pounds.

The problem is that a single standard display does not provide enough scaling information to orient the judgment curve. Only a single value along the angle scale is related by a standard to a single value along the weight scale. In which

direction should the remaining portions of the respective scales be related to one another? That is, should zero angle correspond to zero weight or the largest possible weight? Perhaps the two scales could be oriented by virtue of both being open ended in one direction and closed in the remaining direction (i.e., at zero). But, lifted weight is not open ended. It has a maximum value.[1] Ultimately, angle of lean (or the angle of a hat) also has a maximum value. So this strategy would not work. Given no intrinsic relation, the scales could map either way. Perhaps, in the absence of constraint, observers simply map zeros and directions of increasing magnitude along the two scales. This possibility, however, was undercut by the results of Bingham (1987b, 1992).

Bingham (1987b) investigated lifting motions rather than leaning motions. Lifting motions were isolated from leaning motions by having lifters perform one-arm curls about the elbow while the body and upper arm were supported and immobilized. Heavier weights affected the lifting trajectories by reducing peak velocities reached during the lifts. Thus, larger weights corresponded to smaller peak velocities. Observers consistently were correct in their judgments of lifted weight. Larger kinematic magnitudes were related consistently to smaller values of lifted weight and vice versa.

Ultimately, if, as far as observers are concerned, the relation between kinematic properties and lifted weight is arbitrary, then reversals should be expected in the relative orientation of judged weight magnitudes and kinematic magnitudes. In fact, when we investigated the use of lean to estimate lifted weight, we did obtain reversals in judgments (Bingham, 1992). When we reduced displays to two patches, one on the head and one on the ankle, providing information only about the angle between head and ankle, some observers judged larger angles of lean as corresponding to smaller amounts of lifted weight. We emphasize that these displays provided no information about posture including knee and hip angles or limb activity. The implication of the result was that this angle-of-lean information did not intrinsically signify amounts of lifted weight. In fact, observers were not able to recognize the (reduced) angle-of-lean displays as a person lifting weight. On the other hand, reversals of judgment curves were never obtained in the experiments of either Runeson and Frykholm (1981, 1983) or Bingham (1987b). The implication of those results was that some kinematic property of the respective displays had intrinsic significance concerning amounts of lifted weight. In fact, observers were well able to recognize those displays in all cases as a person lifting weight. Thus, it is unlikely that the identification of lifted weight is at issue in determining whether the lifted-weight results have been artifactual.

The second way in which a standard display might have enabled artifactual judgments was with respect to scaling. What determined where judgment curves

[1]This value varies among lifters although perhaps in principle a maximum for human lifters might be derivable via methods of scale engineering.

appeared within the range of possible values; that is, what determined the slopes and intercepts given the form and orientation of the judgment curves? Although the detectable kinematic properties in these experiments might have had intrinsic significance concerning variations in lifted weight, lifted weight might be judged naturally relative to the lifter's lifting abilities rather than in any absolute scale of weight. Without accessory information about lifter size, lifted weight judgments might be expected to vary freely over a range defined by variations in the size and ability of lifters. The standard may have enabled observers to determine lifter sizes and abilities.

Again, suppose for illustration that angle of lean was the kinematic property detected and used to estimate lifted weight. If so, then information about lifter size was required for accurate judgment, for the following reason: Although the amount of lean is directly proportional to the amount of lifted weight, it is also inversely proportional to the lifter's body weight (and height). Thus, the relation between lean angle and lifted weight is one-to-many. The same angle of lean corresponds to different weights with different-size lifters. Without information about lifter size, lean angle could not provide information about absolute values of lifted weight. Balance-preserving leans are directly analogous in action to a balance beam which only provides information about the relative amounts of weight balanced on the beam. When one of the balanced weights is known, it can be used as a standard allowing any other weight subsequently balanced against it to be judged. The standard trials may have enabled observers to estimate the size or weight of the lifter so that lifter weight could be used, in turn, as an effective standard in subsequent judgment trials. If indeed angle of lean was the kinematic property used by Runeson and Frykholm's observers, then the standard or some other form of accessory information about lifter size would have been required for accurate judgments.

This can be tested by asking observers to judge, without a standard, lift displays similar to Runeson and Frykholm's original displays to see if variability increases to cover the range of weights lifted by lifters of all sizes. Likewise, if the relation between detected kinematic properties and lifted weight was arbitrary, then judgment curves should be expected to vary in orientation. These possibilities were investigated by Bingham (1987b) using one-arm lift displays. Although judgments of those displays made without the benefit of a standard exhibited a good deal of variability, they were hardly arbitrary or totally inaccurate. The tendency was for weights to be overestimated. Subsequently, when the displays were judged again with standard displays preceding judgment trials, mean judgments adjusted downward with an accompanying reduction in variability about the means. The implication of the lack of reversals in judgments and the approximate accuracy of judgments without a standard was that the kinematic property had intrinsic significance for observers concerning amounts of lifted weight. This meant that lifted weight was visually identified when the kinematic property was detected. However, the level of accuracy in

judgments was substantially less than originally obtained by Runeson and Frykholm with displays of lifts in which the entire body was used. Whether the original level of accuracy might be obtained using full-body lift displays without a standard remains to be determined.

Before we engage in such an investigation, however, it is important to consider the potential sources of variability in such a magnitude estimation task. The reason is that *magnitude estimation*—supplying numbers to describe scale in events—is a task peculiar to psychophysical experiments. Gilden and Proffitt (1989) wondered to what extent the ability to judge lifted weight like this might be a "general ability." Presumably, the general ability would be to assess variations in lifted weight, but not necessarily in terms of numbers. The use of numbers is an accessory of the laboratory, intended as an expedient means of measuring the general ability. It is much easier, for instance, than measuring and analyzing the kinematics of lifts performed by observers and used as a response measure. The use of numbers aside, people generally should be expected to have some skill in judging lifted weights because they frequently hand off loads to one another. The range of weights and the types of lifts typically involved in such transactions vary widely depending on the context. Nevertheless, given the need to adjust the mode of lifting depending on the load, together with the potentially dire consequences of lifting a load that is too heavy (viz., back injury and hernia), the average adult should be expected to have some skill in scaling loads handled by other people. Furthermore, we should expect variations in skill depending on past experience. For instance, undergraduates who lift regularly at the gym or who work during their summers in a grain feed store might be more skilled than others, not only in gauging loads with respect to their own and others' lifting capabilities, but also in applying numbers to observed loads.

The laboratory task in lifted-weight studies admits variation in two types of skill. The first is skill in isolating the perceptible property and in resolving variations in scale associated with it. Although resolution has been discussed frequently in the psychophysical literature, especially with respect to variations at the endpoints of a range, the issue of perceptual skill in isolating properties and in resolving magnitudes has been broached less often. To the extent that perception is an activity (J. J. Gibson, 1979/1986) and given the complexity of human movements captured in patch-light displays, we must expect variability in judgment data generated by differences in perceptual skill (E. J. Gibson, 1969; Runeson, 1989; Runeson & Vedeler, in press).

The second skill is in using numbers to express judgments. Many individual differences in the use of numbers have been identified including differences in the range of numbers typically used by individual observers, differences in the number of significant digits typically used, and differences in ability to use numbers in an organized fashion. In judgments assumed to involve mental calculations, calculating abilities are known to vary widely. The latter skill applies to situations in which an observer is required to relate values on one scale

to those on another. All these potential sources of variation might affect judgments of lifted weight. However, these sources of variability are not relevant to the mundane ability to assess lifted weight. They are a product of experimental method and should be treated as measurement noise.

There is no denying that the task of judging lifted weight in extrinsic units is odd and rather difficult, whether the units be metric or British. All participants in these studies initially reacted to this oddity, although most eventually found themselves equal to the task. Normally, people would use the type of information detected in these displays to determine whether or how they themselves should manipulate a load or perhaps to determine whether a co-worker should lift a load. The unit of lifted weight intrinsic to this task would be determined by the manipulative abilities of the observer or of an observed lifter. The extrinsic unit used in the experimental task must be mapped onto an intrinsic unit familiar to observers. Thus, in more mundane circumstances, there would be a single (intrinsic) mapping problem. The observer would have to relate weights relative to an observed lifter's ability to his or her own lifting ability; that is, roughly, weights scaled in terms of lifter effort would be scaled in terms of self-effort. This was ultimately the task intended for study. But the experimental design introduced a second (extrinsic) mapping problem, that of mapping the British scale of weights to both observer and lifter effort. We reserve discussion of this problem for later, noting at this juncture only that difficulty is introduced by topological differences in the scales for lifted weight and weight as used in physics. For instance, lifted weight has a maximum value whereas weight does not.

The current experiments involved lifted-weight judgments made without and with a standard. The variability associated with the two mapping tasks was investigated by relating pounds to lifted weight in four different ways:

1. A weight value in pounds was related through *haptic experience only* to the observer's own abilities by allowing the observer to lift a known amount of weight. This was expected to reduce the variability associated with the mapping of extrinsic units without affecting the intrinsic mapping. This condition should, therefore, yield the best estimate of the variability associated with the intrinsic mapping when performed on the basis of visual information about the weights lifted.
2. A weight value in pounds was related to the observer's abilities through haptic experience and, as a standard, to the observed lifter's abilities via an observed lift. The combination of the *haptic information and the standard* should have affected the intrinsic as well as the extrinsic mapping. An effect on the intrinsic mapping was possible because participants were provided a knowledge of lifted weight values (in pounds) with respect to both their own and the lifter's abilities. The lowest variability and greatest accuracy might be expected in this condition.

3. Observers were forced to rely on their own *past experiences and knowledge* of how weight values in pounds relate to levels of lifted weight, whether lifted by themselves or by others. The greatest variability and least accuracy in the mapping of both extrinsic and intrinsic units would be expected in this case. Any increase in variability over that found in the first condition would be attributable to the extrinsic mapping task.

4. A *standard value* was related to an observed lifter's abilities via an observed lift. This condition replicates the design used by Runeson and Frykholm (1981, 1983). The main effect should have been to reduce variability in the extrinsic mapping. However, once again, an effect on the intrinsic mapping was possible for observers who had a good knowledge of their own lifting abilities.

Finally, visual information about lifter size was provided. By itself, this manipulation was expected to improve the intrinsic mapping without affecting the extrinsic mapping. An evaluation of lifter size is a part of the intrinsic mapping problem. If the standard constrained the intrinsic mapping, then the standard must have provided information about lifter size. By providing alternative information about lifter size, the effect of the standard on the intrinsic mapping should have been annihilated with no change in the effect of the standard on the extrinsic mapping.

EXPERIMENT 1: EFFECT OF THE STANDARD

This experiment was performed to test the possibility that displays without a standard allow accurate judgments of lifted weights and to investigate the effect of a midrange standard. In the first scaling condition, no standard was used. Observers were asked to judge weight in pounds. In the second scaling condition, a midrange standard was used.

Viewing conditions were designed to minimize variability associated with mapping extrinsic units to lifted weight. Before making any judgments, observers lifted a known 35-lb weight. This controlled for variability in the familiarity of observers with lifted weight gauged in pounds.

Half the observers tested the same box as was lifted by the lifter in the displays. The box could be seen in the displays by virtue of reflective material attached to its uppermost edges at the ends. Observers who surmised that the boxes were the same could establish the height of the lifter by comparing the box to the lifter. Although lifter weight is the more important component of lifter size for the scaling of lifted weights, information about lifter height also is relevant to the extent that height and weight covary.[2] To test for the effect of this potential

[2]The extent to which patch-light displays might allow the size and proportions of a person to be apprehended remains to be investigated.

information about lifter height, the remaining half of the observers tested a 35-lb barbell instead of the box. This deprived them of potential information about the size of the box which could, therefore, no longer be compared in the displays to the lifter to determine his height.

Method

Apparatus. For recording, a Sony ¾-in. videotape system was used together with a Panasonic camera. During taping and during the experiment, a 19-in. black-and-white screen monitor was used with the contrast and brightness turned down so that events were shown as bright patches on a dark background. The lifter was dressed in dark, tight clothing with 2.5-cm-wide retroreflective tape attached to white linen strips pinned around the head, wrists, elbows, knees, and ankles, and pinned to the shoulders and hips. Reflective material was also attached to the ends of a 46 cm × 30 cm × 25 cm pine box. The weight of the box was altered using bags of sand. These bags were distributed evenly within the box. The box had comfortable and secure hand grips along the top of each of its short ends.

The box was placed on the floor at a distance of 2.5 m from a 77-cm-high table. The camera was positioned so as to provide a right-angle side view of lifter, box, and table. A movie light with parabolic reflector was placed adjacent to the camera directed at the lifter. The height of the lifter filled approximately 80% of the vertical extent of the monitor screen.

Lifters. One man, the experimenter, acted as the lifter in the study. He was 185 cm (6 ft 1 in.) tall and weighed 75 kg (165 lb). He was moderately experienced in general fitness activities.

Recording Procedure. Each recorded lift began with the visible box resting on the invisible floor. The lifter entered from the right, stood for a moment before the box, then lifted the box and raised it to a natural carrying height. The lift was performed by keeping the back straight and using the legs. With the forearms about horizontal and supporting the box laterally against the abdomen, the lifter took a couple of steps forward, placed the box on the table, and stepped back half a step. Once again, the box was lifted from the table. After turning around and walking back to the original location of the box, the lifter put the box down. The lifter then stood up and stepping over the box, walked off the screen to the right. This sequence was the same as that used in Runeson and Frykholm (1981).

Five weights were used: 5 lb (2.27 kg), 20 lb (9.07 kg), 35 lb (15.87 kg), 50 lb (22.68 kg), and 65 lb (29.48 kg).[3] The maximum weight was 39% of the lifter's

[3]British system units were used in the judgment study because participants were most familiar with them.

weight. The recordings were made in blocks of five lifts, one with each of the five weights taken in a random order. Each block of five lifts was preceded by a 35-lb lift to be used as a standard.

The lifter knew the weight of the box each time before he lifted it. A 2-min. rest was allowed between each lift. Three blocks of lifts were recorded, each in a different random order. The first block was used for practice trials. The remaining two blocks were used for analysis. A total of 18 lifts were recorded.

In the patch-light recordings used in Experiments 1 to 3, patches on the lifter could not always be seen depending on the exact orientation of a reflective patch relative to the camera and the movie light. Thus, as the lifter performed the lifts in the displays, patches winked out of and into view at times when they were not being occluded. Despite these irregularities, the activity of the lifter could be perceived quite easily.

Observers. Twelve (6 men, 6 women) University of Connecticut undergraduates from an introductory course in psychology participated in the experiment for course credit. Twelve (8 men, 4 women) Trinity College undergraduates from an introductory course in psychology also participated in the experiment for course credit. The Connecticut students tested a 35-lb box whereas the Trinity students tested a 35-lb barbell.

Experimental Procedure. The observers were seated, three to five at a time, 2 to 3 m from the monitor in a dimly lit room. Before judgment trials, participants were asked to lift a known 35-lb weight. Participants were first instructed how to perform the lift by bending the knees and keeping the back straight so as to avoid injury. Once all participants had tried the weight, observation of recorded lifts was begun. Observers judged the amount of weight lifted in pounds.

Trials were blocked by scaling condition. The first scaling condition was the *display-only* condition in which no standard was provided. No mention was made of the standard in this condition, and observers judged the standard lifts along with the others. The second scaling condition was the *display-with-standard* condition. In this condition, the midrange 35-lb lift that preceded each block of five lifts was labeled as 35 lb on the protocol sheets. Protocol sheets were collected from observers after each scaling condition, and new sheets were distributed. Participants were told not to communicate with each other during the session nor to disclose any other reactions during the task. Observers were given no other information about the lifter. In particular, no information was provided concerning the sex, size, or physical condition of the lifter.

After the 12 participants at Trinity College had finished making judgments within each scaling condition, they were asked to estimate the height, weight, and sex of the lifter. These judgments were written at the bottom of the protocol sheets.

Design. The main independent variables were weight tested (box, barbell), weight level (1 to 5), and scaling condition (display only, display with standard). Repetition of weight levels was also treated as an independent variable, forming a four-way 2 × 5 × 2 × 2 factorial design with 12 observations in each cell. The first variable was between subjects whereas the rest were within subjects.

Results and Discussion

Mean judged weights, plotted in Figure 1, followed a monotonic increasing trend with actual weight levels. The judgment curve changed in slope over scaling conditions.

That kinematic properties of patch-light displays enable observers to distinguish relative amounts of lifted weight was confirmed by the results. A mixed-design analysis of variance (ANOVA) was performed on the data with weight tested (box vs. barbell), weight level, scaling condition, and repetition of weight levels as variables. The effect of weight level was significant, $F(4, 88) = 133.08$, $p < .001$. Weight accounted for 56% of the variance. In post hoc paired comparisons, all weight levels were different from one another ($p < .05$) according to a t-test. When tested within scaling conditions, all weights were

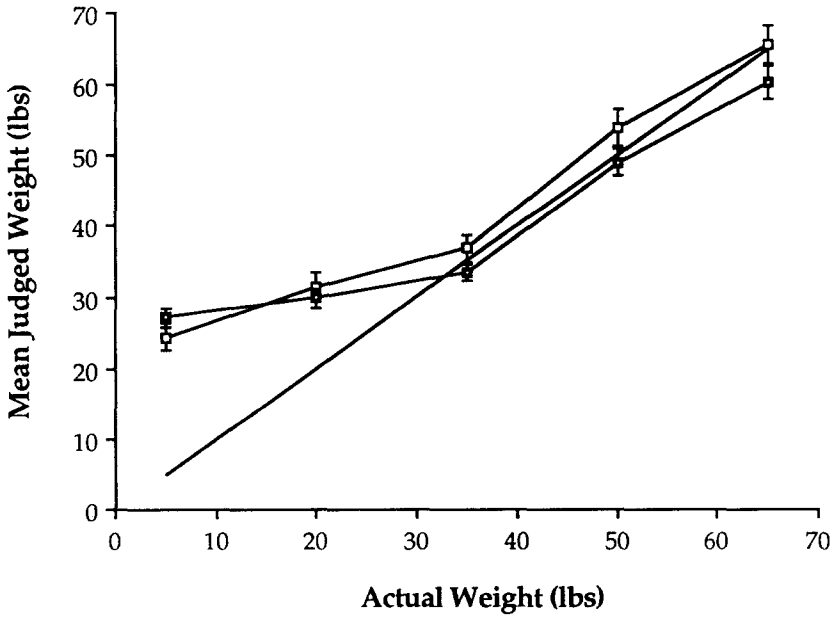

FIGURE 1 Mean judged weights (and standard error bars) from Experiment 1 plotted against actual weights. The diagonal line represents a relation of slope 1 and intercept 0. Conditions were display only (open squares) and display with standard (filled squares).

different ($p < .05$) in the display-only condition, but in the display-with-standard condition, Weight Levels 1 versus 2 and 2 versus 3 were not different ($p > .05$).

The weight-tested variable and interactions involving this variable were all non-significant. This indicated that observers who knew the size of the box did not need to use it to gauge the size of the lifter and, thence, lifted weights.

The same weight levels were judged in two repetitions. The effect of repetition was significant, $F(1, 22) = 7.95$, $p < .01$. Weights were judged as 3 lb heavier on average the second time. The Repetition × Weight Level interaction was also significant, $F(4, 88) = 4.41$, $p < .003$. In simple-effects tests, the first two weight levels differed significantly ($p < .02$), but not the remaining weight levels. The two lightest weights were judged as somewhat heavier in the second set of lifts. This result, together with the shallower slope in the second scaling condition, reflects the well-known "contraction effect" (Cross, 1973; King & Lockhead, 1981; Poulton, 1989; Slack, 1953; S. S. Stevens & Greenbaum, 1966; R. Teghtsoonian & M. Teghtsoonian, 1978). This effect results when, in repetitions of magnitude estimations, judgments become biased increasingly toward the middle range of values presented. Emphasis placed on the middle of the range by a midrange standard can exacerbate the effect. Lighter weight levels would be more subject to the effect due to decreased resolution. As reflected both in these results and in those of Runeson and Frykholm (1981), light weights are discriminated with greater difficulty. Although significantly different in the first scaling condition, the lighter (neighboring) weight levels were not different in the second scaling condition. (Mean judgments were used in subsequent analyses collapsing across the repetition of weight levels.)

The results of a regression of actual weight on judged weight in the display-only condition were comparable to those in Runeson and Frykholm's (1981) study. Runeson and Frykholm's slopes were .75 and .99 for two different lifters. The slope in the display-only condition of this study was .70 with an r^2 of .46, $F(1, 238) = 204.0$, $p < .001$. Rather than r^2s, Runeson and Frykholm reported the percentage of variance that weight levels accounted for in ANOVA. Their value of 68% was somewhat greater than the 56% found in the current study.

Judgments were affected by the inclusion of a standard in the second scaling condition. In ANOVA, the effect of scaling condition was not significant, but the Scaling × Weight Level interaction was significant, $F(4, 88) = 4.19$, $p < .004$. In simple effects tests, the two heaviest weight levels differed significantly ($p < .05$) whereas the remaining weight levels did not. In regression, the r^2 did not change appreciably ($r^2 = .49$), but the slope decreased to .57. Presumably, this change reflected a contraction effect resulting from an emphasis placed on the midrange value (Poulton, 1989). Judgments at all weight levels tended to be biased toward this value with a resulting increase in systematic error. The mean unsigned distance of judgment means from the actual weight levels increased slightly from 7.38 lb for display only to 7.84 lb for display with standard.

Random error, described in terms of the standard errors, decreased significantly over scaling conditions. The standard error values corresponding to each weight level were tested in a two-tailed, paired t test for significant differences between scaling conditions. The overall mean dropped from 2.26 lb to 1.62 lb. A parallel drop in standard errors occurred across weight levels. This change was significant, $t(4) = 6.66$, $p < .003$. With mere repetition of judged weight levels, standard errors increased, $t(4) = -3.12$, $p < .04$, with overall means of 2.77 lb and 3.54 lb, respectively. Thus, the inclusion of a standard appeared to reduce the amount of random error.

These results show that the ability to make estimates of lifted weight from the motion of patches in a display was independent of the availability of a standard. Although the standard improved judgment accuracy with respect to random error, systematic errors were increased by a contraction effect. Although the display-only judgments were more variable, the absolute values of the judgments were in the right ballpark for the observed lifter. The judgment curves were monotonic and the individual slopes were always in the correct direction.

How were observers able to scale judgments of lifted weight from kinematic information? If observers were using degrees of lean as information, then information about the size of the lifter would be required to scale the judgments. Is it possible that the displays contained information about the size of the lifter?

Observers at Trinity College (who tested a barbell) were asked to judge both the body weight and height of the lifter as well as the sex. Nine of the 12 observers correctly judged the sex of the lifter as male. One observer said he was unable to judge and 2 incorrectly judged the lifter as female. In the first scaling condition, the means of judged lifter weights and heights were 170.0 lb (SD = 12.8 lb) and 71.5 in. (SD = 1.4 in.), respectively. The actual weight and height of the lifter were 165 lb and 73 in. These results suggested that some information about lifter size was available in the kinematics. If information about lifter body size was used to scale judgments of lifted weight, then explicitly provided information about lifter size might result in additional improvements in the accuracy of judgments, both without and with a standard. Experiment 2 was performed to test this possibility.

EXPERIMENT 2: EFFECT OF KNOWING LIFTER SIZE

Part of the intrinsic mapping problem is to scale the size of the lifter to the size of the observer. To the extent that the standard improved the intrinsic mapping, observation of lifter size before making judgments might make the standard redundant, annihilating the scaling effect of the standard in the design of Experiment 1. On the other hand, to the extent that the standard improved the extrinsic mapping of pounds to lifted weight, the effect of the standard would

still be seen. Experiment 2 tested these possibilities. The procedure was the same as that employed in Experiment 1. However, observers were first allowed to see the size of the lifter before making judgments.

Method

The displays and procedure were identical to those used in the previous experiment with the following addition. The lifter was recorded in his patch-light outfit in normal lighting conditions with the camera placed so that the lifter filled approximately 80% of the vertical extent of the screen. Participants were informed that the lifter was the experimenter who was standing before them as they were allowed to observe him on the screen. Observers had previously stood next to the experimenter when trying the 35-lb box.

Observers. Ten (4 men, 6 women) University of Connecticut undergraduates from an introductory course in psychology participated in the experiment for course credit.

Results and Discussion

Mean judgments of lifted weight from both scaling conditions are plotted against actual weight levels in Figure 2. The shapes of the curves are similar to those in Experiment 1, although lighter weights were overestimated less. The judgments from the two scaling conditions coincided. No change in mean judgments occurred with the addition of standard displays.

In repeated-measures ANOVA with weight level and scaling condition as variables, the effect of weight level was significant, $F(4, 36) = 241.15$, $p < .001$. In post hoc paired comparisons, all weight levels differed significantly from one another ($p < .05$) according to a t test. Neither the effect of scaling condition nor the Weight Level × Scaling Condition interaction was significant ($p > .5$). Regressions of judged on actual weight produced identical slopes and intercepts for both scaling conditions (slope = .80, intercept = 7.2). In the first condition, r^2 was .52 and, in the second condition, it was .65. These results were an improvement over Experiment 1 and were as good as or better than results of Runeson and Frykholm (1981).

As indicated by these regression results and as shown in Figure 2, systematic errors were smaller in Experiment 1. The mean unsigned distances of judgment means from the actual weight levels were 5.15 lb in the display-only condition and 4.59 lb in the display-with-standard condition, both almost half the values in the previous experiment. The reduction in systematic errors produced by explicit information about the size of the lifter annihilated any changes in slope produced by the standard. However, the random error decreased in the same way as in Experiment 1. Overall mean standard errors dropped from 3.56 lb to

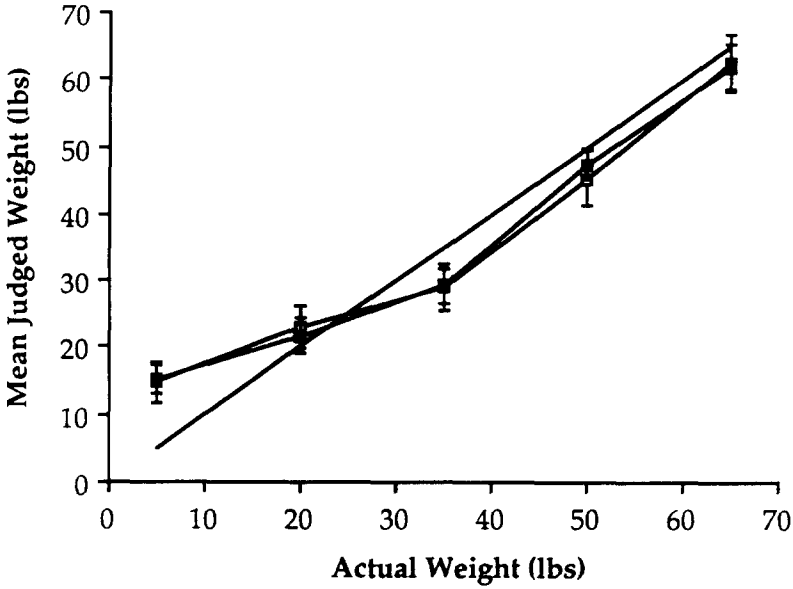

FIGURE 2 Mean judged weights (and standard error bars) from Experiment 2 plotted against actual weights. The diagonal line represents a relation of slope 1 and intercept 0. Conditions were display only (open squares) and display with standard (filled squares).

2.68 lb with a parallel drop at all weight levels. The drop was significant, $t(4) = 5.12$, $p < .007$.

To test the effect of having seen the size of the lifter, a repeated-measures ANOVA was performed comparing the data from Experiments 1 and 2 with experiment as a between-subjects variable and weight level and scaling condition as within-subject variables. The effect of weight level was significant, $F(4, 128) = 207.5$, $p < .001$, as expected. The effect of experiment was not significant. However, the Weight Level × Experiment interaction was significant, $F(4, 128) = 2.5$, $p < .05$. The effect of seeing the size of the lifter was to rotate the mean judgment curve around the largest weight level, lowering the assessment of the remaining weights and increasing the overall slope of the curve. Overall, the judgment curve was closer to the diagonal representing correct judgments.

Knowledge of lifter size eliminated all effect of the standard except on random errors. Mean estimates seem to have been made immune to a contraction bias. Also, slopes increased, showing greater discrimination of successively lighter weights. Why should knowledge of lifter size change the slopes, especially with a rotation around the heaviest value rather than the lightest?

Runeson and Frykholm (1981) suggested that variability in the slopes of individual judgment curves reflected variation in the conservativeness with

which observers used numbers, that is, the distance they were willing to go from the standard value. In the current case, the reference value seems to have been the heaviest weight. This makes sense because the heaviest weight was the weight most readily discriminated and identified. Borg (1962) suggested that muscular force or effort is scaled relative to the maximum values possible (and further, that all sensory scales might be so scaled because they have maximum detectable values). Following on Runeson and Frykholm's analysis, the current results suggest that as performance of the intrinsic mapping task was improved via knowledge of lifter size, observers became more confident and less conservative in their use of numbers below the maximum value. The result was steeper slope and greater immunity to bias by the standard. These results indicate that information about lifter size is an intrinsic part of lifted-weight judgments. By inference from the results of Experiment 1 in the display-only condition, the displays must contain information about lifter size. Furthermore, because angles of lean would not provide information about lifter size, we might infer that amount of lean is not the kinematic property primarily used to make judgments of lifted weight.

EXPERIMENT 3: CONTROL FOR REPEATED OBSERVATION

The design of Experiments 1 and 2 involved repeated observation and judgment of a single set of displays with standard trials added in the second repetition. To ensure the role of the standard display in producing improvement in random errors, a control experiment was performed in which observers repeatedly judged displays without the addition of a standard.

Method

The displays and procedures were the same as in previous experiments except that observers were not provided with a standard display preceding the second set of judgment trials.

Observers. Fifteen (4 men, 11 women) Indiana University undergraduates from an introductory course in psychology participated in the experiment for course credit.

Results and Discussion

The shapes of the curves were similar to those from Experiments 1 and 2. No change in mean judgments or in random errors occurred with mere repetition of judgments.

In repeated-measures ANOVA with weight level and repetition as variables, the effect of weight level was significant, $F(4, 56) = 81.93$, $p < .001$. Neither the repetition effect nor the Weight Level × Repetition interaction was significant ($p > .5$). In Experiments 1 and 2, decreases in random error were obtained with the inclusion of a standard trial in the second repetition. The results in Experiment 3 indicate that decreases in random error can indeed be attributed to the standard. The random error did not change significantly between the two repetitions, $t(4) = -1.90$, ns. Overall mean standard errors were 2.82 lb and 3.16 lb, respectively.

EXPERIMENT 4: VARYING LIFTER SIZE AND WEIGHT RANGE

Perhaps, without information about lifter size, observers produce values corresponding to the average-size lifter and the lifter in Experiments 1 to 3 just happened to fall near this average. Experiment 4 was designed to replicate Experiment 1 using a smaller lifter lifting a smaller range of weights. The design was identical to that of Experiment 1 except that observers did not test a 35-lb weight prior to making judgments. Greater variability, associated with the extrinsic mapping task, was anticipated in the display-only scaling condition as a result, but no change in systematic error was anticipated. The main question was whether the smaller range of lifted weights would be recognized by observers judging only on the basis of information in the displays?

Method

Patch-light recording of lifts were made using methods as in Experiment 1.

Lifters. One man acted as the lifter in the study. He was 179 cm (5 ft 10 in.) tall and weighed 63.5 kg (140 lb). He was highly trained as a long distance runner.

Apparatus and Recording Procedure. Five weights were used: 5 lb (2.27 kg), 20 lb (9.07 kg), 30 lb (13.61 kg), 40 lb (18.14 kg), and 50 lb (22.68 kg). The maximum weight was 36% of the lifter's weight. The weight used in standard trials was 30 lb. Due to the use of somewhat larger patches (6.35-cm wide retroreflective strips), these displays were more stable than those in the previous experiment. Patches were only deleted from the display when occluded.

Observers. Seventeen (10 men, 7 women) Indiana University undergraduates from an introductory course in psychology participated in the experiment for course credit.

Results and Discussion

Mean judged weights, plotted in Figure 3, followed a monotonic increasing trend with actual weight levels. Mean judged weights largely underestimated actual weights in the display-only condition. With the addition of the standard, the slope decreased as the judgment curve rotated, moving all mean judged weights toward the standard weight value. A repeated-measures ANOVA was performed on the data with weight level and scaling condition as variables. The effect of weight level was significant, $F(4, 64) = 67.1$, $p < .001$. In post hoc paired comparisons, all weight levels were significantly different from one another ($p < .05$) according to a t test.

The result of a regression of actual weight on judged weight in the display-only condition was comparable to that in Experiment 1 as well as in Runeson and Frykholm (1981). The slope in the display-only condition of the present study was .71 with an r^2 of .38 in comparison to a slope of .70 with an r^2 of .46 in Experiment 1. Greater variability was anticipated given the absence of the pretest of a midrange weight. With the inclusion of a standard in the second scaling condition, the slope decreased to .58 with an r^2 of .46 as compared to a slope of .57 with r^2 of .49 in Experiment 1. The increased variability in the display-only condition associated with the extrinsic mapping task was reduced

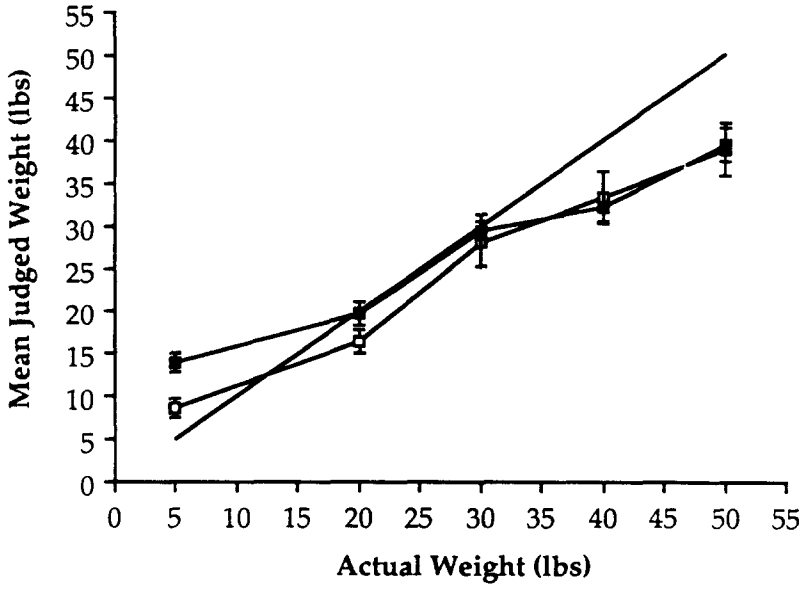

FIGURE 3 Mean judged weights (and standard error bars) from Experiment 4 plotted against actual weights. The diagonal line represents a relation of slope 1 and intercept 0. Conditions were display only (open squares) and display with standard (filled squares).

in the display-with-standard condition to levels comparable with Experiment 1. The standard reduced extrinsic mapping variability to previously obtained levels.

In repeated-measures ANOVA, neither the effect of scaling condition nor the Scaling Condition × Weight Level interaction was significant. (In simple-effects tests, the lightest weight level differed significantly, $p < .002$, whereas the remaining weight levels did not.)

Random error decreased over scaling conditions. The coefficients of variation corresponding to each weight level dropped between scaling conditions at all weight levels, with a mean decrease of .19. This change was significant in a two-tailed, paired t test, $t(4) = 5.78$, $p < .004$. The test of a difference in standard errors was only marginal, $t(4) = 2.34$, $p < .08$, with a drop in the mean standard error from 2.27 to 1.65.

The effect of the standard on systematic errors can be seen in Figure 3. Once again, with inclusion of the standard, the slope of the judgment curve decreased. Because the initial judgments tended to underestimate actual weights in this instance, the rotation of the curve was anchored at the high rather than the low end as it was previously. Mean estimations of heavier weights were unchanged whereas the mean estimation of the lightest weight was moved farther from the diagonal. The mean unsigned distance of judgment means from the actual weight levels increased slightly from 5.35 lb for display only to 5.56 lb for display with standard.

Overall, the results replicated those of Experiment 1. An ANOVA was performed comparing the data from Experiments 1 and 4 with lifter as a between-subjects variable and scaling condition and weight level as within-subject variables. The effect of weight level was significant, $F(4, 156) = 159.5$, $p < .001$. Scaling condition did not yield a significant effect, but the Scaling Condition × Weight Level interaction was significant, $F(4, 156) = 5.1$, $p < .001$, reflecting the rotation of the judgment curve by the standard. The Lifter × Weight Level interaction was significant, $F(4, 108) = 4.9$, $p < .001$. The actual and judged amounts of weight were the same for the two lifters at the lightest weight levels and then diverged for heavier weight levels to 65 lb versus 50 lb, respectively. This result shows that the ranges of weights lifted by the two lifters were discriminated as, by inference, were the different sizes of the two lifters.

Alternatively, the correct judgments of respective actual weight values were shown by the absence of a Lifter × Weight Level interaction when judgments were regressed against actual weight in a multiple-regression analysis performed on the data from Experiments 1 and 4 with vectors coding (orthogonally) for lifter, scaling, the 3 two-way interactions, and the 1 three-way interaction. The overall regression was significant ($r^2 = .535$, $p < .001$), but only weight, lifter, scaling, and the Weight × Scaling interaction were significant. The analysis, performed with the nonsignificant interaction vectors removed (Pedhazur, 1982), was also significant, $r^2 = .533$, $F(4, 815) = 232.8$, $p < .001$. Weight was

significant, partial $F = 146.5$, $p < .001$, with a beta weight of .63 and a slope of .64. Scaling was significant, partial $F = 4.5$, $p < .04$, with a beta weight of $-.10$ and a slope of -1.95. (The standard reduced weight estimates by about 2 lb.) The Weight × Scaling interaction was also significant, partial $F = 8.4$, $p < .004$, with a beta weight of .135. Separate simple regressions performed on each scaling condition produced decreasing slopes, from .75 to .61, and increasing intercepts, from 10.8 to 14.7. The two lines crossed at a weight of 28 lb. On average, the standard rotated the judgment curves around this midrange point yielding a classic contraction bias.

Finally, lifter was significant, partial $F = 50.2$, $p < .001$, with a slope of 5.65 and a beta weight of .28. Weight estimations for the second lifter were 5.6 lb less on average than those for the first lifter. However, the displays of the second lifter were judged without the benefit of having tested a known amount of weight. Random error associated with the extrinsic mapping had been expected to increase as a result of this, but no prediction had been made concerning systematic error. The next experiment was performed to test for the possibility that testing a known weight in advance of judgment might affect systematic errors as well as random errors.

EXPERIMENT 5: EFFECT OF TESTING A KNOWN WEIGHT

Observers were asked to judge displays of the first lifter without being given a standard and without testing a known amount of weight in advance. These judgments were compared with display-only judgments of the first lifter from Experiment 1 in which observers had first tested a known amount of weight. Additional observers were asked to judge displays of the second lifter without being given a standard but with testing of a known amount of weight in advance. These were compared with display-only judgments of the second lifter from Experiment 4 in which observers had not tested a known weight in advance. The two sets of judgments for each lifter were also compared to one another.

Method

A first group of observers judged displays of the first lifter from Experiment 1. These observers did not test a known weight before making judgments. A second group of observers judged displays of the second lifter from Experiment 4. Preceding judgment trials, these observers lifted a box which they knew to weigh 30 lb. All observers judged the displays without standards (and without seeing the size of the lifter beforehand).

Observers. Fifteen (4 men, 11 women) Indiana University undergraduates from an introductory course in psychology participated for course credit as observers of the first lifter.

Ten (2 men, 8 women) Indiana University undergraduates participated as observers of the second lifter. They were paid $5.00 for their efforts.

Results and Discussion

The effect of the pretest of a known weight was to increase mean judgments in all cases by about 8 lb as well as to decrease random error. A comparison of mean judgments for the two lifters with and without the pretest appears in Figure 4, which includes data from the display-only condition of Experiments 1 and 4.

A multiple-regression analysis was performed on these data with vectors for actual weight, lifter, pretest, the 3 two-way interactions, and the 1 three-way interaction. The result was significant, $F(7, 532) = 71.5$, $r^2 = .485$, $p < .001$, with only actual weight ($p < .001$) and pretest ($p < .003$) significant. In repetitions of the analysis, the interaction vectors failed to approach significance in any combination. When the analysis was performed without the interaction vectors, it was significant, $F(3, 536) = 167.7$, $r^2 = .484$, $p < .001$, and actual

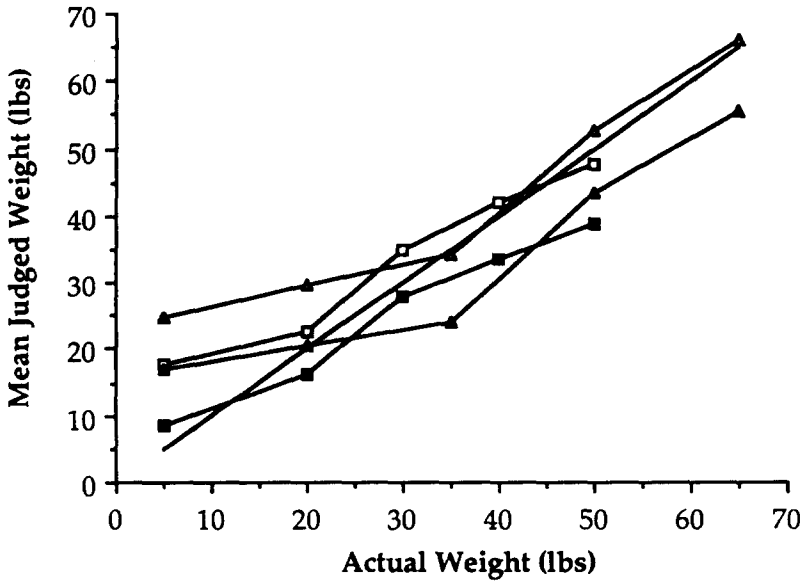

FIGURE 4 Mean judged weights for two lifters from Experiment 5 plotted against actual weights. The diagonal line represents a relation of slope 1 and intercept 0. The first lifter is represented by triangles, the second lifter by squares. Judgments were made without a pretest of a known weight (filled points) or with a pretest of a known weight (open points).

weight (partial $F = 414.3$, $p < .001$, $\beta = .64$), lifter (partial $F = 7.7$, $p < .006$, $\beta = -.09$), and pretest (partial $F = 45.0$, $p < .001$, $\beta = .21$) were all significant. The overall slope was .69. The effect of the pretest was to increase judgments by 8.6 lb on average. Judged weights for the first lifter were an average of 3.6 lb heavier at corresponding actual weight levels than for the second lifter. As apparent in Figure 4, most of this difference occurred at the lighter weight levels.

Three aspects of the results shown in Figure 4 were noted. First, the different ranges of weights lifted by the two lifters were recognized by observers. Even if the 3.6-lb-heavier estimation for the first lifter was to be added to second-lifter means, the respective heaviest weights of 65 lb and 50 lb would remain well discriminated. Second, these results were obtained in conditions where no standard was provided and even when no pretest of a known weight was performed. Thus, the information enabling observers to make these estimates was intrinsic to the displays. Finally, without the haptic experience of a lifted weight, estimates regularly tended to underestimate actual weight levels, whereas with the haptic experience, mean estimates were accurate at weight levels from 30 to 65 lb. Weights below 30 lb tended to be overestimated.

Random errors were comparable for the two lifters and were reduced in parallel by inclusion of the pretest. Two-tailed, paired t tests were performed comparing coefficients of variation for the second lifter without versus with the pretest (mean difference = .18), $t(4) = 5.0$, $p < .007$. These results were comparable to those for the first lifter (mean difference = .16), $t(4) = 4.2$, $p < .01$. Furthermore, when variability in the no-pretest condition for the second lifter was compared to that for the first lifter, the two-tailed, paired t was not significant (mean difference = .05). Similarly, a two-tailed, paired t test comparing coefficients of variation for the first and second lifters with pretest was not significant (mean difference = $-.03$). Finally, coefficients of variation for the second lifter without a standard but with a pretest and with a pretest but without a standard (second scaling condition of Experiment 4) were compared. The two-tailed, paired t was not significant (mean difference = .007). This last result indicates that the same reduction in random errors produced by inclusion of a standard was produced by inclusion of the pretest. The implication is that the random variability eliminated by the standard was associated with the extrinsic mapping task.

The two manipulations were equivalent in their effect on random errors, but they were not the same in their effect on systematic errors. Inclusion of the standard produced rotation of the judgment curves and shallower slopes reflecting the contraction effect. Inclusion of the pretest produced no change in the slopes of judgment curves (the interactions were nonsignificant in the multiple regression) as the medium-heavy mean judged weights moved closer to actual weights.

The effect of the pretest was to adjust the extrinsic mapping of pounds to lifted weight by a constant amount at all weight levels. Weights were labeled as almost

10 lb heavier then they were otherwise. Without the experience of a known amount of lifted weight, observers apparently were very conservative in their use of weight values.

EXPERIMENT 6: JUDGING THREE LIFTERS WEIGHING FROM 115 TO 190 LB

When naive observers judged patch-light displays of different-size lifters lifting different ranges of weight, mean judgments tracked actual weight values reasonably well. However, a couple of uncontrolled circumstances in the display generation may have contributed to the results. Two different tables were used in recording the displays. The one used with the smaller lifter was 50% of lifter height whereas the table used with the larger lifter was only 42% of his height. Furthermore, the width of the patches on the smaller lifter was 3.7% of his height whereas the width of those on the larger lifter was only 1.4% of his height. These discrepancies may have contributed to judgments apparently distinguishing the sizes of the two lifters. Different observers judged the two lifters. A within-subject design might be more sensitive. Also, weights below about 30 lb were not well discriminated. If a very small lifter were to lift a range of weights below 30 lb, would they be discriminated? Experiment 6 addressed these possibilities using methods designed to yield an accurate assessment of average abilities to judge lifted weight without a standard.

A single set of observers judged three lifters who varied in weight from 115 lb to 190 lb. Table height, apparent box height, and patch sizes were adjusted to constant proportions of lifter height. Observers lifted a weight known to be 30 lb before making judgments, but no standard was used.

Lifters each lifted a range of weights extending from 5 lb to a maximum weight equal to about 33% of their body weight. If observers were to apply a single range of weight values to all three lifters or if observers judged effort instead of lifted weight values, then the results should appear as in Figure 5. Weight values for the smallest lifter should be strongly overestimated as compared to those for the largest lifter. If, in contrast, lifted weights were judged correctly, all judgments would be on a single line of slope 1, intercept 0.

Successive weight levels were not equally spaced in a couple of cases to control for the possibility that observers might simply divide a judged range of weights into five equal weight intervals (Poulton, 1979, 1989; J. C. Stevens, 1958). In particular, the largest weight of the smallest lifter and the smallest weight of the middle lifter were spaced at two or three times the interval used for remaining weights in the respective ranges.

Method

Apparatus. For recording, a Panasonic AG-170 VHS camcorder was used. The tape was displayed during experiments using a Panasonic NV-8950 VHS

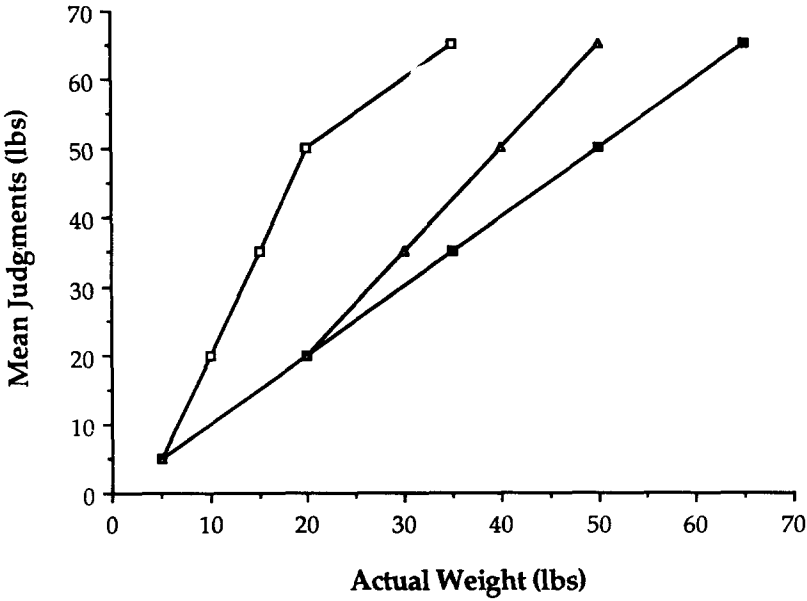

FIGURE 5 Predicted judgments for three lifters plotted against actual weights. The first lifter is represented by filled squares, the second lifter by open triangles, and the third lifter by open squares.

videocassette recorder. During taping and during the experiment, a Panasonic CT2580VY 24-in. monitor was used with the contrast, brightness, and color turned down so that events were shown as bright white patches on a dark background. The low light capabilities of the camera made it somewhat difficult to get the patch-light effect. This was achieved finally by placing a vertical strip of retroreflective tape about .3 m in front of the camera so that it appeared as a 2- to 3-cm bright strip forming the leftmost border of the display. This leveled the contrast values producing a good patch-light image. Black linen cloth was also draped over the rear wall of the room which served as the backdrop of the filmed events. The lifter was dressed in dark clothing with retroreflective tape attached to black linen strips pinned around the head, wrists, elbows, knees, and ankles, and pinned to the shoulders and hips. The width of the retroreflective tape was adjusted to 3.4% of lifter height (from 5.3 cm to 6.3 cm). Retroreflective tape was attached to the ends of the same pine box used in the previous experiments. The size of the patches was adjusted just as for patches on the lifters. The height of the patches on the box was adjusted to maintain a constant proportion of 16% of the lifter height (from 25.5 cm to 30 cm). The weight of the box was manipulated using bags of sand as before.

The box was placed on the floor at proportional distances from a table. The height of the table was adjusted to be 50% of lifter height (from 80 cm to 92.5

cm). The camera was positioned so as to produce a right-angle side view of lifter, box, and table and at a distance so that the event filled the left to right extent of the display while the lifter filled 80% of the vertical extent of the display. A 120–V 500-W blue movie light with parabolic reflector was placed adjacent to the camera directed at the lifter (and the strip of retroreflective tape in front of the camera).

Lifters. There were three lifters who volunteered for participation in the study. The first was a woman, 160 cm (5 ft 3 in.) in height, and weighed 52.2 kg (115 lb). The second was a man 179 cm (5 ft 10 in.) in height, and weighed 63.5 kg (140 lb). The third was a man, 185 cm (6 ft 1 in.) in height, and weighed 86.2 kg (190 lb). All were at least moderately fit.

Recording Procedure. Each recorded lift proceeded exactly as described in Experiment 1. Five different weight levels were used for each lifter. For Lifter 1, the weights were 5 lb (2.27 kg), 10 lb (4.54 kg), 15 lb (6.80 kg), 20 lb (9.07 kg), and 35 lb (15.88 kg); the maximum weight was 31% of the lifter's weight. For Lifter 2, the weights were 5 lb (2.27 kg), 20 lb (9.07 kg), 30 lb (13.61 kg), 40 lb (18.14 kg), and 50 lb (22.68 kg); the maximum weight was 36% of the lifter's weight. For Lifter 3, the weights were 5 lb (2.27 kg), 20 lb (9.07 kg), 35 lb (15.88 kg), 50 lb (22.68 kg), and 65 lb (29.48 kg); the maximum weight was 34% of the lifter's weight.

As before, the recordings were made in blocks of five lifts, one with each of the five weights taken in a random order. Each block of five lifts was preceded by a 30-lb lift to be used as a standard in other experiments. Three blocks of lifts were recorded each in a different random order. The first block was used for practice trials. The remaining two blocks were used for analysis. A total of 18 lifts was recorded per lifter. The lifter knew the weight of the box each time before he or she lifted it. A 2-min rest was allowed between each lift.

Observers. Sixteen (6 men, 10 women) Indiana University undergraduates from an introductory course in psychology participated in the experiment for course credit.

Experimental Procedure. Before making judgments, observers tested a known 30-lb weight just as they did in Experiment 1. The observers were then seated 3 to 5 m from the monitor in a dimly lit room. Observers were instructed to judge the amount of weight lifted in pounds. Participants were told not to communicate with each other during the session nor to disclose any other reactions during the task. Observers were given no other information about the lifter. In particular, no information was provided concerning the sex, size, or physical condition of the lifter.

Design. The main independent variables were lifter(1 to 3), weight level (1 to 5), and repetition of weight levels, forming a three-way 3 × 5 × 2 factorial design with 16 observations in each cell. All variables were within subjects.

Results and Discussion

As before, mean judged weights exhibited a monotonic increasing trend when plotted against actual weights as shown in Figure 6. All mean judgments lay close to a diagonal of slope 1, intercept 0, rather than being spread apart as in Figure 5.

When judgments were regressed on actual weights in a multiple-regression analysis with a vector coding for lifter and the Lifter × Actual Weight interaction, the result was significant, $F(3, 476) = 165.6$, $p < .001$, $r^2 = .51$. Only actual weight was significant. This was also true when the analysis was performed again without the interaction vector. Thus, mean judgments for the different lifters were not significantly different in slope or in intercept. When judgments were regressed on actual weight alone, the result was significant, $F(1, 478) = 493.7$, $p < .001$, $r^2 = .51$, with a slope of .74 and an intercept of 9.3. When simple regressions were performed independently for each lifter, they

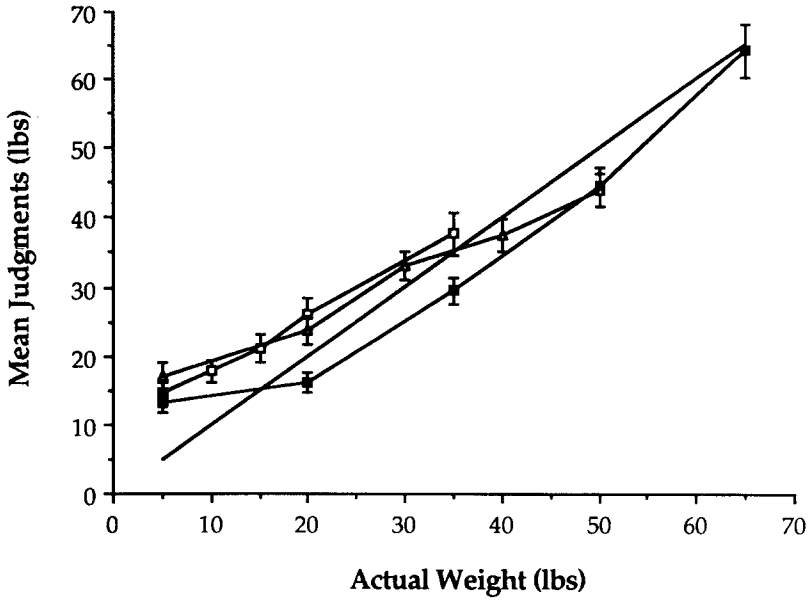

FIGURE 6 Mean judged weights (and standard error bars) for three lifters plotted against actual weights. The diagonal line represents a relation of slope 1 and intercept 0. The first lifter is represented by open squares, the second lifter by open triangles, and the third lifter by filled squares.

were significant ($p < .001$) in all cases with slopes of .78, .61, and .87 for Lifters 1 to 3, respectively.

Lifters lifted different amounts of weight at each of the 5 weight levels except the first, for which all lifters lifted 5 lb. Were weights at remaining levels distinguished among lifters? To test for this, a repeated-measures ANOVA was performed on judgments with lifter, weight level (1 to 5), and repetition as variables. The weight-level effect was significant, $F(4, 60) = 100.0$, $p < .001$, and accounted for 42% of the variance. Both the effect of lifter, $F(2, 30) = 11.2$, $p < .001$, and the Lifter × Weight interaction, $F(8, 120) = 13.3$, $p < .001$, were significant. When plotted against weight level, mean weight judgments for the three lifters were almost the same at the first weight level from which they increasingly diverged at successive weight levels. In simple-effects tests, lifter was significant ($p < .001$) at all weight levels except the first ($p > .1$), whereas weight level was significant ($p < .001$) at all levels of lifter. Thus, actual weights lifted at each weight level by different lifters were judged as the same when they were the same and as different when they were different. In particular, the maximum weight levels for the three lifters were recognized appropriately as such.

Irregular spacing of actual weight levels did not result in significant departures from linearity in the respective judgment curves. In particular, weights for the smallest lifter were spaced at intervals of 5 lb up to the heaviest weight which was spaced at 15 lb from the previous weight level. Despite this, the judgment curve continues to be approximately linear along its whole length.

Finally, discrimination of successive weight levels continued to be difficult at weights below 20 to 25 lb, even for weights lifted by a smaller lifter. In post hoc paired comparisons by t test, all weight levels were different from one another ($p < .01$) for the largest lifter except Weight Levels 1 versus 2. All weight levels were different from one another ($p < .05$) for the middle lifter except Weight Levels 3 versus 4. All weight levels were different from one another ($p < .05$) for the smallest lifter except Weight Levels 1 versus 2, 2 versus 3, and 3 versus 4 (nevertheless, for instance, 1 was different from 3).

These results are representative of abilities of university undergraduates to judge lifted weights. Performance might well improve with trained or practiced observers. However, without special training, the average undergraduate was able to recognize the different ranges of weights lifted by lifters of different size. These judgments must have been based on the detection of a kinematic property of the displays because all static sizes and proportions in the displays had been controlled and because these observers performed judgments without standards. First, the detectable kinematic property must have had intrinsic significance concerning amounts of lifted weight. Second, the property must have been scaled intrinsically to amounts of lifted weight (rather than, for instance, to effort).

The results show conclusively that the ability of human observers to judge lifted weight from the kinematics of the event is not an experimental artifact.

GENERAL DISCUSSION

Our experimental paradigm took a task that must commonly be performed with some accuracy and skill and made it more complicated than it would normally be. In mundane circumstances, there is a single (intrinsic) mapping problem in which an observer must relate observed lifted weights to his or her own lifting ability. Our magnitude estimation method introduced a second (extrinsic) mapping problem in which weight values in pounds had to be related both to observer capabilities and to lifter performance. This complication introduced the need to sort out variability associated with the extrinsic mapping task to establish performance levels for the intrinsic mapping task. We asked observers to judge lifted weight in patch-light displays both without and with the benefit of a preceding standard display. The difficulty was that such a standard might have affected both the intrinsic and extrinsic mapping tasks. To isolate the possible effect on the extrinsic mapping task, we included a condition in which observers tested a known weight before making judgments. This should have enabled observers to better understand how pounds relate to lifted weight without affecting their ability to evaluate the amounts of weight lifted in the displays.

Main Findings

Effect of No Pretest and No Standard. The greatest variation in relating pounds to amounts of lifted weight would be expected when observers were left to depend on their own background experience. This situation was investigated in Experiment 5 where lifted weight was judged with no standard and no pretest of a known weight. This did result in the greatest random errors with mean coefficients of variation of .51 and .57 for the two lifters, respectively.

Effect of Testing a Known Weight. Allowing observers to lift a known weight in advance of making judgments was expected to reduce the extrinsic mapping variability without affecting the ability to evaluate lifted weight as such. This condition should have yielded the best estimate of intrinsic mapping variability. When the haptic experience of a known weight was included in Experiment 5, the random error was reduced significantly with mean coefficients of variation of .35 and .39 for the two lifters, respectively. These coefficients must have included some residual extrinsic mapping variability as well as that generated by variations in skill across participants.

The addition of haptic experience relating the extrinsic unit to the observer's abilities also affected systematic error. Underestimation was changed to slight overestimation via a change in intercept with no change in slope. Without recent experience of known amounts of weight, observers simply seem to have been conservative in their weight estimates. Systematic errors with pretest of a

known weight should reflect the underlying intrinsic mapping accuracy achieved on the basis of visual information alone. Mean judgments were fairly precise for weights greater than 20 lb and overestimated weights 20 lb or less, as shown in Figure 6 and in the upper judgment curves of Figure 4.

Effect of a Standard. The second way in which the random error associated with the extrinsic mapping was reduced was via a standard. The effect of a standard was tested without haptic information in Experiment 4 in which the random error dropped to levels that were significantly below levels without the standard, but equivalent to levels achieved via the haptic information alone. The mean coefficient of variation in Experiment 4 with the standard was .38. A standard was equivalent to a test of a known weight in reducing random error.

However, the standard had a different effect on systematic errors than did the haptic information. Adding the standard affected both the slopes and the intercepts of judgment curves. The standard value seems to have acted as an attractor for all judgments, which accordingly exhibited a strong contraction effect. The result was an increase in overall systematic error.

Unlike the haptic information alone, the standard alone might have affected the intrinsic as well as the extrinsic mapping. An effect on the intrinsic mapping was possible for observers who had a good knowledge of their own abilities in terms of pounds of lifted weight. However, in Experiment 5, the variability associated with judgments made without haptic information and without a standard together with the significant effect of the haptic information in reducing that variability indicated that observers, in general, do not have adequate background knowledge for a standard to produce more than a very weak effect on the intrinsic mapping (i.e., without the pretest of a known weight).

Effect of a Pretest of a Known Weight Used as a Standard. When participants were given haptic experience of a given weight level which also was associated as a standard with a midrange weight lifted by the observed lifter, the random error was reduced to the lowest levels obtained overall. In the display-with-standard condition of Experiment 1, the mean coefficient of variation dropped to .28. This reduction was significant and about 25% lower than levels obtained with either haptic information or the standard alone. Presumably in this condition, the standard could indeed affect the intrinsic mapping because observers were given explicit information about their own lifting abilities in terms of the extrinsic unit, pounds. However, the contraction effect was produced by the standard in this condition as well, increasing systematic errors.

Summary. Both random and systematic errors were increased when lifted weight displays were judged without a standard display. However, as shown by the substantial decrease in both types of error with the inclusion of a pretest of

a known weight, most of the errors were produced by difficulties in relating an extrinsic scale of units to amounts of lifted weight. Reductions in random error were equivalent with inclusion of either a standard or a pretest of a known weight. However, systematic errors were actually increased by a standard due to the contraction effect. (The contraction effect was strengthened in our design which involved repeated judgments.) No contraction effect resulted from the pretest of a known weight. The implication of these results was that the original results of Runeson and Frykholm (1981, 1983) obtained using a standard display, were representative of a general ability to judge lifted weight. Runeson and Frykholm did not require repeated judgments, so their results would have been less subject to the contraction effect.

The Form of the Judgment Curves

Judgment curves were not truly linear. All exhibited a shallower slope at lighter weights and steeper slopes at medium to heavy weights. When medium to heavy weights were judged accurately, lighter weights were overestimated. The form of these curves simply reflects the nature of lifted weight. Lighter weights are of less consequence to a lifter and are thus less easily discriminated by an observer. What might this imply for the lifted-weight scale? An interval scale is obviously insufficient. Is a ratio scale possible with such difficulties in resolution around zero? In fact, the scale for lifted weight is quite different from the scale for weight as used in physics. Nevertheless, an absolute scale is required if information about lifted weight is to be useful in selecting and preparing modes of action and in preventing injury or harm.

We cited individual differences in the use of numbers as an expected source of variability in the extrinsic mapping task. More important, variability should be expected because of the differences between the scales for *lifted* weight and weight in general, differences that may have contributed to the initial oddity of the task. In physics, weight is bounded only at zero whereas lifted weight is bounded at both ends. The British scale of weight has no maximum value whereas lifted weight does for a given individual. Skill in the extrinsic mapping task must require the use of numbers to be tuned to the lifted-weight scale.

Although the natural anchor for a general weight scale is zero, anchoring lifted weight at zero would be a poor strategy because zero lifted weight is functionally and perceptually ill-defined. A person's limb segments have considerable mass. For instance, an adult hand weighs about a pound. Performing lifting movements empty handed is not to lift zero weight. The consequences of performing an empty-handed lifting movement 100 times are not discontinuous with those of lifting 0.5 kg or even 20 g 100 times. On the other hand, when estimating lifted weight, should we include a watch or a glove worn by the lifter? Surely not. When assessed functionally, the scale of values of lifted weight descends into a neighborhood of functionally equivalent zeros, a relatively

broad tolerance region. Despite this circumstance, lifted weight is meant strictly to refer to "the lifting of weight in addition to one's own weight." Thus, in principle, there is a zero. The continuity of functional consequences means that distinguishing the zero becomes extremely difficult.

Indeterminacy near the bottom of the range of sensitivity would be expected to yield considerable variability of intercepts. Given a floor effect for values near zero, a regular tendency for overestimation of low values would be expected. Indeed, the distributions generally are positively skewed for magnitude estimations of values near zero (Poulton, 1989; Slack, 1953; S. S. Stevens, 1956). Given heavier weights that are discriminated more easily and lighter weights that are discriminated only with effort and skill, the implication is that judgments should be anchored at maximum values. Furthermore, given potential skill variations in discriminating lighter weights in particular, variability in slopes and intercepts of otherwise well-formed individual judgment curves would be expected.

In all the experiments, the r^2 values for individual observers were higher than those for mean judgments. The mean individual r^2 in the display-only condition of Experiment 1 was .85 (SD = .10) and for each of the three lifters observed in Experiment 6, the mean individual r^2s were .77 (SD = .22), .80 (SD = .18), and .91 (SD = .07), respectively. The forms of the judgment curves were predominantly the same as the mean curves. Thus, observers were able to generate well-formed judgment curves that were subject to variability in determining intercepts and slopes.

Mean slopes were about .70 in all cases. The tendency for somewhat shallow slopes and high intercepts presumably was produced by readily detectable maximum values combined with poorly resolved minima that were overestimated as part of both a floor effect and a contraction toward a focal maximum value. This pattern was consistent with the results of Experiment 2 where, with improved information about lifter size, judgment curves rotated about the maximum value yielding higher slopes, lower intercepts, and less overestimation of lighter weights. The maximum value was clearly the anchor point.

The Scaling Problem

Whether and how variations in resolution might determine the nature of sensory scales have perhaps been the core issues in psychophysics (Gescheider & Bolanowski, 1991; Krantz, 1972; Luce, 1990, 1991; Luce & Edwards, 1958; Luce & Krumhansl, 1988; Poulton, 1989; S. S. Stevens, 1956). Related questions have been whether sensory scales should be considered to have zeros, whether sensory scales should be anchored at inevitable maximum values, and whether the latter is an alternative to the former (Birnbaum, 1980; Borg, 1962; Poulton, 1989; Veit, 1978). Fechner's (1860/1965) founding assumption was that scaling is determined by resolution, but this notion is currently held in low repute.

Scaling is known to be problematic near the ends of the dynamic range of the sensory systems where resolution deteriorates (Gescheider, 1976; Luce & Edwards, 1958; Luce & Krumhansl, 1988; Poulton, 1989). Because of this, Poulton (1989), among others (Birnbaum, 1980; Veit, 1978), suggested that an expectation of ratio scaling in psychophysical judgment should be abandoned in favor of interval scaling. Poulton argued at length that observers inherently perform judgments in terms of differences rather than ratios, at least when using unfamiliar units. He has implied, in addition, that accurate judgment along an absolute scale is possible with the use of familiar units. Unfortunately, how a transition might be achieved from interval scales (with arbitrary intercept) using unfamiliar units, to absolute scales (with zero intercept) using familiar units, has remained rather elusive in his account.

S. S. Stevens, in one of his foundational articles on ratio scaling and magnitude estimation, suggested that if an experimenter were to provide two standards or anchors, then an observer would be forced to perform judgment on an interval rather than a ratio scale (S. S. Stevens, 1956, p. 6). If this is the case, then the same must be true once an observer has established two values along a sensory continuum either by making two successive judgments or by making an initial judgment combined with a single standard provided by the experimenter. The implication is that observers are always forced to make judgments on an interval scale.[4] Accuracy of estimates in terms of an absolute scale would depend on the values of the anchors.

This line of thought naturally focuses attention on the anchors, because the accuracy of subsequent judgments would depend on them. The question would be: What determines the accuracy of the anchors? As long as zero is assumed to be a primary anchor from which the rest of the scale might be built, determination of the anchor value seems unproblematic. As soon as a nonzero anchor is assumed or required, the entire problem of scaling is reprised in determining the value of the anchor. Resolution may play a role in determining optimal anchor points. It may also contribute to a determination of the form of the judgment curves. However, resolution could not be responsible for establishing the value of an anchor.

In the visual perception of lifted weight, maximum-lifted-weight values appear to play the role of an anchor. How is the value of the maximum lifted weight established for a given lifter of given size and capability? How is the maximum weight value recognized and once recognized, how is its absolute value determined? This is the scaling problem.

[4]Instructions in magnitude estimation methods have often attempted to dissociate successive judgment values. See, for instance, M. Teghtsoonian (1965) who instructed her participants, who were judging the areas of circles, as follows: "Don't worry too much about being consistent or trying to remember what you assigned a circle before; just judge each one as it comes along." Whether these attempts are really successful is, of course, debatable.

The Role of the Standard and Lifter Size

The main question addressed by these investigations was whether a standard was necessary for judgments of accuracy comparable to those obtained in Runeson and Frykholm (1981). The slopes obtained in the display-only condition of Experiments 1 to 6 were all about .70. Using a standard, the slopes obtained by Runeson and Frykholm were .75 and .99 for their two lifters. The percentages of variance accounted for by weight in Experiment 6 and in the display-only condition of Experiment 1 were 49% and 46%, respectively—that is, about 20% less than the 68% reported by Runeson and Frykholm for their two lifters combined. The display-with-standard condition of Experiment 4 was comparable to Runeson and Frykholm's design for the absence of haptic test of a known weight. There the percentage of variance accounted for by weight was 46% which was also less than Runeson and Frykholm's figure by about 20%. Perhaps the Swedish observers tested by Runeson and Frykholm were more skilled in performing the task. In addition to being Swedish as opposed to American, they were 7 to 17 years older than the undergraduates in the current studies. From personal experience, both differences suggest that Runeson and Frykholm's observers might have approached the task with greater seriousness and concentration. However, comparisons of percentages of variance like this across experiments with somewhat different designs should be treated only as a rough means of comparison.

A second difference between our results was that slopes were lower in the display-with-standard condition in all cases (except in Experiment 2). The slopes were between .55 and .60. Providing the standard, with its emphasis on the midrange value, for a repeated set of judgments appears to have increased the strength of the contraction effect. This is consistent with current understanding of the contraction effect (Cross, 1973; Jones, 1986; King & Lockhead, 1981; Poulton, 1989; Slack, 1953; S. S. Stevens, 1956; R. Teghtsoonian & M. Teghtsoonian, 1978) although we could not find a comparable design and result in the literature.

Overall, we concluded that the standard was not essential to obtain results roughly comparable to those of Runeson and Frykholm, assuming that variability associated with the extrinsic mapping task has been adequately controlled. We also noted that in Experiment 2 with improved information about lifter size, observers' estimates were the most accurate and the most stable. Individual r^2 values were among the highest overall with a mean individual r^2 of .91 (SD = .06) in the display-only condition. The intriguing implication of this result when combined with the results of Experiments 5 and 6 was that observers were able to appreciate lifter size from the kinematics of the events captured in the displays. A new problem may be emerging. Can observers truly judge the size of a person from observation of their motions in a patch-light display? If so, what is the detectable kinematic property and the informational basis?

ACKNOWLEDGMENTS

This work was supported by a Summer Faculty Fellowship from Indiana University and by National Science Foundation Grant BNS-9020590.

I thank Sverker Runeson, Reinoud Bootsma, William Mace, David Rosenbaum, David Gilden, and three anonymous reviewers for their comments and very helpful advice.

REFERENCES

Bingham, G. P. (1987a). Dynamical systems and event perception. *Perception/Action Workshop Review, 2*(1), 1–14.

Bingham, G. P. (1987b). Kinematic form and scaling: Further investigations and the visual perception of lifted weight. *Journal of Experimental Psychology: Human Perception and Performance, 13*, 155–177.

Bingham, G. P. (1991). *The identification problem in visual event perception: Part I. Rate structures in optic flow and the degrees of freedom problem* (Cognitive Science Report Series No. 52). Bloomington: Indiana University.

Bingham, G. P. (1992). Leaning posture as information about lifted weight. Unpublished manuscript, Indiana University, Bloomington.

Birnbaum, M. H. (1980). Comparison of two theories of "ratio" and "difference" judgments. *Journal of Experimental Psychology: General, 109*, 304–319.

Borg, G. (1962). Physical performance and perceived exertion. *Psychologica et Paedagogica, 11*, 1–64.

Cross, D. V. (1973). Sequential dependencies and regression in psychophysical judgments. *Perception & Psychophysics, 14*, 547–552.

Fechner, G. T. (1965). Elements of psychophysics. In R. J. Hernstein & E. G. Boring (Eds.), *A source book in the history of experimental psychology* (pp. 66–75). Cambridge, MA: Harvard University Press. (Original work published 1860)

Gescheider, G. A. (1976). *Psychophysics: Method and theory*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Gescheider, G. A., & Bolanowski, S. J. (1991). Final comments on ratio scaling of psychological magnitude. In S. J. Bolanowski & G. A. Gescheider (Eds.), *Ratio scaling of psychological magnitude* (pp. 295–311). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Gibson, E. J. (1969). *Perceptual learning and development*. New York: Century-Crofts.

Gibson, J. J. (1986). *The ecological approach to visual perception*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc. (Original work published 1979)

Gilden, D. L., & Proffitt, D. R. (1989). Understanding collision dynamics. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 372–383.

Jones, L. A. (1986). Perception of force and weight: Theory and research. *Psychological Bulletin, 100*, 29–42.

King, M. C., & Lockhead, G. R. (1981). Response scales and sequential effects in judgment. *Perception & Psychophysics, 30*, 599–603.

Krantz, D. H. (1972). A theory of magnitude estimation and cross-modality matching. *Journal of Mathematical Psychology, 9*, 168–199.

Luce, R. D. (1990). "On the possible psychophysical Laws" revisited: Remarks on cross-modal matching. *Psychological Review, 97*, 66–77.

Luce, R. D. (1991). What is a ratio in ratio scaling? In S. J. Bolanowski & G. A. Gescheider (Eds.), *Ratio scaling of psychological magnitude* (pp. 8–18). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Luce, R. D., & Edwards, W. (1958). The derivation of subjective scales from just noticeable differences. *Psychological Review, 65,* 222–237.

Luce, R. D., & Krumhansl, C. L. (1988). Measurement, scaling, and psychophysics. In R. C. Atkinson, R. J. Herrnstein, G. Lindzey, & R. D. Luce (Eds.), *Stevens' handbook of experimental psychology* (2nd ed., pp. 3–74). New York: Wiley.

Pedhazur, E. (1982). *Multiple regression in behavioral analysis.* New York: Holt, Rinehart & Winston.

Poulton, E. C. (1979). Models for biases in judging sensory magnitude. *Psychological Bulletin, 86,* 777–803.

Poulton, E. C. (1989). *Bias in quantifying judgments.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Runeson, S. (1989, July). *An invariant-based model for the acquisition of perceptual skills.* Paper presented at the Fifth International Conference on Event Perception and Action, Oxford, OH.

Runeson, S., & Frykholm, G. (1981). Visual perception of lifted weight. *Journal of Experimental Psychology: Human Perception and Performance, 7,* 733–740.

Runeson, S., & Frykholm, G. (1983). Kinematic specification of dynamics as an informational basis for person and action perception: Expectations, gender recognition, and deceptive intention. *Journal of Experimental Psychology: General, 112,* 585–615.

Runeson, S., & Vedeler, D. (in press). The indispensability of precollision kinematics in the visual perception of relative mass. *Perception & Psychophysics.*

Slack, C. W. (1953). Some characteristics of the "range effect." *Journal of Experimental Psychology, 46,* 76–80.

Stevens, J. C. (1958). Stimulus spacing and the judgment of loudness. *Journal of Experimental Psychology, 56,* 246–250.

Stevens, S. S. (1956). The direct estimation of sensory magnitudes—loudness. *American Journal of Psychology, 69,* 1–25.

Stevens, S. S., & Greenbaum, H. B. (1966). Regression effect in psychophysical judgment. *Perception & Psychophysics, 1,* 439–446.

Teghtsoonian, M. (1965). The judgment of size. *American Journal of Psychology, 78,* 392–402.

Teghtsoonian, R., & Teghtsoonian, M. (1978). Range and regression effects in magnitude scaling. *Perception & Psychophysics, 24,* 305–314.

Veit, C. T. (1978). Ratio and subtractive processes in psychophysical judgment. *Journal of Experimental Psychology: General, 107,* 81–107.