COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 45 (2021) e12940

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12940

Instructed Hand Movements Affect Students' Learning of an Abstract Concept From Video

Icy (Yunyi) Zhang,[a] Karen B. Givvin,[a] Jeffrey M. Sipple,[a]
Ji Y. Son,[b] James W. Stigler[a]

[a]*University of California, Los Angeles*
[b]*California State University, Los Angeles*

Received 13 December 2019; received in revised form 13 October 2020; accepted 22 December 2020

Abstract

Producing content-related gestures has been found to impact students' learning, whether such gestures are spontaneously generated by the learner in the course of problem-solving, or participants are instructed to pose based on experimenter instructions during problem-solving and word learning. Few studies, however, have investigated the effect of (a) performing instructed gestures while learning concepts or (b) producing gestures without there being an implied connection between the gestures and the concepts being learned. The two studies reported here investigate the impact of instructed hand movements on students' subsequent understanding of a concept. Students were asked to watch an instructional video—focused on the concept of *statistical model*—three times. Two experimental groups were given a secondary task to perform while watching the video, which involved moving their hands to mimic the placement and orientation of red rectangular bars overlaid on the video. Students were told that the focus of the study was multitasking, and that the instructed hand movements were unrelated to the material being learned. In the *content-match* group the placement of the hands reinforced the concept being explained, and in the *content-mismatch* group it did not. A *control* group was not asked to perform a secondary task. In both studies, findings indicate that students in the content-match group performed better on the posttest, and showed less variation in performance, than did students in the content-mismatch group, with control students falling in between. Instructed hand movement—even when presented as an unrelated, secondary task—can affect students' learning of a complex concept.

Keywords: Instructed hand movements; Gesture; Video-learning; Statistics learning; Embodied cognition

Correspondence should be sent to Icy (Yunyi) Zhang, UCLA Psychology Department, Los Angeles, CA 90095-1563. E-mail: yunyi9847@g.ucla.edu

1. Introduction

In recent years, we have witnessed an increased interest in the role of gesture in teaching and learning. In classrooms, gesture is commonly and spontaneously used by both teachers and students as they carry out their educational activities (Alibali & Nathan, 2011; Kontra, Goldin-Meadow, Beilock, & Sian, 2012; Novack & Goldin-Meadow, 2015; Richland, Stigler, & Holyoak, 2012). Although gesture can serve many functions—with the most commonly studied being its use during problem-solving—our interest is in the role that gesture plays in the development of conceptual understanding in complex academic domains.

Kita, Alibali, and Chu (2017), in an extensive review of the literature, cite a number of studies in which spontaneous gestures have been found to impact learning. For example, simply encouraging students to gesture while explaining their incorrect solutions to a problem can facilitate learning (Broaders, Cook, Mitchell, & Goldin-Meadow, 2007) and lead to the emergence of new ideas (Novack & Goldin-Meadow, 2015). Likewise, preventing students from gesturing while problem-solving can hold them back compared with students who are free to use their hands while they think (Alibali & Kita, 2010; Alibali, Spencer, Knox, & Kita, 2011). Even just watching gestures in an instructional video can impact students' learning (Rueckert, Church, Avila, & Trejo, 2017; Son, Ramos, DeWolf, Loftus, & Stigler, 2018).

It is not only spontaneously generated gestures that play a role in learning. Instructed gestures—that is, those generated not spontaneously, but based on specific instructions—can affect learning as well. For example, instructing participants to imitate content-related gestures used by a presenter in an instructional video improved learners' long-term recall of anatomical structure names and locations compared with participants not instructed to imitate gestures (Cherdieu, Palombi, Gerber, Troccaz, & Rochet-Capellan, 2017). And students who were instructed to gesture about a strategy for solving a series of math problems retained more of what they had learned a month later than students not instructed to gesture (Cook, Mitchell, & Goldin-Meadow, 2008).

A growing body of evidence reveals that it is not just the general movement of the hands, but the content of the gesture that is critical for learning. In one striking example, participants instructed to perform a specific gesture designed to activate a correct problem-solving strategy outperformed participants instructed to perform a similar gesture representing a strategy that was only partially correct (Goldin-Meadow, Cook, & Mitchell, 2009). Similarly, learners instructed to perform specific iconic gestures while repeating novel words retained significantly more verbal material over time than learners instructed to perform meaningless gestures (Krönke, Mueller, Friederici, & Obrig, 2012; Macedonia, Mueller, & Friederici, 2010).

In most studies, instructed gestures are accompanied by speech and designed to serve a direct, communicative purpose. However, some studies have shown that simply telling learners to move their hands in a certain meaningful way—without accompanying speech, or knowing the purpose of the hand movement, or even both—can produce an effect on learning. Brooks and Goldin-Meadow (2016) showed children a series of math problems

and asked them to move their hands in a fashion that was either relevant or irrelevant to a strategy for solving the problem. They were not asked to solve the problem, only to move their hands in prescribed ways. Children whose instructed hand movements represented a correct strategy benefited more from subsequent instruction on how to solve the problems than children instructed to hold their hands in irrelevant positions.

Similar findings have been shown in the embodied cognition literature. A growing number of studies have indicated that people understand and process information through their physical movements, with those movements not only reflecting their cognition, but also informing it (Dove, 2018; Walkington, Chelule, Woods, & Nathan, 2019; Wilson, 2002). Some studies have found a relationship between directed physical actions (e.g., posing) and participants' interpretation and meaning-making. Both prompted and unprompted posing (e.g., raising one's arm to mimic a figure in a painting) helped with participants' meaning-making of the art piece (Steier, 2014). Not only can physical actions help the interpretive process; they can sometimes even *alter* the interpretive process. Smith (2005) found that toddlers who were given an object and asked to move it up and down vertically judged the object they were holding to be more similar to a similar but taller object than to a similar but wider object. (Toddlers asked to move the object back and forth horizontally showed the opposite result.)

Two studies, in particular, bridge the embodied cognition literature and the gesture literature. Thomas and Lleras (2009) found that participants instructed to perform arm movement exercises related to a strategy for solving Maier's two-string problem were later more successful in solving the problem than were participants completing different arm movement exercises. Similarly, Nathan et al. (2014) found that participants who were instructed to perform arm movements related to a particular strategy for solving of a mathematical task were more likely to develop key insights into the problem than were participants who were asked to perform random arm movements.

In the research discussed so far, the instructed gestures studied were related to solution strategies for specific kinds of problems, and they often served a representational or communicative purpose. But our focus here is a little different. Instead of focusing on problem-solving as the outcome, we are interested in the role that instructed hand movements can play in improving students' conceptual understanding in complex domains. In this sense, our study is similar to that of Ping and Goldin-Meadow (2008), who investigated the role teachers' gestures played in students developing understanding of the Piagetian concept of conservation. They found that instructions that included gestures representing important dimensions of the context—in this case gestures that described the shape of a glass—resulted in children learning more about conservation than children in a comparison group who received only spoken instructions. These results demonstrate the potential of gesture to facilitate abstract thinking and understanding, in addition to specific problem solution strategies.

The research we report here is in the spirit of Ping and Goldin-Meadow. However, instead of having students observe gestures delivered as part of verbal instructions, we direct students themselves to move their hands in certain ways, and we investigate the role such hand movements play in students' developing understanding of the concept of

statistical model. In addition, the hand movements students are asked to perform in the current study do not serve any immediate communicative purpose for the learner; indeed, learners are told that the instructed hand movements are extraneous to their comprehension of the concept, and only included as part of a multitasking study. Thus, the present study will look at the effect of directing students to move their hands in ways that do and do not map onto crucial features of an abstract concept, but without drawing participants' attention to the mapping, or lack thereof.

1.1. The concept of a statistical model

The current work focuses on students' learning of introductory statistics, a content area with many abstract concepts and which often requires some scaffolding to expose underlying structures and contexts to students. Although students learn many statistical concepts and procedures in typical introductory courses, they often have trouble seeing the structure that ties everything together and, as a consequence, have difficulty transferring their knowledge to new situations. Our interest is in how to leave students with a coherent representation of the domain, and in the possible benefits of students using their hands as an additional resource to support understanding (Fries, Son, Givvin, & Stigler, 2020; Son et al., 2018; Stigler et al., 2020).

In the studies reported here we focus on the concept of *statistical model* as a core concept that can be connected to all parts of an introductory statistics course. Although most people get introduced to the concept of modeling when they take graduate-level statistics, the concept is rarely taught in the introductory undergraduate course. The current studies are situated within a broader project in which we are exploring what it would look like to teach everything in connection with modeling, and then to understand the effects of such an approach on students' understanding and transfer (Stigler et al., 2020; Son, Blake, Fries, & Stigler, 2020).

When we first introduce students to the arithmetic mean, we teach it as the simplest of all statistical models. All of our students know how to calculate a mean, but they never have thought of it in terms of modeling. A statistical model, in simplest terms, is a function that can be used to generate a prediction about a future observation. If we randomly choose one more observation from a population, our best prediction of what it will be, assuming that we want to reduce error and that we know nothing else, is the mean. As soon as we cast the mean as a statistical model, we can see variation around the mean as error, and we therefore introduce concepts such as Sum of Squares and Standard Deviation as statistics used to quantify the amount of error around a model. From that point, we connect everything to the overarching idea that DATA = MODEL + ERROR.

In the instructional video students watched in the current studies, we introduced them to the concept of statistical model (as outlined in the previous paragraph), and then proceeded to develop a slightly more complex model, which we called the two-group model. Using the two-group model, we based our prediction of a future observation on the group from which the observation would be drawn (i.e., from one of two levels of a categorical predictor variable). If we are looking at thumb length in millimeters as a function of sex,

for example, our two-group model would simply predict the mean of females for a newly selected female, and the mean of males for a newly selected male.

In the video, we present distributions using histograms, helping students to see where, in a histogram, they can see the model, and where they can see the error. We chose histograms because we know from other research that students often misinterpret the meaning of the vertical axis on a histogram (delMas, Garfield, & Ooms, 2005). However, the vertical axis on a histogram actually represents the number of observations falling within a particular range of scores, a common misconception among students is that it represents the value of scores on the outcome measure being graphed. This prior research helped us to think about the kinds of gestures that might be useful for helping students overcome their misconceptions by directing hand movements to certain features that are important to conceptual understanding.

1.2. Overview of the study

In the two studies reported here, we asked whether instructing students to place and move their hands in specific ways while watching a video can impact their understanding of, and learning from, the video. The video is distinctive in that it does not intend to teach students to solve a particular kind of problem, but instead was designed only to increase understanding of a concept, in this case the concept of statistical model, as described above. We wanted to explore hand movement without any relation to either speech production or the goal of solving a particular type of problem. Our outcome measures focused on understanding the concept of statistical model and extending the concept to a new situation.

Participants were brought into the laboratory and asked to watch an instructional video. They watched the video three times and were told at the outset that they would be tested on the content at the end. Participants were also told that the goal of the study was to assess the effect of multitasking on learning, so as to obfuscate its actual intent. Students in the experimental conditions were asked to perform a secondary task while watching the video. That task involved placing their hands in front of the video in alignment with one or two red, rectangular bars that appeared superimposed on the video and moving their hands to follow the bars as they moved around the screen.

In one condition (the *content-match* condition), the red bars led participants to place their hands in a way that could support a connection between the concept being explained and the histogram being shown in the video. For example, when the video talked about the distribution having a larger variation, the content-match condition would first place their hands in the middle of the distribution and then simultaneously move their left hand to the left and right hand to the right, representing the idea of variation. In a second condition (the *content-mismatch* condition), the red bars were placed in a configuration opposite to that of the content-match condition, thus leading participants to place their hands in a way that was not supportive of the concept being explained. For example, during the discussion of the variation in a distribution described above, the content-mismatch group moved their hands up and down instead of left and right. Finally, in a third, *control*,

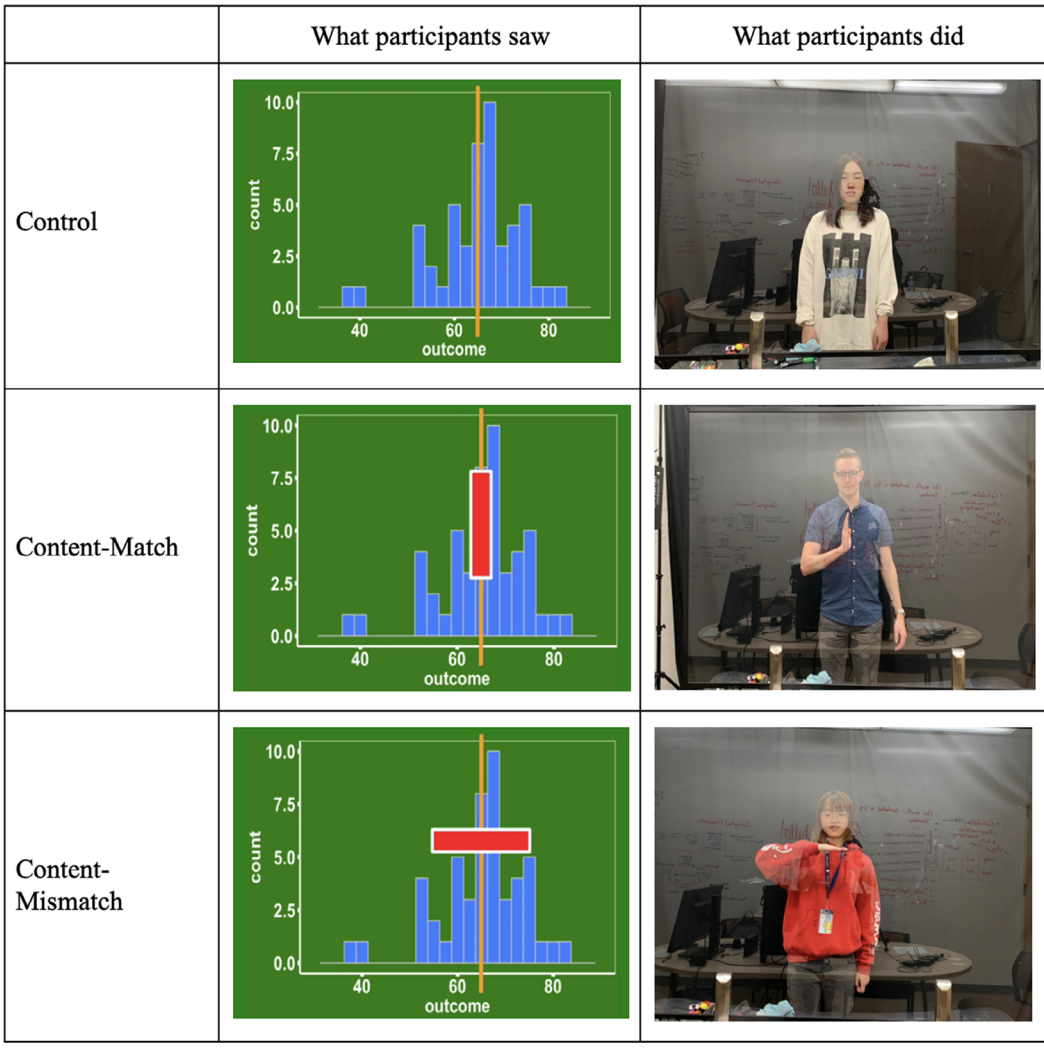

Fig. 1. Example frames from the instructional video discussing the mean of a distribution (left), and the associated instructed hand movements, for control, content-match, and content-mismatch conditions.

condition, the video did not include red bars at all, and students were not given any instructions about what to do with their hands. (See Fig. 1, e.g., frames from the video as it appeared in the three conditions.)

If participants' instructed hand movements serve only to orient their attention to the video in a *general* way, then we would expect no difference between the content-match and content-mismatch conditions, and students in both of these conditions would be expected to learn more than those in the control condition simply because their hands are directing their attention toward the slides in the video. However, if the *specific* orientation of the hands matters—that is, if we have succeeded in designing instructed hand

movements to help overcome students' possible misconceptions—then we would expect more learning from students in the content-match condition than in the content-mismatch condition. Students in the content-mismatch group would perform worse than those in the control group to the extent that the prescribed hand movements in that group reinforce their misconceptions.

If, on the other hand, students must perceive a representational or communicative purpose for their instructed hand movements or if the hand movements have to activate a problem solution strategy in order to affect learning, then we would expect students in the control condition to learn just as much as students in the experimental conditions. In fact, control students might learn more than those in the experimental conditions to the extent that instructed hand movement without a clear purpose truly does amount to multitasking, taking resources away from the primary task of understanding the video. If we observe that the content-match group outperforms both the control group and the content-mismatch group, it will suggest that hand movements are not only useful as a means of suggesting a solution strategy in the context of problem-solving, but also can provide an important resource to help with the construction of understanding. This will, in turn, shed light on how instructed hand movements might be used in the teaching and learning of abstract ideas in complex domains.

2. Study 1

2.1. Methods

2.1.1. Participants

The participants were 60 undergraduate students from the University of California, Los Angeles (UCLA). They were a diverse sample, with an ethnic composition similar to that of UCLA as a whole (28% Asian, 27% White, 22% Latino, 3% Black). They participated in the experiment either to fulfill a course requirement or to get extra credit, and they did not receive any other form of compensation for their participation. The only exclusion criterion was that they could not have taken Psychological Statistics with specific professors because the content presented in the study was taught in those classes. All participants had normal or adjusted-to-normal vision and were able to stand for 40 min without impairment. They were all fluent in English and able to understand the instructions clearly.

2.1.2. Materials and procedure

Procedure: Participants were randomly assigned to one of the three conditions (control, content match, or content mismatch). After participants signed the consent form, they were asked to stand facing a monitor, above which was a camera, which recorded them from the waist up for the duration of the study. Participants read the instructions on the monitor that introduced the study's purpose: how multitasking affects learning from

instructional video. Participants were instructed to try their best to understand the contents of the video because there would be a quiz at the end. They were additionally told that they may or may not be asked to engage in a secondary task while watching the video (i.e., the multitasking part of the study).

After reading the initial instructions, participants in both the content-match and content-mismatch conditions were trained in how to perform the instructed hand movement task. While watching the monitor, participants were shown a series of four slides on which there appeared either one or two red bars. Participants were asked to align their hands to match the placement of the red bars, moving their hands each time the bars moved to a different position. In the control condition, participants did not receive any hand-placement training. Participants in all three groups then watched the instructional video three times in a row, either with or without the hand movement task (as prescribed by their experimental condition).

Video: The video was 7 min and 10 s in length and consisted of an audio presentation with accompanying slides introducing the concept of statistical model. The video started with a simple one-group model (i.e., using the mean as a model) and then moved on to introduce a two-group model. The audio and slides were identical across the three conditions except for the red bars, which indicated, for the two experimental groups, where participants should place their hands. Fig. 1 shows an example frame from the video as it appeared in the three conditions, along with photos showing examples of how participants were instructed to place their hands as they watched the video. In the experimental conditions, the red bars appeared at 18 time points during the span of the video.

Decisions about the timing of the red bars and where to place them in the content-match condition were guided by observing the gestures produced by the presenter when initially recording the audio and by research on how students understand and interpret histograms (delMas et al., 2005). When the presenter used a gesture to connect a graph with a concept—such as holding her hand in a vertical position to indicate the middle of the distribution pictured in Fig. 1, or moving her hands from the middle to the left and right to represent variation—we inserted a red bar that would place the participants' hand in approximately the same position (see middle panel of Fig. 1). Thus, during the study, participants would be placing and moving their hands in roughly the same position that the presenter had used when creating the audio stimulus.

Red bars in the content-mismatch condition were placed in a contradictory position (as in the lower panel of Fig. 1), close in location but opposite in orientation to the bar in the content-match condition. (All three versions of the video are available online: http://bit.ly/2BIyOPL; a transcript of the audio portion of the video is included in Appendix A.) Because the position of the red bars depends on the content being discussed, we did not control for the particular orientation of the hands. Although the hand(s) was/were always paused in a vertical orientation in the content-match condition and in a horizontal orientation in the content-mismatch condition, the direction in which the hands moved to follow the red bars in the video varied. In both conditions, participants moved their hands in both vertical and horizontal directions.

2.1.3. Posttest measures

After watching the video for the third time, participants answered, verbally, a set of nine questions (five multiple-choice and four free response; see the complete list of posttest questions in the Supplementary Material file). The questions were presented, one at a time, on the monitor, and in the same order for each participant. The questions either asked students about the concepts discussed in the video or required participants to apply what they had learned in the video to a new situation. For each of the multiple-choice questions, participants were asked not only to select the correct answer but also to explain their selection.

Participants were videotaped throughout the experiment from a point behind the monitor. Their answers to posttest questions, as well as their instructed hand movements and gestures, were coded from the video by four trained coders. The coding of the gestures was conducted separately from the coding of posttests, and coders remained blind to condition. For each multiple-choice question, coders recorded both the correctness of the participant's choice and the quality of their explanation. Coders judged the quality of each multiple-choice explanation and free response answer using a 3-point scale: 2 points for responses that were both relevant and correct; 1 point for responses that were relevant but only partially correct; or 0 points for responses that were neither relevant nor correct. At least two coders coded each participant's video, with coding disagreements discussed by the four coders in weekly meetings until a consensus was reached.

Each participant was assigned four outcome scores based on their responses to the posttest questions:

- **Number correct on multiple choice (MC Correct):** One point for each of the five questions answered correctly; possible range from 0 (none correct) to 5 (all correct).
- **Quality of explanations for multiple choice (MC Explain):** Score of 0 to 2 for each of the five explanations; possible range from 0 to 10.
- **Quality of answers to free response questions (Free Response):** Score of 0 to 2 for each of the four free response questions; possible range from 0 to 8.
- **Total score on posttest questions:** Sum of each of the previous three subscores; possible range from 0 to 23.

2.2. Results

2.2.1. Effect of condition on posttest performance

The distributions of total posttest scores for the three conditions are presented in Fig. 2. Participants in the content-match group scored higher than those in the other two groups. A one-way ANOVA found a significant effect of condition on total score ($F(2, 56) = 4.56$, $p = .015$). Pairwise contrasts showed that the content-match group scored significantly higher than the content-mismatch group ($t = 3.12$, $p = .003$) but not significantly higher than the control group ($t = 1.83$, $p = .076$). There was no significant difference between the control group and the mismatch group ($t = 1.22$, $p = .231$).

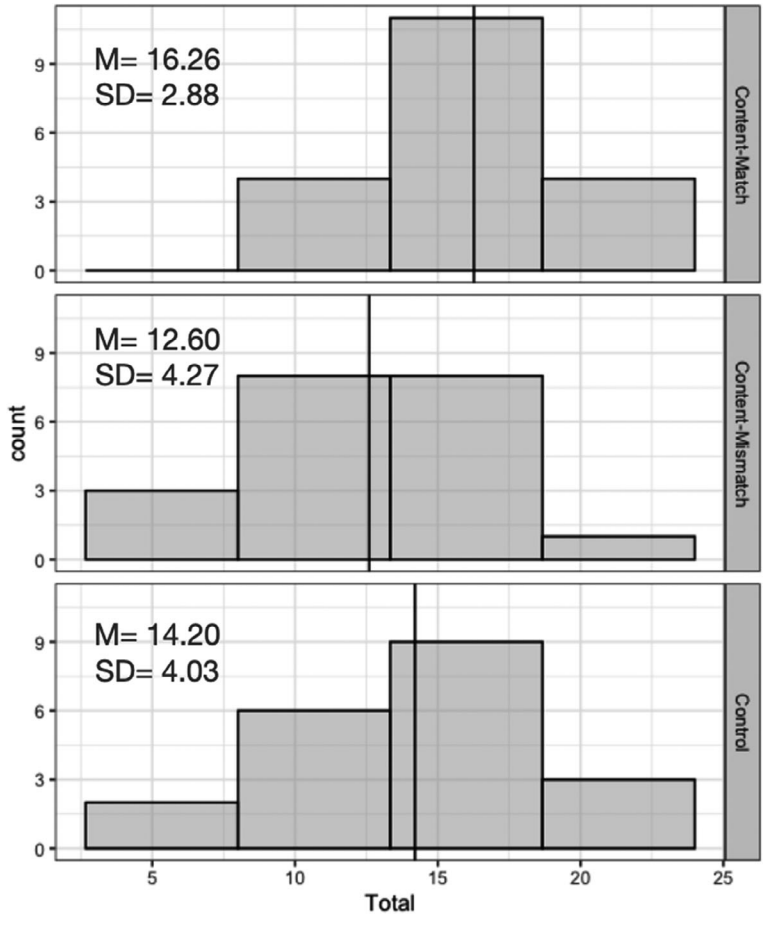

Fig. 2. Distribution of total scores, by group.

We next analyzed the three component subscores separately in order to better understand the source of the effect. In general, the pattern of results looked similar across the three subscores: Content-match scores were highest, followed by control, with content mismatch the lowest, in all cases (see Table 1). Overall *F*s were significant for MC Correct and Free Response, but not for MC Explain. Pairwise contrasts for both MC Correct and Free Response showed a significant advantage of content match over content mismatch. No other effects were statistically significant.

2.2.2. Variance in posttest performance across conditions

It appears from the histograms that scores in the content-match group were less variable than those in the other two groups. Further analysis showed this pattern in both the MC Correct and MC Explain subscores, but not in the Free Response scores (see Table 1). Levene's test for equality of variance showed this difference in variance to be statistically significant for MC Correct overall ($F(2, 56) = 3.28$, $p = .045$). Pairwise

Table 1
Posttest performance for overall score and three component subscores

Measure	Control		Content Match		Content Mismatch		Overall ANOVA		Match Versus Mismatch	
	M	*SD*	*M*	*SD*	*M*	*SD*	*F*	*p*	*t*	*p*
Total score	14.20	4.03	16.26	2.88	12.60	4.27	4.56	.015	3.12	.003
MC correct	3.95	1.00	4.16	0.60	3.40	1.05	3.65	.032	2.79	.009
MC explain	6.50	2.24	7.32	1.38	6.05	2.31	1.95	.153		
Free response	3.75	1.86	4.79	1.87	3.15	1.69	4.08	.022	2.87	.007

comparisons showed variance within the content-match group was significantly lower than within the content-mismatch group ($F(1, 38) = 7.12$, $p = .011$). Other pairwise differences were not statistically significant: Content match versus control, $F(1, 38) = 3.89$, $p = .056$; content mismatch versus control, $F(1, 39) = 0.24$, $p = .631$. For MC Explain scores, Levene's test for equality of variance showed the variance of the content-match group to be significantly lower than both the control group ($F(1, 38) = 4.21$, $p = .047$) and the content-mismatch group ($F(1, 38) = 4.59$, $p = .039$). There was no significant difference between the variance of the content-mismatch group and that of the control group ($F(1, 38) = 0.02$, $p = .903$). There was also no significant difference of variance across conditions for participants' performance on free response questions ($F(1, 38) = 0.42$, $p = .659$).

2.2.3. Gesture production during posttest

Participants answered all posttest questions verbally while standing in front of the monitor on which the questions were presented. Although participants were not given any instructions regarding the use of their hands during this phase of the study, most participants did produce spontaneous gestures while answering the questions. We wondered if the previous instructions regarding instructed hand movements would affect the quantity of gestures produced by participants while answering the posttest questions.

Fig. 3 shows the number of gestures participants produced while answering the questions. Participants in the control group produced the fewest gestures. A one-way ANOVA found a significant difference across the three conditions in the number of gestures participants produced ($F(2, 56) = 5.17$, $p = .006$). Pairwise contrasts showed that both the content-match group and the content-mismatch group performed significantly more gestures while answering the questions than did the control group ($t = 2.35$, $p = .024$ and $t = 3.26$, $p = .002$, respectively). There was no significant difference between the content-match and content-mismatch groups ($t = -.62$, $p = .536$).

2.3. Discussion of Study 1

Significant group differences were found for both multiple-choice and free response questions. Participants in the content-match group scored significantly higher than those

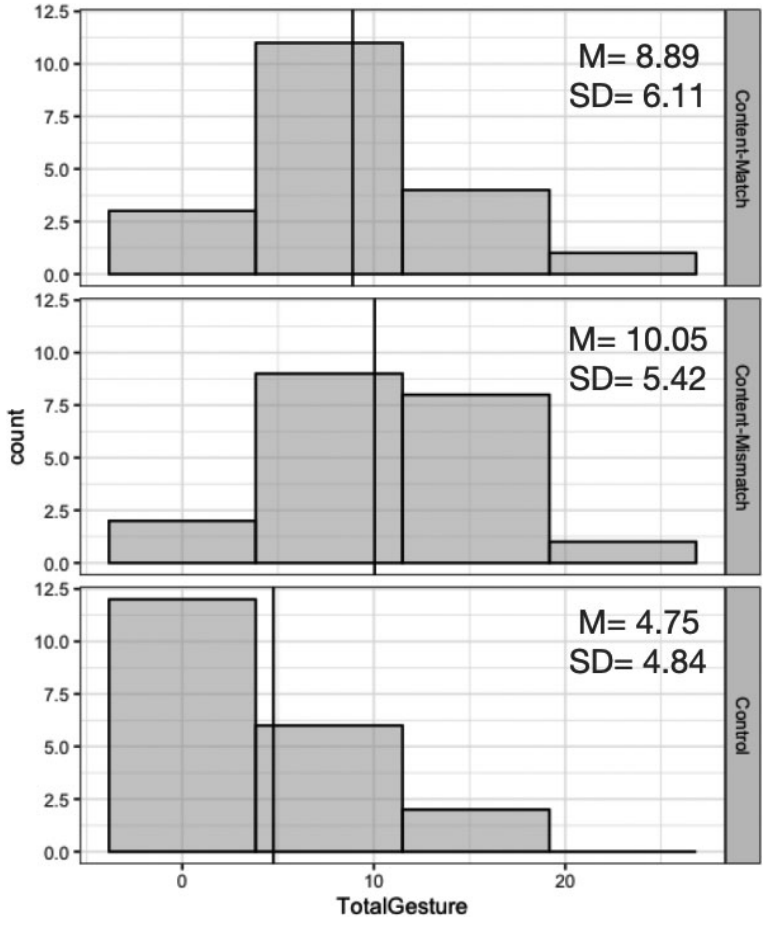

Fig. 3. Number of gestures produced during posttest, by group.

in the content-mismatch group, with control participants falling between the other two groups. In general, this pattern of results supports the interpretation that specific hand orientation matters, even when students are not given a purpose for their instructed hand movements. If the role of the hands was merely to draw attention to the slides, then content-mismatching placements should be as effective as the content-matching ones, but this clearly was not the case. If the hand movement task interfered with learning (i.e., functioning as a secondary task), then the control group should have scored higher. Again, this was not what happened.

Something we did not anticipate, but which we find interesting, is the lower variation observed in the content-match condition compared with the other two conditions for both multiple-choice questions and explanations. We believe that this reduced variation among the content-match group represents an asymmetry in the effect of matching gestures on students' thinking. If a student is already understanding the concept being explained, then

they are not further helped by the matching gesture. But if they are struggling to understand, then the gesture helps.

In this particular case, a student who has a correct understanding of the meaning of the vertical axis of a histogram might not be distracted by a gesture that seems inconsistent with their understanding. But a student who is struggling with this concept might need the support of a matching gesture to nudge them toward a correct interpretation. If this interpretation were correct, the reduction in variation in the content-match group would result from the bottom students being brought up, not the top students being brought down.

It is worth noting here that an educational intervention that simply asks students to place and move their hands in certain orientations during instruction—even if the movement makes no sense to them *yet*—could constitute a small step toward the elusive goal of reducing gaps between high and low achievers, especially in STEM fields. Although this tiny intervention might have no more than a limited positive impact on the learners at the high end of the distribution, it does appear to help the struggling students. Indeed, OECD (2012) has suggested that the reduction of such gaps, while improving achievement overall, should be a core part of our definition of what it means to be a high-performing educational system.

Students in both the content-match and content-mismatch groups gestured more while answering the posttest questions than did students in the control group. This made sense to us, given the lack of mention of hand movements in the instructions provided to control participants. A possible mechanism—multimodal gist representation—was discussed by Galati and Samuel (2011). They found that participants were more likely to encode a target action required by the study in their own gestures if they had seen a gesture about the action before, and more so if the gesture they saw was congruent than if it was incongruent. (It is worth noting that in our study, both content-match and content-mismatch conditions produced more gestures than the control.) They speculated that the mechanism behind this may be that students activated and relied on their multimodal gist representation of the video content in their gesture because they had seen gestures when those representations were formed.

In our study, participants performed gestures rather than observing them. However, because performing gestures has been shown to be more powerful than simply observing gestures (Wakefield, Hall, James, & Goldin-Meadow, 2018), the same mechanism may hold (or be even stronger). Because students in the two experimental conditions in our study performed gestures during encoding, they may have relied more on their multimodal gist representation of the content than students who did not perform any gestures. This mechanism helps to explain the surprising finding that some of the students in the content-mismatch group actually produced content-mismatching gestures—the gestures they had been asked to mimic in their instructed hand movements—when answering the posttest questions, even though such gestures worked against their understanding of the concept. For example, they moved their hands up and down vertically when talking about variation within a group, consistent with the aforementioned misconception about the

meaning of the vertical axis. They might have been relying on their multimodal gist representation of the video content, however, erroneous.

Participants in the content-mismatch group seemed to raise questions at the end of the experiment more often than did students in the other two groups, although we did not formally measure this. In particular, they asked questions to try to make sense of the instructed hand movements they were instructed to make, trying to figure out how they related to the content of the video. Even though these students were explicitly told that the instructed hand movements were a secondary task, they may nonetheless have believed that they were related to the video. Thus, aside from the multimodal gist representation hypothesis, it is also possible that the higher number of gestures produced by the content-mismatch group while answering the posttest questions may have resulted from an attempt to integrate the instructed hand movements they had made during the experiment with their understanding of the concept of statistical model because they believed the movements to be related to the content. These all revealed to us the importance of having a posttest question that measures participants' perceptions of the purpose of the study—an omission we remedied in Study 2.

3. Study 2

The primary aim of Study 2 was to replicate the posttest results with a larger sample. It was our hope that having greater power would enable us to differentiate the control from one (or both) of the experimental groups. We also wanted to ask students about their prior experience with statistics, and afterwards, their perceptions about our intervention. Finally, we wanted to revise the protocol to one that was easier to run, requiring less input from an experimenter. So, instead of participants standing in front of a video camera, the study was automated, with participants interacting one-on-one with a laptop on which materials were presented.

3.1. Methods

3.1.1. Participants

The participants were 148 undergraduate students from the University of California, Los Angeles (UCLA). (A power analysis using standard errors from Study 1 indicated that we could detect an effect size of 0.3 on posttest scores, with power of 0.90, with sample sizes of $n = 48$ in each group.) As with Study 1, their ethnicity was representative of the UCLA population as a whole. Participants who had taken Psychological Statistics with specific professors were excluded because the content presented in the study was taught by those professors. We also excluded students who had participated in Study 1. All participants were fluent in English and able to use a laptop to complete the study.

3.1.2. Materials and procedure

Participants were randomly assigned to one of the three conditions (control, content match, or content mismatch). Six laptops were setup in the laboratory, with two laptops designated for each condition. The study was hosted on Qualtrics and administered completely through the laptops. After participants entered the laboratory, they were instructed to sit in front of a laptop and to remain seated throughout the study. Similar to Study 1, students read the instructions on the laptop.

After answering the presurvey, participants in the two experimental conditions completed a brief training video in which they practiced moving their hands to match the red bars that appeared on the laptop screen, as described in Study 1. The training in Study 2 included additional written instructions in the video to replace instructions provided by the experimenter in Study 1 (e.g., telling participants they could use either hand in segments where there was only one red bar). Participants in all three groups then watched the instructional video three times in accordance with their assigned condition.

As in Study 1, participants in the two experimental groups engaged in the hand movement task while watching the video. The videos and instructed hand movements were identical to those used in Study 1. (See Fig. 4 for an image of the laptop station setup.) After each viewing of the video in Study 2, participants were asked to rate how much of the video they understood. Participants wore headphones and barriers were placed between them to minimize distractions.

3.1.3. Posttest measures

After watching the video three times, participants completed a posttest. The posttest contained the same five multiple-choice questions as in Study 1. (Because in Study 1

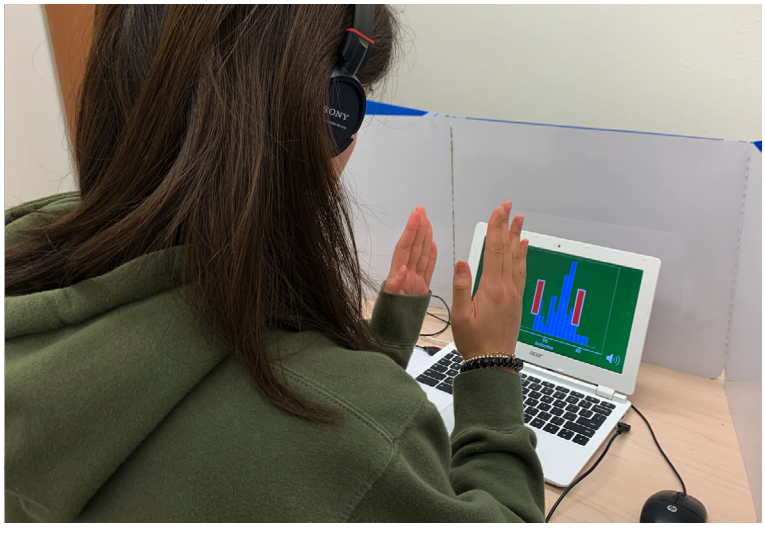

Fig. 4. The experimental setup for Study 2. A student is watching video (slides plus audio) on a laptop screen, while holding her hands to match the red bars.

results from the free response posttest questions were consistent with those from the multiple-choice questions, we used only the multiple-choice questions in Study 2.) Participants saw the questions one at a time and were not able to go back to change their answer once they proceeded to the next question.

After completing the five multiple-choice questions, participants were asked, in an open response question, what they thought was the purpose of the study. Finally, participants in the two experimental groups received one additional question asking them, in open-response format, whether they thought the instructed hand movements were helpful for their learning of the content.

Eighteen participants were excluded from the study because they either failed to pay attention to the video (e.g., they fell asleep or appeared distracted), or because they did not follow the study instructions correctly. The latter situation included such behaviors as not performing the requested hand movements, not performing the hand movements correctly, or watching the video more or fewer than three times.

Multiple-choice questions were auto-scored by Qualtrics and yielded a total number correct from 0 to 5 for each student. Answers to the question of what the students believed the purpose of the study to be were coded into one of four mutually exclusive categories: how repeated video watching influences learning or memory; how multitasking (while watching video) relates to learning; how hand movements, or physical actions influence learning; and other. Answers by students in the two experimental groups to the question of whether they felt the instructed hand movements were helpful to their learning were coded as either helpful or unhelpful. Open response questions were coded by a single coder, blind to condition.

3.2. Results

3.2.1. Posttest scores

The distributions of posttest scores for the three conditions are presented in the right panel of Fig. 5. (Study 1 results are presented in the left panel, for purposes of comparison.) Participants in the content-match group scored higher than those in the other two groups. A one-way ANOVA found a significant effect of condition on posttest score ($F(2, 127) = 5.67$, $p = .004$). Pairwise contrasts showed that the content-match group scored significantly higher than both the content-mismatch group ($t = 3.37$, $p = .001$) and the control group ($t = 2.54$, $p = .013$). There was no significant difference between the control group and the content-mismatch group ($t = 0.83$, $p = .407$).

The distribution of posttest scores looks highly similar across the two studies, but distinctly different across the experimental conditions. The similarities in distributions across the two studies are particularly noteworthy because of the difference in experimental procedures between the two studies. In Study 2, students were sitting down instead of standing up, and holding their hands over a laptop screen instead of in front of a larger monitor.

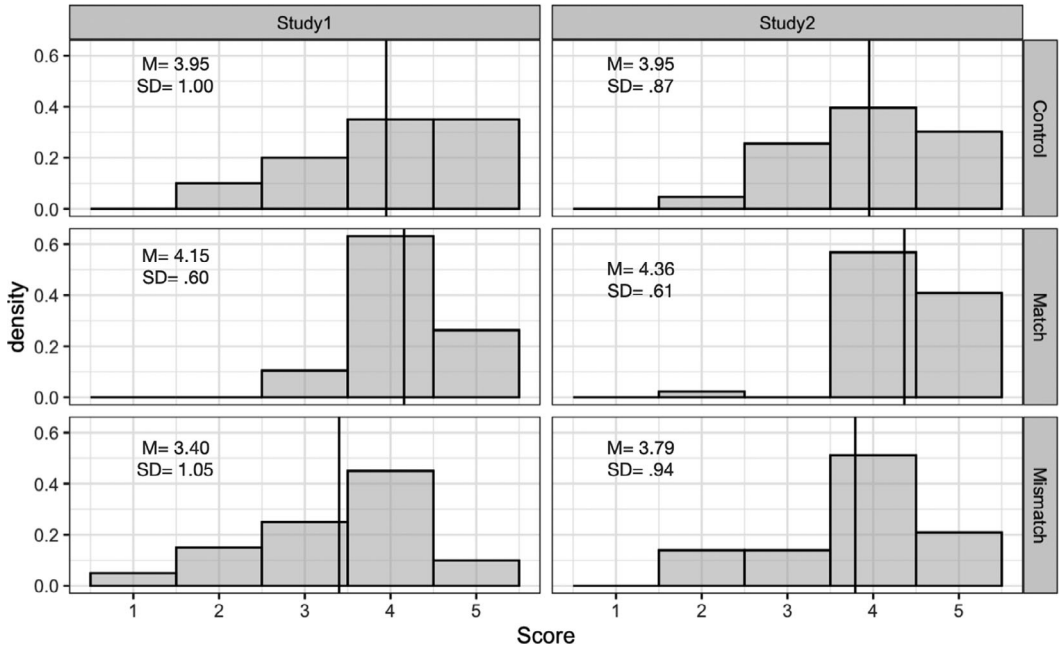

Fig. 5. Distributions of posttest scores by condition for Study 1 (left) and Study 2 (right).

3.2.2. Variance in posttest performance across conditions

One interesting finding from Study 1 was that the content-match condition not only improved learning but also reduced variation, apparently by raising the performance of students in the lower tail of the distribution. The reduction in variation was replicated in Study 2 (see Standard Deviations in Fig. 5). Levene's test of equality of variances was not significant for Study 2 (Levene's Statistics = 1.884, $p = .156$). However, given the distribution of the posttest scores in this study—that is, all students scored one of four scores (2, 3, 4, or 5)—it might be that Standard Deviation is not a sensitive enough indicator of variability in this case.

Looking at the middle panel of Fig. 5, it is clear that the percentage of students scoring 2 or 3 is attenuated in the content-match condition compared with both of the other conditions (2.3% in the content-match condition vs. 27.9% in the content-mismatch condition and 30.2% in the control). A comparison of 10,000 randomly bootstrapped pairs of samples ($n = 44$ and $n = 43$) showed that the differences between the percentage of students scoring below 4 in the content-match condition versus the other two conditions were highly unlikely to have occurred by chance ($p < .002$). It is also possible, however, that this reduction in variance is due to a ceiling effect for students in the content-match group.

3.2.3. Perceived purpose of the study

A chi-square test of independence revealed a significant relationship between the condition participants were in and what they perceived to be the purpose of the study ($\chi^2(6,$

Table 2
Percentage of participants in each group reporting each of four different perceived purposes of the study

	Perceived Purposes			
	Hand Movements/ Physical Actions	Multitasking	Other	Repeated Video Watching/Memory
Control	0.0	2.3	16.3	81.4
Content Match	34.1	22.7	22.7	20.5
Content Mismatch	21.0	48.8	11.6	18.6

130) = 66.82, $p < .001$). As seen in Table 2, this was mainly due to the difference between the control group and the other two groups. Because control participants were not assigned a hand movement task, they naturally thought the study was about the effect on learning of watching a video multiple times. Students in the content-match group more frequently guessed the purpose to be to investigate the role of instructed hand movements in learning, whereas students in the content-mismatch group were more likely to believe us when we told them the focus was on multitasking. This difference between the two experimental groups, however, was not statistically significant ($\chi^2(3, 87) = 7.12$, $p = .068$).

3.2.4. Did students who correctly guessed the purpose of the study learn more?

One of our research questions was whether students needed to be aware of the possible helpfulness of the instructed hand movements in order to benefit from them in learning. Although we specifically told participants that the hand movements were not related to the content of the video, some participants did not believe us by the end, correctly guessing that the hand movements were intended to affect learning. Did these participants, who correctly guessed the purpose of the study, benefit more from the instructed hand movements than the other (clueless) participants?

To answer this question, we looked at posttest performance of students who did versus did not guess that the hand movements were designed to impact their learning in some way. The results broken down this way are presented in Fig. 6, with those who guessed the purpose in the bottom row. The distributions of outcome scores for students in both the content-match and content-mismatch conditions look remarkably similar between those who guessed and those who did not. A two-way ANOVA revealed a main effect for condition ($F(1, 84) = 4.92$, $p = .004$), but not for perceived purpose ($F(1, 84) = 0.55$, $p = .460$).

3.2.5. Perceived usefulness of the instructed hand movements

Students differed across the two experimental conditions in their opinions of how useful the assigned hand movements were for learning. Eighty-two percent of the students in the content-match group found the hand movements helpful, compared with only 44% in the content-mismatch group. This difference was statistically significant ($\chi^2(1, 87) = 13.25$, $p < .001$).

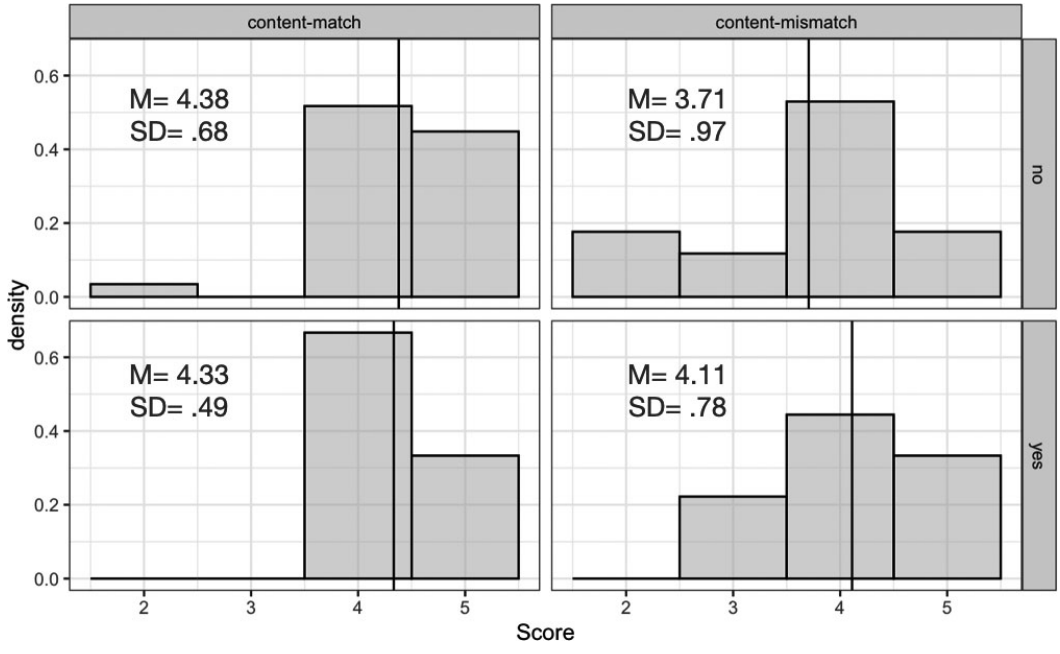

Fig. 6. Distributions of posttest scores, by condition, for participants who correctly guessed the purpose of the study (top) and participants who did not (bottom).

Similar to the comparison of those who guessed the purpose of the study and those who did not, we also compared performance on the outcome test between students who found the hand movements useful and those who did not. Results are presented in Fig. 7. Just as with the previous analysis, the distribution of outcome scores did not appear to differ based on whether they reported the hand movements as useful or not for learning, for either the content-match or content-mismatch conditions. A two-way ANOVA found a main effect for condition ($F(1, 84) = 10.38$, $p = .001$), but not of perceived usefulness ($F(1, 84) = 0.11$, $p = .736$).

3.2.6. Students' self-ratings of their understanding

After each viewing of the video, participants rated their understanding of the video on a 100-point scale. The results are shown in Fig. 8. The content-match group and the content-mismatch group showed an overall upward trend, as expected. That is, they reported understanding a little more after each successive viewing of the video. After their second viewing, all participants in the content-match group and 91% of participants in the content-mismatch group rated their level of understanding as either the same or higher than they had rated it after their first viewing. Furthermore, 93% of students in those groups continued the upward trend after their third viewing.

The control group, on the other hand, showed a different pattern. Although they tracked with the other two groups after the second viewing of the video, they reported a

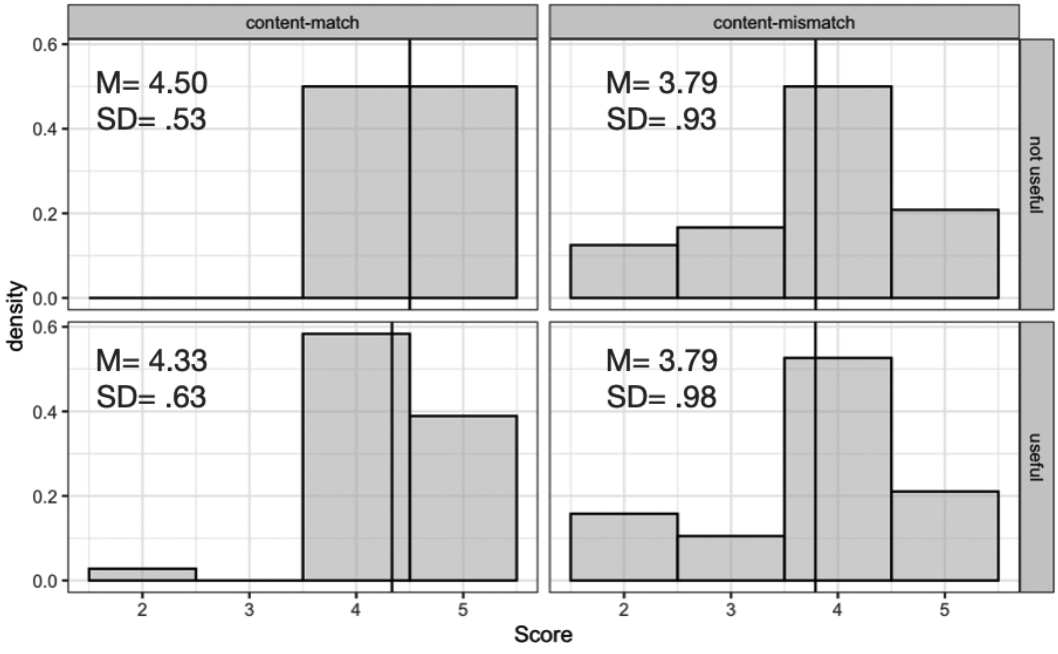

Fig. 7. Distributions of posttest scores, by condition, for participants who did not perceive them as helpful (top) and participants who perceived the instructed hand movements as helpful.

steep downturn in their understanding after the third viewing, with 33% of the participants rating their understanding as lower after watching the video for the third time than they did after the second time through. This pattern was confirmed by a 3 within-subjects x 3 between-subjects ANOVA, which found a significant time by group interaction, $F(4, 254) = 13.495$, $p < .001$. Our interpretation of this result is that control participants felt frustrated at having to watch the same video three times and acted out this frustration in their ratings. Perhaps having had the hand-placement task to keep them busy prevented the other two groups from feeling this frustration.

If we exclude time point three from the analysis (by running a 2 within-subjects × 3 between-subjects ANOVA), we see a pattern in ratings of understanding, with main effects for both time point ($F(1, 127) = 76.71$, $p < .001$) and condition ($F(2, 127) = 4.827$, $p = .010$), and no interaction between the two ($F(2, 127) = 1.128$, $p = .327$). With time point three excluded from the analysis, all three groups rated their understanding of the video as higher after viewing it the second time than after viewing it the first time. The content-match group ($t = 2.20$, $p = .030$) and control group ($t = 2.86$, $p = .005$) both rated their understanding as higher than did the content-mismatch group, which is consistent with their performance on the posttest. However, the content-match group's ratings were not significantly different from those of the control group ($t = 0.60$, $p = .552$), even though they performed significantly higher than the control group on the posttest. The correlation across the whole sample between ratings of understanding, averaged across time points 1 and 2, and learning was $r = .37$, $p < .001$.

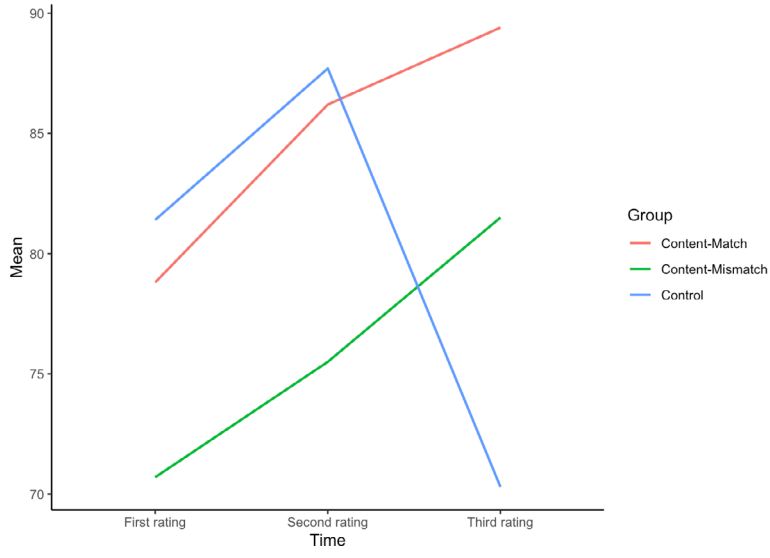

Fig. 8. Average self-ratings of how well participants in each group "understood the video" after each of the three viewings.

4. General discussion and conclusion

Across two studies, the ways in which participants moved their hands while they learned a complex concept led to significant differences in their performance on a posttest that required them to apply that concept. Specifically, participants who placed their hands in ways that supported a correct interpretation of the content they were learning scored significantly higher than those who placed their hands in ways that supported a common misconception of the concept (or which were otherwise misaligned with a correct interpretation). Thus, hand orientation matters for learning, even when students are not given a communicative purpose for their instructed hand movements. In large part, prior work has focused on gestures that represent or hint at solution strategies in a problem-solving context. However, the two studies reported here demonstrate that instructed hand movements that simply reinforce visual representations of key ideas can help students better understand difficult concepts in statistics, apart from any particular problem-solving task. The current studies extend the literature in an additional way. Prior work has shown that learning can be impacted by spontaneous and/or co-speech gestures (i.e., the gestures that people generate to accompany what they say; Andric, & Small, 2012). Our studies show that hand movements that occur independent of both the performer's speech and the performer's knowledge of their purpose can also facilitate learning.

In both Study 1 and Study 2, participants performed instructed hand movements without speaking. When, in Study 2, participants were queried about what they perceived to be the purpose of the study, the majority misidentified it as being about the effect of repeated video watching or multitasking. Thus, purposeful generation of hand movements

to accompany speech is not a necessary condition for hand movements to impact learning. Learners do not even need to understand the meaning of their instructed hand movements in order for their learning to be enhanced.

If the effectiveness of gesture lies not in its suggestion of a solution strategy, nor in its co-occurrence with speech, nor in its spontaneous generation, nor even with the performer's understanding of its purpose, what mechanism might explain its effectiveness? Findings from the embodied cognition literature support the idea that people's bodily movements, even purposeless instructed hand movements without accompanying speech, can impact cognition and learning (Smith, 2005). However, whereas the bodily movements from previous studies have been related to a specific problem-solving strategy or emotional signal (Kraft, & Pressman, 2012), the instructed hand movements in our studies are simply representations of central features of an idea or concept.

Our findings showed that the content-match group participants' physical representation of an important idea may have served to support their integration of this idea into their efforts to understand a concept. Similarly, the fact that the content-mismatch group performed worse on the posttest might be due to the reinforcement of a misconception activated by their instructed hand movements. For the content-mismatch group, it might even have been that incongruence between the audio to which the participants listened and the instructed hand movements they were asked to produce did, in fact, turn their activity into multitasking. The extra cognitive load may have dampened their learning.

This finding has implications for the design of mathematics learning technologies. Goldstone, Marghetis, Weitnauer, Ottmar, and Landy (2017) point out the importance of taking perceptual and spatial information into account when developing new technological tools designed to teach students about mathematical expressions. Use of perceptual and spatial information can help students develop valid intuitions and overcome misconceptions. Our finding suggests that an intervention as simple as placing your hands at the right place can help students learn to interpret histograms. The finding might extend beyond statistics education to education in general. It supports the use of instructed hand movements as a resource for teaching abstract ideas in complex domains. Activating the learners' physical representation of the concept appears to help with their construction of conceptual understanding.

Another contribution of the present pair of studies is derived from analyses of participants' self-ratings of their understanding. Across the three video viewings, for the two experimental groups, the trajectory of perceived understanding played out as one might expect: Self-ratings of understanding increased with each video viewing. If it is true that matching instructed hand movements to content improves learning, and if participants can accurately assess their own levels of learning, one would expect significant differences between the perceptions across groups and across time, favoring participants in the content-match group. This, too, was borne out in the data.

What is perhaps more interesting is how control group participants' self-ratings of understanding compared to those of the experimental group participants. In a pattern markedly different from the experimental groups, the control group participants' ratings dropped precipitously from their second to third video viewing. It seems unreasonable

to assume that they felt they understood the material in the video less well because of the additional video viewing. What we think is more likely is that they were acting out, frustrated by having to watch the video yet again. The interesting question is: why did not participants in the experimental groups—the groups given something to do while they watched the instructional video—respond in this way? We think it is possible that the secondary task kept them engaged in the primary task. This serendipitous finding has implications for creating learning activities for students that keep frustration at bay and increase their willingness to continue to engage with instruction. The studies reported here have other implications for learning environments outside of the laboratory. Although Study 1 called for an experimenter to be present and for participants to engage in a rather artificial learning setting, the procedures used in Study 2 demonstrated that students can profit from following hand movement instructions even when they sit alone, watching an instructional video on a laptop. Future studies of online learning from video should test this application further, adding hand movement instructions to teaching materials in other domains in which concepts can be represented visually.

4.1. Possible concerns, alternative explanations, and limitations

It is possible that, from the learners' perspective, the animation of the red bars in the videos could have served the same function as a speaker's gestures would have. That is, watching the red bars move was akin to watching a teacher's hands move, and it was watching the bar movement—not producing the hand movements—that led to group differences in learning. Although we did not include a condition in which students watched a video that included the red bars without moving their hands and thus we cannot distinguish in the current study the effect of seeing the bars from the effect of using them as instructions for where to place one's hands, we do not believe this explanation is likely. Several studies have compared the effect of gesture with that of animated highlighting; all have found that gesture produces better learning than does animated highlighting (Bem et al., 2012; Louden et al., 2015). Furthermore, adding just animated highlighting to videos of narrated slides has not been shown to improve learning compared with just the voice and slides without the animation (Bem et al., 2012; Louden et al., 2015). Based on these findings, we do not believe that the red bars served as a stand-in for a teacher's gesture.

Another issue stems from whether students believed the cover story about the study being related to multitasking. From the post survey, we know that a majority of students assumed that the purpose of the study was either multitasking or repeated video watching, which suggests that they did not perceive a relationship between what they were doing with the red bars and the concepts being taught in the video. But it is worth noting that more students in the content-match condition (34.1%) perceived the purpose of the study as being related to gesture or hand movement than did students in the content-mismatch condition (21%). The concern is that even a 13% difference in perceived purpose of the study could have driven the difference in the outcome measures. If correctly guessing the

purpose of the study had a significant impact on students' learning, then we should see a difference in posttest scores between students who had correctly guessed the purpose of the study and students who did not. However, neither the set of posttest histograms nor our statistical analysis revealed such differences, suggesting that even for the students who detected the purpose of the study—and perhaps even the relationship between the red bars, the corresponding hand movements, and the video concept—their correctly guessing the purpose of the study did not lead to better posttest performance.

In addition, because we designed the timing, position, and direction of the gestural instructions (i.e., the red bars) based on the instructor's spontaneous gestures and common student misconceptions, one limitation of our design is that we did not control for the orientation of the red bars. The static hand position was always vertical for the content-match condition and horizontal for the content-mismatch condition. Thus, another possible explanation for the observed differences in posttest performance across groups is that they were caused by the particular position in which the hands were placed. Students in the content-match condition always positioned their hands vertically, whereas students in the content-mismatch condition always positioned their hands horizontally. Positioning hands horizontally could have interfered with students' ability to process the vertical lines in the histogram (e.g., the lines for the means or the bins of the histograms), which would place the content-mismatch condition at a disadvantage, compared not only to the content-match condition, but also to the control condition.

The two studies reported here have several other limitations that need to be considered. First, data were collected from participants during a single laboratory session, thus preventing us from making claims about long-term retention of concepts learned, and from knowing how the intervention might play out in a live classroom environment. Also, data were collected from a narrow sample of students: undergraduates enrolled in a small number of psychology classes. Especially because the intervention appeared particularly effective for students with lower levels of understanding, extending the subject pool to a more diverse population is worthy of investigation.

4.2. Future directions

Many questions, of course, are left unanswered. We find the decrease in variation of the posttest scores of the content-match group interesting, because it indicates that instructed hand movements that match the content of the instruction can help some people more than others. In future studies, we are interested to see if instructed hand movements are more powerful in helping students with less incoming knowledge or with more misconceptions about the concept being taught. Because the decrease in variation of the posttest scores might be attributable to a ceiling effect, we would like to design more difficult posttest questions (e.g., questions that require more transfer).

The finding of increased gestures in posttest questions for the two experimental conditions in Study 1 is also worth further exploration. In particular, we are interested in investigating the degree to which students' gestures are congruent with what they were instructed to perform while watching the video and how that interacts with their learning.

If Galati and Samuel's (2011) theory of multimodal gist representation applies, then we expect the gestures to be congruent with what participants were instructed to perform. Moreover, we expect that among participants who were instructed to perform mismatched hand movements, those who generate content-match gestures will demonstrate better learning outcomes than those who continue to rely on their initial mismatch representations.

In the two studies presented here, participants were asked to watch the video three times, which can be hard to enforce in real-life learning situations. Thus, we would like to know what the effect of the instructed hand movements would be if participants watch the video only once or twice to see if the effect is only manifested after some threshold number of viewings. Related to this, we want to further explore the effect of instructed hand movements on students' self-ratings of their understanding. For example, will we be able to replicate the effect hand movement tasks can play in preventing students from becoming frustrated by repeated video-watching? And finally, we would like to extend the work beyond the concept of statistical model to explore the effect of this type of gestural intervention in different concepts and disciplines.

4.3. Conclusion

Our study demonstrates that instructed hand movements that are related to central features of the underlying structure of the concept, but unrelated to a specific solution strategy, can promote learning of abstract ideas in statistics. The study sheds light on the possibility of using gestures not only as an additional pathway to deliver solution strategies, but also as an additional resource for students to explore and understand abstract ideas and develop conceptual understanding. In this framework, the use of instructed hand movements in education makes students not passive recipients of information, but active participants who can take advantage of instructed hand movements to promote their own learning.

Acknowledgments

The authors gratefully acknowledge the support of the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation (DRL-1229004) and the California Governor's Office of Planning and Research (contract OPR18115).

Open Research badges

This article has earned Open Data and Open Materials badges. Data and materials are available at http://re3data.org/.

References

Alibali, M. W., & Kita, S. (2010). Gesture highlights perceptually present information for speakers. *Gesture*, *10*(1), 3–28. https://doi.org/10.1075/gest.10.1.02ali

Alibali, M. W., & Nathan, M. J. (2011). Embodiment in mathematics teaching and learning: Evidence from learners' and teachers' gestures. *Journal of the Learning Sciences*, *21*(2), 247–286. https://doi.org/10.1080/10508406.2011.611446

Alibali, M. W., Spencer, R. C., Knox, L., & Kita, S. (2011). Spontaneous gestures influence strategy choices in problem solving. *Psychological Science*, *22*(9), 1138–1144. https://doi.org/10.1177/0956797611417722

Andric, M. W., & Small, S. L. (2012). Gesture's neural language. *Frontiers in Psychology*, *3*(99), 1–12. https://doi.org/10.3389/fpsyg.2012.00099

Bem, J., Jacobs, S. A., Goldin-Meadow, S., Levine, S., Alibali, M. A., & Nathan, M. (2012). Gesture's benefit for instruction: Attention coordination or embodied cognition? Paper presented at the Jean Piaget Society (JPS) conference. Toronto, CN.

Broaders, S. C., Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2007). Making children gesture brings out implicit knowledge and leads to learning. *Journal of Experimental Psychology: General*, *136*(4), 539–550. https://doi.org/10.1037/0096-3445.136.4.539

Brooks, N., & Goldin-Meadow, S. (2016). Moving to learn: How guiding the hands can set the stage for learning. *Cognitive Science*, *40*, 1831–1849. https://doi.org/10.1111/cogs.12292

Cherdieu, M., Palombi, O., Gerber, S., Troccaz, J., & Rochet-Capellan, A. (2017). Make gestures to learn: Reproducing gestures improves the learning of anatomical knowledge more than just seeing gestures. *Frontiers in Psychology*, *8*, 1–15. https://doi.org/10.3389/fpsyg.2017.01689

Cook, S., Mitchell, Z., & Goldin-Meadow, S. (2008). Gesturing makes learning last. *Cognition*, *106*, 1047–1058. https://doi.org/10.1016/j.cognition.2007.04.010

delMas, R., Garfield, J., & Ooms, A. (2005). Using assessment items to study students' difficulty reading and interpreting graphical representations of distributions. In K. Makar (Ed.), *Proceedings of the Fourth International Research Forum on Statistical Reasoning, Thinking and Literacy* (pp. 1–18). Auckland, New Zealand: University of Auckland.

Dove, G. (2018). Language as a disruptive technology: Abstract concepts, embodiment and the flexible mind. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*, 20170135. https://doi.org/10.1098/rstb.2017.0135

Fries, L., Son, J. Y., Givvin, K. B., & Stigler, J. W. (2020). Practicing connections: A framework to guide instructional design for developing understanding in complex domains. *Educational Psychology Review*, 1–24. https://doi.org/10.1007/s10648-020-09561-x

Galati, A., & Samuel, A. G. (2011). The role of speech-gesture congruency and delay in remembering action events. *Language and Cognitive Processes*, *26*(3), 406–436. https://doi.org/10.1080/01690965.2010.494846

Goldin-Meadow, S., Cook, S., & Mitchell, Z. (2009). Gesturing gives children new ideas about math. *Psychological Science*, *20*, 267–272. https://doi.org/10.1111/j.1467-9280.2009.02297.x

Goldstone, R. L., Marghetis, T., Weitnauer, E., Ottmar, E. R., & Landy, D. (2017). Adapting perception, action, and technology for mathematical reasoning. *Current Directions in Psychological Science*, *26*(5), 434–441. https://doi.org/10.1177/0963721417704888

Kita, S., Alibali, M. W., & Chu, M. (2017). How do gestures influence thinking and speaking? The gesture-for-conceptualization hypothesis. *Psychological Review*, *124*(3), 245–266. https://doi.org/10.1037/rev0000059

Kontra, C., Goldin-Meadow, S., Beilock, S., & Sian, L. (2012). Embodied learning across the life span. *Topics in Cognitive Science*, *2012*(4), 731–739. https://doi.org/10.1111/j.1756-8765.2012.01221.x

Kraft, T. L., & Pressman, S. D. (2012). Grin and bear it: The influence of manipulated facial expression on the stress response. *Psychological Science*, *23*(11), 1372–1378. https://doi.org/10.1177/0956797612445312

Krönke, K., Mueller, K., Friederici, A., & Obrig, H. (2012). Learning by doing? The effect of gestures on implicit retrieval of newly acquired words. *Cortex*, *49*, https://doi.org/10.1016/j.cortex.2012.11.016

Louden, D., Saucedo, M., Latif, D., Gorczynski, L., Burns, C., Rueckert, L., & Alibali, M. (2015). Factors that determine the embodiment of mathematical concepts: The effect of instruction with gesture on learning and muscle movement. Poster presented at the Jean Piaget Society (JPS) Conference in Toronto, CA.

Macedonia, M., Mueller, K., & Friederici, A. D. (2010). The impact of iconic gestures on foreign language word learning and its neural substrate. *Human Brain Mapping*, *32*, 982–998. https://doi.org/10.1002/hbm.21084.

Nathan, M. J., Walkington, C., Boncoddo, R., Pier, E. L., Williams, C. C., & Alibali, M. W. (2014). Actions speak louder with words: The roles of action and pedagogical language for grounding mathematical proof. *Learning and Instruction*, *33*, 182–193.

Novack, M., & Goldin-Meadow, S. (2015). Learning from gesture: How our hands change our minds. *Educational Psychology Review*, *27*, 405–412. https://doi.org/10.1007/s10648-015-9325-3

OECD. (2012). *Equity and quality in education: Supporting disadvantaged students and schools*. Paris: OECD. https://doi.org/10.1787/9789264130852-en

Ping, R. M., & Goldin-Meadow, S. (2008). Hands in the air: Using ungrounded iconic gestures to teach children conservation of quantity. *Developmental Psychology*, *44*(5), 1277–1287. https://doi.org/10.1037/0012-1649.44.5.1277

Richland, R. E., Stigler, J. W., & Holyoak, K. J. (2012). Teaching the conceptual structure of mathematics. *Educational Psychologist*, *47*(3), 189–203. https://doi.org/10.1080/00461520.2012.667065

Rueckert, L., Church, R. B., Avila, A., & Trejo, T. (2017). Gesture enhances learning of a complex statistical concept. *Cognitive Research: Principles and Implications*, *2*(1), 2. https://doi.org/10.1186/s41235-016-0036-1

Smith, L. B. (2005). Action alters shape categories. *Cognitive Science*, *29*, 665–679. https://doi.org/10.1207/s15516709cog0000_13

Son, J., Blake, A., Fries, L., & Stigler, J. (2020). Modeling first: Applying learning science to the reaching of introductory statistics. *Journal of Statistics Education*, 1–34. https://doi.org/10.1080/10691898.2020.1844106

Son, J., Ramos, P., DeWolf, M., Loftus, W., & Stigler, J. (2018). Exploring the practicing-connections hypothesis: Using gesture to support coordination of ideas in understanding a complex statistical concept. *Cognitive Research: Principles and Implications*, *3*, 1–13. https://doi.org/10.1186/s41235-017-0085-0

Steier, R. (2014). Posing the question: Visitor posing as embodied interpretation in an art museum. *Mind, Culture and Activity*, *21*(2), 148–170.

Stigler, J., Son, J., Givvin, K. B., Blake, A., Fries, L., Shaw, S., & Tucker, M. (2020). The Better Book approach for education research and development. *Teachers College Record*, *123*(2), 1–32.

Thomas, L. E., & Lleras, A. (2009). Covert shifts of attention function as an implicit aid to insight. *Cognition*, *111*, 168–174. https://doi.org/10.1016/j.cognition.2009.01.005

Wakefield, M., Hall, C., James, K. H., & Goldin-Meadow, S. (2018). Gesture for generalization: Gesture facilitates flexible learning of words for actions on objects. *Developmental Science*, *21*(5), e12656. https://doi.org/10.1111/desc.12656

Walkington, C., Chelule, G., Woods, D., & Nathan, M. (2019). Collaborative gesture as a case of extended mathematical cognition. *Journal of Mathematical Behavior*, *55*, 1–20. https://doi.org/10.1016/j.jmathb.2018.12.002

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*, 625–636. https://doi.org/10.3758/BF03196322

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article:

Data S1. Video material for Study 1 and Study 2.

Appendix A: Transcript for the instructional video used in Study 1 and Study 2

Transcript of instructional video

So statistics is the study of variation. But then we need to really unpack, what is variation? And in statistics we often are not just talking about any old kind of variation; we're really talking about variation in data. And so let's take this outcome variable that we've measured out in the real world, and we can see there's variation here. Let's say we've measured people on some outcome variable, and we can see that not all people are the same; they vary. And so, even though a lot of people have around 61 or 62 as their score on this outcome variable, there's a bit of variation here. There are some people who are scoring all the way up to 70, ans some people scoring all the way down to 50, and so, there's variation.

Now, if we had taken another measurement, we might see a very different pattern of variation. For example, here we see a lot more variation in this distribution. They seem to have roughly similar centers; for instance, it seems to be around 64, low 60s perhaps, but there's a lot more variation in this particular distribution. We see people scoring all the way up to 80 and all the way down to 40. Now, what do statisticians do with this variation we see in distributions? Often, we create a model, but that begs the question: what is a model? And so in this case we're just going to draw from our regular old understanding of models. For instance, if we have a White House, the real White House, we might create a scale model of it to help us understand some things about it. For instance, this scale model is not the same as the White House; it's not even the same color, but from it we could represent certain parts of this White House. For instance, it has four columns and the real White House has four columns.

Now, what do we do with a statistical model? It's not just a small version of the distribution. In fact, it's just going to be simplified somehow and often it's simplified by having fewer things about that distribution. For example, a simple model is the mean as the model of this whole distribution. It's simpler because it's just one number. This distribution has a whole bunch of numbers, but the mean is just one, simple number.

In this case, the mean seems to be around 64, and you can see that 64 is not a perfect representation of this distribution. This distribution is a lot more varied, it has a lot more details to it, and it totally misses this idea that some people scored 80 on this outcome

variable. But it's a good representation because it roughly captures something about this distribution that's important, namely the center of it.

Now, one of the things we could do with a simple statistical model is that we could use it to predict the next data point that we gather from the world. Let's say I told you, "We're going to go out there and measure one more person on this outcome variable." And given that you have no other information about this person, you might think, "Well, maybe they're going to be about 64 on this outcome variable." It won't be perfect, it might be off by some amount, but this might be one of the better guesses that you could make, given this information.

Now, there are going to be situations where that simple model is a better model and some cases where it's a worse model. Take, for instance, this top distribution; we could see that because there's less variation around the model, this model is quite a bit better. But in this bottom distribution, there's more variation around the model and because of that, it's not going to be quite as good of a prediction, not going to be quite as good of a representative of this whole distribution. Now, one of the things we're doing here is, we're making a subtle shift. Instead of just looking at how much variation there is; we're looking at how much error there is from the model. If you predict that the next person is going to be 64, in the case of this top distribution here, you might not be off by that much. Your error will be smaller. But if you guess, "This next person is going to be 64," and the bottom distribution is a good representation of the world, then you will be off by more on average.

So let's break down this idea of error from the model a little bit more concretely with this one data point that we've colored white. So here we see this distribution, there's variation, and here's this simple model, the mean, 64, but you can see this data point is not exactly our model. Our model is wrong. Well, how wrong is it? We can represent how wrong it is by this distance between the data point and the model. So you could think of all data points as being made up of two components. You could think of each data point as being, part of it is our model prediction, and part of it is how off that model is from the actual data point.

Now that works well if the world is a simple place that could be described with just one number, but let's say that we know a little bit more about the world and its complexity. Can we make our simple model a little more complex? Let's say we know that this distribution of people actually is made up of two different groups of people: the red group and the yellow group. Now the red group, you can see, they vary, too. Just because you're in the red group doesn't mean that you're all the same and just because you're in the yellow group doesn't mean that you're all the same either. How can we make use of this little bit more information that we have?

Well, one of the things you can see is that the red group, although it's not exactly perfect, they tend to have lower scores on this outcome variable than the yellow group. And that's not true for every single person, but it's kind of true as a whole. So if we put up the means for the red and yellow groups, you could see that the red group's mean is smaller or lower than the yellow group's mean. Now, one of the ways that we might use that information in our two group model is by saying, "If I knew what group you came

from, then I can make a different prediction about you." So, if I knew that the next person we are going to measure comes from the red group, we might say, "I predict that their score is going to be a little bit smaller. Maybe their score is going to be around 61." But if I knew this person was going to come from the yellow group, I might make a different prediction. I might say, "I think this person's score is going to be closer to 70."

And so, this model is very similar to the simple model of the mean we saw before; now we're just using two means to represent the two different groups. And because of that, we are going to call this the "two group model," because it's slightly more complex than the one group model.