Journal of Numerical Cognition
jnc.psychopen.eu | 2363-8761

PsychOpen publishing psychology

Research Reports

Visual Comparison of Two Data Sets: Do People Use the Means and the Variability?

Robin S. S. Kramer*[ab], Caitlin G. R. Telfer[b], Alice Towler[b]

[a] Department of Psychology, Trent University, Peterborough, Canada. **[b]** Department of Psychology, University of York, York, United Kingdom.

Abstract

In our everyday lives, we are required to make decisions based upon our statistical intuitions. Often, these involve the comparison of two groups, such as luxury versus family cars and their suitability. Research has shown that the mean difference affects judgements where two sets of data are compared, but the variability of the data has only a minor influence, if any at all. However, prior research has tended to present raw data as simple lists of values. Here, we investigated whether displaying data visually, in the form of parallel dot plots, would lead viewers to incorporate variability information. In Experiment 1, we asked a large sample of people to compare two fictional groups (children who drank 'Brain Juice' versus water) in a one-shot design, where only a single comparison was made. Our results confirmed that only the mean difference between the groups predicted subsequent judgements of how much they differed, in line with previous work using lists of numbers. In Experiment 2, we asked each participant to make multiple comparisons, with both the mean difference and the pooled standard deviation varying across data sets they were shown. Here, we found that both sources of information were correctly incorporated when making responses. Taken together, we suggest that increasing the salience of variability information, through manipulating this factor across items seen, encourages viewers to consider this in their judgements. Such findings may have useful applications for best practices when teaching difficult concepts like sampling variation.

Keywords: informal inferential reasoning, comparing groups, mean difference, pooled standard deviation, variability

Journal of Numerical Cognition, 2017, Vol. 3(1), 97–111, doi:10.5964/jnc.v3i1.100

Received: 2016-11-03. Accepted: 2017-03-14. Published (VoR): 2017-07-21.

*Corresponding author at: Department of Psychology, Trent University, Peterborough, Ontario K9J 7B8, Canada. E-mail: remarknibor@gmail.com

When deciding whether two groups are different on some measure, one of the most important concepts to understand is the mean or "average". Indeed, many teachers have focussed on determining the best ways to convey this idea to students at an early age, both through calculation and visual impression (Gal, 1995; Watson & Moritz, 1998). Even in adulthood, we are often presented with mean values in newspapers or television adverts (e.g., comparing the miles per gallon of two car models or the battery life of two smartphones) and expected to decide whether there is a meaningful difference. Problematically, research suggests that we place more weight than we should on these types of average ratings (e.g., de Langhe, Fernbach, & Lichtenstein, 2016). Statistically, information about the means alone is insufficient for making such decisions. One also requires knowledge of the variances (or some other measure of the "spread" of the two groups) in order to determine the size and importance of any difference. Unfortunately, understanding the concept of sampling distributions is both difficult for people (delMas, Garfield, & Chance, 1999) and under-researched (Meletiou, 2000).

For several years, studies have investigated what has been termed 'informal inferential reasoning', where judgements are made based on prior knowledge but not formal statistical procedures. Here, we consider whether people are sensitive to both the average and the spread of data when asked to make such judgements. Recent research provides some evidence that both the mean differences and the set variances correctly influence decisions about which of two groups is larger when the data are presented as lists of raw values (Morris & Masnick, 2015). However, with a greater number of mean and standard deviation conditions, and longer lists of numbers, Saito (2015) found that participants correctly judged an increase in effect size for larger mean differences but perceived incorrectly that effect size also increased as standard deviations increased. Obrecht, Chapman, and Gelman (2007) showed that participants gave little consideration to either the sample size or the standard deviation when presented with two lists of raw data. Instead their judgements were primarily driven by the mean differences. In addition, researchers have shown that between-group variability (the difference between the means) influences decisions when displaying raw data but participants did not respond to changes in within-group variability (each group's spread) when comparing groups (Masnick & Morris, 2008). Even experience with introductory-level statistics did not guarantee that within-group variability would be given sufficient consideration (Trumpower, 2015; Trumpower & Fellus, 2008; Trumpower, Hachey, & Mewaldt, 2009). In fact, after extensive training, secondary-school mathematics teachers continued to show difficulties with the concept of sampling distributions in the context of comparing two groups (Makar & Confrey, 2004).

So far, the raw data for the two groups have been presented as lists of numbers from which participants were expected to make summary judgements. In addition, several studies have investigated the potential effects of displaying summary statistics such as the mean, sample size, and standard deviation (e.g., using visual analogue scales; Obrecht, Chapman, & Suárez, 2010). For instance, work with box plots as a presentation method has shown that using these in combination with in-depth instruction may facilitate students' understanding of sampling variability, along with how to compare two sets of data (Bakker, Biehler, & Konold, 2005; Pfannkuch, 2006; Pfannkuch, Arnold, & Wild, 2015). Reading and Reid (2005, 2006) model the learning progression in students as a shift from thinking about only the means to a strong consideration of variation.

Also of relevance to the current research is how people understand visual representations of data more generally. As mentioned, little is known about how people compare two sets of raw data presented visually. However, evidence suggests that even our interpretations of simple line graphs, depicting three variables, are often incomplete and incorrect, and such graphs require complex processes in order to comprehend (Carpenter & Shah, 1998; Shah & Carpenter, 1995). Further, presenting the data as bar versus line graphs, for example, appears to influence viewers' interpretations – the former encourages descriptions of *x*-*y* interactions while the latter results in descriptions of main effects and *z*-*y* interactions (Shah & Freedman, 2011). Indeed, bar graphs also suffer from a 'within the bar' bias, where data points falling within the area of the bar itself are seen as more likely than those appearing outside the bar (Newman & Scholl, 2012). As one would predict, the particular layout and details of the graph play an important role in how the data are interpreted, as does participants' graph-related prior knowledge (Okan, Galesic, & Garcia-Retamero, 2016; Okan, Garcia-Retamero, Galesic, & Cokely, 2012; Shah & Freedman, 2011).

To our knowledge, only one study has displayed raw data (that is, each value separately) visually. Fouriezos, Rubenfeld, and Capstick (2008) used a cluster of vertical bars to represent each group (each bar was a data point) and asked participants to judge which of the two clusters were taller. The results showed that the mean

difference had a large effect on participants' responses, while the sample sizes and standard deviations showed statistically significant, but far smaller, effects on decisions. However, there is a large variety of options available when depicting datasets (e.g., Cleveland & McGill, 1985). Here, we investigate whether displaying the two groups' data as parallel dot plots may help participants to incorporate information about spread into their decisions. Previous research has shown that observers are capable of accurately comparing the means of two groups when displayed on a single scatter plot (Gleicher, Correll, Nothelfer, & Franconeri, 2013). However, participants were not required to make decisions regarding the size of the differences between groups, and within-group variability was not manipulated or investigated.

In the current work, we investigate the possibility that viewers previously failed to incorporate information regarding variability because data were presented as lists of numbers, summary values, or in graphically inaccessible ways. We hypothesise that visually presenting the data using dot plots may provide this type of information in a readily accessible format. In two experiments, we investigate whether this presentation method will result in both the means and standard deviations influencing participants' responses.

Experiment 1

In this first experiment, we investigated whether the means and standard deviations of the simulated data would affect responses when tested *across participants*. That is, each participant was presented with only one version of the graph, and was asked to make a single rating. In this way, our design is similar to real-world decision-making, whereby information on two products are compared and one is chosen.

Method

Participants

We recruited participants from three sources in order to incorporate a wide range of ages and education levels. The first group (n = 77) comprised members of the public who attended an interactive psychology event, open to everyone and held in the city centre. There was no cost of admission for the event and participants were not compensated for taking part. The second group (n = 17) comprised students at a university in the northeast of England, who received either course credits or money for their participation. The third group (n = 71) comprised secondary school students and teachers who attended an open day at the university. These participants received no compensation for taking part.

The three groups represented convenience samples that, when combined, would provide a spread of ages and education levels. Large amounts of variation along these dimensions were also present within groups, in particular the first group comprising members of the public. As such, the uneven sizes of these nominal groups, of itself, was not important. In total, 165 people (112 women; age M = 25.30 years, SD = 12.97 years) participated. Complete demographic information can be found in the Supplementary Material.

All participants provided verbal consent before taking part, and were given both a verbal and written debriefing after completion. The experiment's design and procedure were approved by the university psychology department's ethics committee (identification number 484) and conform to the Declaration of Helsinki.

Materials

We gave participants a pen-and-paper questionnaire describing a study in which a new product, 'Brain Juice', was being tested for its memory-boosting ability. Participants were informed that in this fictional study, one group of 20 children drank water before their memory test while a second group of 20 children drank Brain Juice. The two groups were reported as identical in all other ways. The children's memory test scores were then presented in a graph on the questionnaire for the participants to examine (see Figure 1).

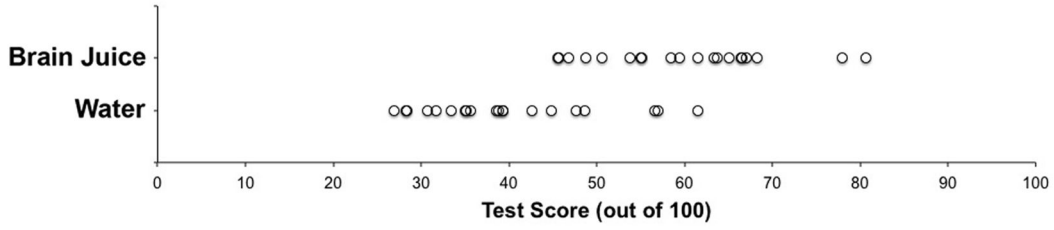

Figure 1. Example version of a graphical representation of the data shown to participants in Experiment 1 (Condition 4 in Table 1).

We created seven versions of this pen-and-paper questionnaire (see Table 1 and the Supplementary Material), varying only the data parameters according to those used by Saito (2015). The use of seven conditions represented a compromise between achieving variation in mean differences and standard deviations while limiting the number of groups, given the between-participants nature of the design.

Table 1

Summary of the Seven Versions of the Graph and Participants' Responses

		Water		Brain Juice			
Condition	***n***	***M***	***SD***	***M***	***SD***	**Cohen's *d***	**Participants' Ratings**
1	24	40	10	45	10	0.5	3.21 (2.02)
2	24	40	20	50	20	0.5	3.67 (1.76)
3	24	40	10	50	10	1.0	3.83 (1.55)
4	24	40	10	60	10	2.0	5.29 (1.49)
5	23	40	5	50	5	2.0	3.96 (1.85)
6	23	40	10	80	10	4.0	6.17 (1.64)
7	23	40	2.5	50	2.5	4.0	4.35 (1.72)

Note. n is the number of participants that completed each version of the questionnaire. Participants' ratings, given in response to the rating question, were calculated during the analysis and are presented as *M* (*SD*).

The children's test scores were produced using customised MATLAB software. For each set of values, 20 normally distributed random numbers were generated and then standardised, resulting in a mean of zero and a standard deviation of one. These were then multiplied by the standard deviation specified for that condition (see Table 1), and then the corresponding mean value was added. Therefore, all condition means and standard deviations were exact even though the values were originally generated using random numbers.

Procedure

Each participant received one version of the questionnaire only, determined by the order in which they took part – the first person was given Condition 1, the second Condition 2, and so on, with the eighth starting at 1 again. Table 1 summarises how many people completed each version of the questionnaire.

After reading the description of the fictional study and examining the graph that followed, the questionnaire then asked, "If you were asked to give a rating, how large do you think the Brain Juice improvement was?" Participants circled their answer on a labelled rating scale from 0 ('none') to 9 ('very large') on a 10-point scale. Next, participants read, "If a reporter asked you if the Brain Juice group did better than the children who drank water, what would you say?" Answers were given by circling either 'yes' or 'no'. This second question was designed to investigate real-world outcomes, where participants are required to make a decision regarding, for example, the purchase of one particular car over another.

Finally, demographic information was collected: age, sex, and how much mathematics/statistics education participants had previously received. For this question, several options were provided (e.g., secondary school, undergraduate degree) for participants to select, or they could choose 'other' and provide an open response.

Throughout testing, participants were instructed not to confer, and were not shown other versions of the questionnaire.

Results

Participant data and visual stimuli can be found in the Supplementary Material.

The data were analysed using multiple regression in order to determine which factors predicted participants' judgements of the Brain Juice improvement. First, the amount of statistics education that participants had received was converted to an ordinal variable, ranging from 1 (primary school) to 5 (PhD). Next, several regression models were explored with participants' ratings as the dependent variable, and condition variables (mean difference, pooled standard deviation, Cohen's *d*) as predictors. The mean difference was simply *M*(Brain Juice) minus *M*(water), whereas the pooled standard deviation always equalled the standard deviation of either group since both had equal standard deviations. Cohen's *d*, a statistical measure of the size of the difference between the two groups, was given by the mean difference divided by the pooled standard deviation.

Modelling Average Ratings

First, we averaged ratings across participants for each condition (see Table 1, final column). Using regression models, we then explored whether condition variables predicted mean responses for our participant sample as a whole. Note that averaging across conditions, while smoothing out potential noise due to individual differences in responses, resulted in only seven observations. As such, any conclusions from these analyses are limited in their scope.

In Models 1-3 (see Table 2), we find that the mean difference between the two groups (water versus Brain Juice) is the best individual predictor of participants' ratings. In Model 5, we include all three variables (the collinearity statistics remain within acceptable ranges) and find a decrease in performance in comparison with Model 1. While Model 4 suggests a slight improvement (higher adjusted R^2) to Model 1 if the pooled standard deviation is included, this increase in explanatory power is not statistically significant, R^2_{change} = .030,

$F_{change}(1,4) = 1.53$, $p = .284$. Therefore, the best model includes only the mean difference as a predictor of averaged ratings. Equally, a stepwise linear regression, initially entering all three variables, also results in Model 1 as the best solution.

Table 2

Results of the Multiple Regression Analyses on Average Ratings in Experiment 1

Model	Variable	Beta	*t*	*p*	Adjusted R^2	*F*
1	Mean difference	0.944	6.42	.001	.870	41.27
2	Pooled standard deviation	-0.143	-0.32	.760	.176	0.10
3	Cohen's *d*	0.742	2.48	.056	.461	6.14
4	Mean difference	0.950	6.79	.002	.883	23.58
	Pooled standard deviation	-0.173	-1.24	.284		
5	Mean difference	0.829	3.10	.054	.858	13.08
	Pooled standard deviation	-0.056	-0.22	.844		
	Cohen's *d*	0.186	0.55	.620		

Note. The mean difference, pooled standard deviation, and Cohen's *d* refer to the characteristics of the two groups that participants judged. The adjusted R^2 and *F* values refer to the model.

Modelling Individual Ratings

Next, we used multiple regression in order to model participants' individual ratings. This approach allows us to consider the potential influence of individual differences (sex, age, education level).

In Models 1-4 (see Table 3), we find that the mean difference between the two groups (water versus Brain Juice) is the best predictor of participants' ratings.

Table 3

Results of the Multiple Regression Analyses on Individual Ratings in Experiment 1

Model	Variable	Beta	*t*	*p*	Adjusted R^2	*F*
1	Mean difference	0.464	6.69	<.001	.211	44.78
2	Mean difference	0.466	6.73	<.001	.213	23.21
	Pooled standard deviation	-0.085	-1.22	.223		
3	Cohen's *d*	0.365	5.01	<.001	.128	25.12
4	Mean difference	0.409	3.39	<.001	.210	15.53
	Pooled standard deviation	-0.029	-0.25	.806		
	Cohen's *d*	0.089	0.59	.558		
5	Mean difference	0.590	2.33	.021	.198	6.74
	Education	0.089	0.69	.489		
	Sex	-0.079	-0.61	.544		
	Age	0.022	0.18	.860		
	Mean difference x Education	-0.237	-0.94	.350		
	Mean difference x Sex	0.229	0.81	.420		
	Mean difference x Age	-0.145	-0.66	.508		

Note. The mean difference, pooled standard deviation, and Cohen's *d* refer to the characteristics of the two groups that participants judged, and can be calculated from the information presented in Table 1. The adjusted R^2 and *F* values refer to the model.

While Cohen's *d* is a significant predictor of ratings (Model 3), the variance is better explained by the mean difference (Model 1). In Model 4, we include all three variables (the collinearity statistics remain within acceptable ranges) and find that only the mean difference is a significant predictor. Finally, the addition of the pooled standard deviation to Model 1 does not produce a statistically significant improvement, $R^2_{change} = .007$, $F_{change}(1,162) = 1.50$, $p = .223$. Equally, a stepwise linear regression, initially entering all three variables, also results in Model 1 as the best solution.

In Model 5, the demographic variables and their two-way interactions with the mean difference were included in the model, along with the mean difference itself. However, only the mean difference was a significant predictor of ratings. Indeed, if these additional variables and interactions are added to Model 1 in a second step, they provide no significant improvement over the original model, $R^2_{change} = .010$, $F_{change}(6,156) = 0.35$, $p = .911$.

Yes/No Decision

We carried out an independent samples *t*-test to compare participants who responded 'yes' versus 'no' when asked whether the Brain Juice group did better than the group that drank water. As expected, those who responded 'yes' had also given higher ratings ($M = 4.80$) than those who responded 'no' ($M = 2.87$), $t(163) = 5.95$, $p < .001$, $d = 1.09$. Indeed, a logistic regression with the participants' decisions ('yes' versus 'no') as the dependent variable and ratings as the predictor produced a statistically significant model, $\chi^2(1) = 31.83$, $p < .001$, Nagelkerke $R^2 = 0.264$. The odds ratio for the participants' ratings (1.84) meant that, for a one-unit increase along the rating scale, we expect an 84% increase in the odds of a 'yes' response. These results suggest that, although separable, the switch from responding 'no' to 'yes' occurred over a relatively small interval on our scale.

We also considered how well the three condition variables predicted participants' decisions. As individual predictors, we can say that Cohen's *d* provides a better fit to the data (Nagelkerke $R^2 = 0.094$) in comparison with the mean difference (0.047) and the pooled standard deviation (0.065). However, by considering combinations of these predictors, we find that a model including the mean difference and the pooled standard deviation provides the best fit, $\chi^2(2) = 12.30$, $p = .002$, Nagelkerke $R^2 = 0.108$. The addition of Cohen's *d* to this model produced no significant benefit, $\chi^2(1) = 0.12$, $p = .730$, while adding the mean difference ($p = .801$) or the pooled standard deviation ($p = .368$) to a model with Cohen's *d* as the predictor produced no significant improvements. Even considering our best model, we see that the mean difference and the pooled standard deviation together have a notably weaker influence on yes/no decisions in comparison with the rating given. Finally, adding these two predictors to the 'ratings alone' model also provides no significant improvement ($p = .058$).

Experiment 2

In the second experiment, we investigated whether the means and standard deviations of the simulated data would affect responses when tested *within participants*. That is, each participant was presented with 16 versions of the data/graph, and was asked to give their ratings for all versions. In this way, our design allows us to consider a more varied set of parameters, but also potentially highlights important changes (the means and spreads of the data) to the participants, given that they now see these parameters vary across the different graphs (in comparison with viewing only one graph in Experiment 1).

Method

Participants

Thirty-three students (31 women; age *M* = 18.76 years, *SD* = 1.77 years) at a university in Ontario, Canada, took part in exchange for course credits. Complete demographic information can be found in the Supplementary Material.

All participants provided written informed consent, and were given both a verbal and written debriefing after completion. The experiment's design and procedure were approved by the university psychology department's ethics committee and conform to the Declaration of Helsinki.

Materials

The materials used here were similar to those presented in Experiment 1, with a few important differences. First, all text and graphs were presented on a computer using custom MATLAB software, and responses were collected using the keyboard. Second, a fully crossed design (mean differences: 5, 10, 20, 40; pooled standard deviation: 2.5, 5, 10, 20) with all possible value combinations was used to create 16 conditions. In this experiment, each participant was presented with all 16 graphs (in a randomised order), and was required to provide a rating for each of them.

The children's test scores were generated as in Experiment 1 for each condition. However, new sets of test scores could be produced for each participant since this was a computer-based task. As such, every participant saw a graph where the mean difference was 10 and the pooled standard deviation was 20 (for example), but the raw data were newly generated for each instance. Using different graphs across participants, we were able to rule out any influence of particular distributions (e.g., the presence of outliers which might affect perceptions), in comparison with Experiment 1, where the same graph was always presented for a given condition.

Finally, only one response was required for each graph, and the question itself was reworded from Experiment 1. Here, participants were asked, "If you were asked how much Brain Juice improves memory, what would you say?" This rewording was used to give a more inferential tone, encouraging participants to predict general/future outcomes rather than simply describing the data presented. Participants entered their responses using a rating scale from 0 ('not at all') to 9 ('very large') on a 10-point scale.

Procedure

After reading the description of the fictional study onscreen and examining the graph depicted below, participants were required to respond to the rating-scale question using the keyboard. No time constraints were imposed. Once a response had been recorded, the raw data on the onscreen graph changed to reflect a new condition. Only the graph itself changed, while the text remained unaltered. Participants rated all 16 graphs in a random order.

Finally, demographic information was collected: age, sex, and how much statistics training/education participants had previously received (in months/years). In addition, participants were invited to provide written responses to two open-ended questions: "What were the main changes you noticed across the graphs that you rated?" and "How did the groups seem to differ from each other (when they did)?" Lastly, following Saito (2015), four statistical terms (mean, variance, standard deviation, effect size) were presented and participants were

asked to choose one of four options which best described their knowledge of each term: (1) don't know it, (2) have heard of it, (3) have learned it, and (4) can calculate it.

Results

Participant data and visual stimuli can be found in the Supplementary Material.

As in Experiment 1, the data were analysed in order to determine which factors predicted participants' judgements of the Brain Juice improvement. Several models were explored with participants' ratings as the dependent variable and condition variables (mean difference, pooled standard deviation, Cohen's *d*) as predictors. Unfortunately, age and sex could not be considered in our analyses since our student sample was largely homogeneous with regard to these variables.

Modelling Average Ratings

First, we averaged ratings across participants for each condition (data presented in the Supplementary Material). Using regression models, we then explored whether condition variables predicted mean responses for our participant sample as a whole. Note that averaging across conditions, while smoothing out potential noise due to individual differences in responses, resulted in only 16 observations. As such, any conclusions from these analyses are limited in their scope.

In Models 1-3 (see Table 4), we find that the mean difference between the two groups (water versus Brain Juice) is the best individual predictor of participants' ratings. While Cohen's *d* is a significant predictor of ratings (Model 3), the variance is better explained by the mean difference (Model 1). In Model 5, we include all three variables (the collinearity statistics remain within acceptable ranges) and find that both the mean difference and the pooled standard deviation are significant predictors. Indeed, the addition of the pooled standard deviation to Model 1 produces a statistically significant improvement, $R^2_{change} = .179$, $F_{change}(1,13) = 104.44$, $p < .001$, with the resulting model explaining 97.4% of the variance in average ratings (see Model 4). In comparison, the addition of Cohen's *d* to Model 1 also produces a significant improvement, $R^2_{change} = .092$, $F_{change}(1,13) = 10.94$, $p = .006$, but the resulting model explains only 87.4% of the variance. Therefore, the best model includes the mean difference and the standard deviation as predictors of averaged ratings. Equally, a stepwise linear regression, initially entering all three variables, also results in Model 4 as the best solution.

Table 4

Results of the Multiple Regression Analyses on Average Ratings in Experiment 2

Model	Variable	Beta	*t*	*p*	Adjusted R^2	*F*
1	Mean difference	0.894	7.46	<.001	.785	55.66
2	Pooled standard deviation	-0.423	-1.75	.103	.120	3.05
3	Cohen's *d*	0.799	4.98	<.001	.613	24.75
4	Mean difference	0.894	21.61	<.001	.974	285.68
	Pooled standard deviation	-0.423	-10.22	<.001		
5	Mean difference	0.874	13.63	<.001	.973	178.41
	Pooled standard deviation	-0.406	-6.91	<.001		
	Cohen's *d*	0.032	0.42	.684		

Note. The mean difference, pooled standard deviation, and Cohen's *d* refer to the characteristics of the two groups that participants judged. The adjusted R^2 and *F* values refer to the model.

Modelling Individual Ratings

Next, we used generalised linear mixed models in order to investigate participants' individual ratings, with each participant's unique ID included in the model as a random term to account for data collected repeatedly from the same person. Condition variables (mean difference, pooled standard deviation, Cohen's *d*) were included in the model as fixed factors.

In Models 1-3 (see Table 5), we find that the mean difference between the two groups (water versus Brain Juice) is the best individual predictor of participants' ratings, resulting in the lowest value for the corrected Akaike information criterion (AICc) of the three models. The inclusion of the pooled standard deviation as an additional predictor (Model 4) significantly improves this model, with a decrease in AICc of more than 10 (Burnham & Anderson, 2002). In comparison, we see no benefit if we add Cohen's *d* to Model 1, AICc = 2102.86. Although the small difference in AICc values for Models 4 and 5 means we cannot rule out Model 5 completely, we can say that there is "considerably less" support for Model 5 (Burnham & Anderson, 2002, p. 70), and Cohen's *d*, when added, remains a non-significant predictor.

Table 5

Results of the Generalised Linear Mixed Models on Individual Ratings in Experiment 2

Model	Variable	Beta	*t*	*p*	AICc	*F*
1	Mean difference	0.690	25.39	<.001	2177.44	644.64
2	Pooled standard deviation	-0.326	-8.46	<.001	2521.68	71.63
3	Cohen's *d*	0.617	20.21	<.001	2289.91	408.35
4	Mean difference	0.690	30.14	<.001	2014.37	555.87
	Pooled standard deviation	-0.326	-14.26	<.001		
5	Mean difference	0.675	19.63	<.001	2019.26	370.22
	Pooled standard deviation	-0.313	-9.95	<.001		
	Cohen's *d*	0.024	0.60	.548		

Note. The mean difference, pooled standard deviation, and Cohen's *d* refer to the characteristics of the two groups that participants judged. AICc is the Akaike information criterion, corrected for finite sample sizes. The *F* values refer to the model.

Next, we consider the inclusion of participants' individual differences as predictors. Using Model 4 (above) as our starting point, we find that the addition of how much statistics training each participant has had, along with this variable's interactions with the two original predictors, produces no improvement in the model (AICc = 2044.34, all additional coefficient *p*s > .16). Similarly, we find no benefit when we add in participants' rated knowledge of the terms 'Mean' and 'Standard Deviation', along with the interactions with their respective condition variables (AICc = 2020.30, all additional coefficient *p*s > .06). We find a very similar result when we consider participants' knowledge of 'Variance' and its interaction with the pooled standard deviation instead (AICc = 2021.74). As such, we find no evidence for improvements beyond Model 4.

Open-Ended Response Questions

Participants' written responses to the two open-ended questions ("What were the main changes you noticed across the graphs that you rated?"; "How did the groups seem to differ from each other (when they did)?") were not analysed formally. These questions were simply included in order to determine whether viewers noticed specifically that the spread of the data varied across the graphs. Although such observations do not guarantee

that the information was correctly utilised in the responses they had previously given, they at least confirm that this manipulation was salient to participants.

From reading the written responses, it is clear that participants noticed this particular change. (Remember that only two factors varied across the 16 graphs: the mean difference and the pooled standard deviation.) There were numerous mentions of “closer together”, “how far apart”, “spread out”, “distribution”, “grouped together (not scattered)”, and so on. We conclude from this coarse evidence that participants were explicitly aware that the variability of the data sets changed across trials. Our regression analyses (presented above) confirm that such information was used when participants gave their responses.

General Discussion

The aim of this research was to determine how people compare sets of data when these are presented visually in a way that was hypothesised to make both the averages and the variance within each group salient. By making this information accessible, we predicted that participants would utilise within-group variance in their decisions. Previous research has shown that people are heavily influenced by the mean difference (between-group variability) but place less (if any) importance on within-group variability (Masnick & Morris, 2008; Obrecht et al., 2007; Saito, 2015), although the majority of this research involved data presented simply as a list of values. Using a novel method of data visualisation, the results of Experiment 1 confirm that the mean difference predicted people’s decisions, while the pooled standard deviation did not appear to affect judgements. This was true both when ratings were averaged across participants and when individual responses were modelled. While the first experiment was a one-shot design (participants saw only one trial), the results of Experiment 2 demonstrate that participants use both pieces of information when responding to several trials in a sequence (again, individually and when ratings were averaged across the sample).

These two seemingly contrasting findings are likely the result of the two different experimental procedures used here. Experiment 1 suggests that, when faced with two sets of data, people do not naturally incorporate a consideration of variability when making decisions – their ratings are driven solely by the mean difference. Experiment 2, however, supports the idea that viewing 16 graphs in succession encourages the (correct) use of the pooled standard deviation. From their written responses, we know that participants noticed the changes in variability across trials. Viewers are, therefore, able to utilise information regarding the spread of the data in principle, but it may be important to draw their attention to this feature (here, through its manipulation over the course of the experiment). Unfortunately, we were unable to determine whether participants in Experiment 1, when presented with a single graph, considered variation and simply failed to incorporate it into subsequent decisions. This would be an important question for future research.

In Experiment 1, when forced to make a binary decision regarding the outcome of the Brain Juice intervention, we found that participants’ ratings were a strong predictor of their subsequent choices. However, our results also suggest that the mean difference *and* the pooled standard deviation were, to a lesser extent, predictors of their yes/no responses. This result might represent an interesting caveat. Future experiments could consider whether the nature of the response has an effect on what information is incorporated – perhaps, when forced to make a coarse, binary decision, participants implicitly take into account variation, having failed to do so during a

more fine-grained ratings judgement. Of course, the original result should first be replicated before any conclusions can be drawn.

The wording of the question in Experiment 1 ("…how large do you think the Brain Juice improvement was?") may have been interpreted as descriptive rather than inferential. Perhaps participants' responses were limited to the specific samples rather than a more generalised statement about the Brain Juice product and its effects. Importantly, showing that participants fail to incorporate variability information even in this situation is informative. In the second experiment, we reworded the question to perhaps imply a consideration of future outcomes beyond the specific samples presented ("if you were asked how much Brain Juice improves memory…"). Although a subtle difference, this may have helped participants to think more generally about treatment effects in a broader context. Unfortunately, we are unable to quantify the effects of this change (if any) within the current data, but further investigation might consider the influence of this type of framing on judgements.

We predicted that displaying the raw data visually would help participants to use the spreads of the two groups in their judgements. While such information may be perceived more easily than when simple lists of numbers were presented in previous work (Trumpower, 2015; Trumpower & Fellus, 2008; Trumpower et al., 2009), we did not make this comparison directly. It is worth noting that, even if variability information is made salient, participants are still required to understand the inverse relationship between within-group variability and effect size (i.e., larger spreads equate to smaller Brain Juice effects). However, Experiment 2 suggests that people do seem able to comprehend the effects of increased sampling variability. Further research might consider a direct comparison between data displayed using our parallel dot plots versus presenting lists of values.

In both experiments described here, we found no predictive effect of the level of statistical education that participants had previously received, or their knowledge of specific statistical terms, in line with previous research (Trumpower, 2015; Trumpower & Fellus, 2008; Trumpower et al., 2009). For Experiment 1, we attempted to represent a variety of ages and education levels in our sample, but the majority of participants reported mathematics/statistics education at the level of secondary school (i.e., until approximately 18 years of age). Ideally, in order to better investigate the influence of statistics education on judgements, we would include a more even spread of participants across different levels of education. In addition, perhaps a more detailed set of questions would allow us to distinguish between fine-grained educational achievements (e.g., GCSE, A-Level, etc.). For Experiment 2, we collected more detailed information regarding participants' specific statistics knowledge, but interestingly, this did not appear to influence responses either. In the current work, participants used both the means and the standard deviations to inform their decisions (when these concepts were made salient), but their use was not improved with increased statistical knowledge.

Perhaps another way to encourage participants to consider variability when making their judgements is to draw their attention to it explicitly. For two groups of participants, we might ask only one group to first judge whether the level of variability is the same or different in the two parallel dot plots. We predict that this group, who first considered the variability before making their difference judgements, would show a greater variability influence in their subsequent ratings. Again, this presumes that people would know what to do with this type of information, but Experiment 2 (presented here) suggests that this may well be the case.

Previous research has started to explore the effect of sample sizes on judgements (e.g., Obrecht et al., 2007). Here, each of our fictional groups comprised 20 children, and remained unchanged across conditions. Although

effect size measures are unaffected by sample sizes, researchers found that contingency judgements were lower for smaller samples (Clément, Mercier, & Pastò, 2002) and confidence in group differences decreased (Obrecht et al., 2007). Extending this idea, we might predict that larger samples would lead participants to judge group differences as larger (all other factors being held constant).

In conclusion, we extend previous research showing that participants fail to take into account the importance of within-group variability when making decisions about group differences. While information regarding the variability within the data does not seem to influence perceptions in isolation, we see that manipulating variability across situations may increase the salience of this factor, encouraging viewers to consider this additional source of information. Our results have important implications for statistical educational approaches, where visual displays may be a more suitable format for highlighting variability changes across items. In an applied context, these types of data sets might prove useful as a tool for conveying difficult concepts like variability to students.

Supplementary Materials

Participant data and visual stimuli. doi:10.6084/m9.figshare.4751095

Funding

The authors have no funding to report.

Competing Interests

The authors have declared that no competing interests exist.

Acknowledgments

The authors have no support to report.

References

Bakker, A., Biehler, R., & Konold, C. (2005). Should young students learn about box plots? In G. Burrill & M. Camden (Eds.), *Curricular development in statistics education* (pp. 163-173). Voorburg, The Netherlands: International Statistical Institute.

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodal inference: A practical information-theoretic approach* (2nd ed.). New York, NY, USA: Springer.

Carpenter, P. A., & Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied, 4*(2), 75-100. doi:10.1037/1076-898X.4.2.75

Clément, M., Mercier, P., & Pastò, L. (2002). Sample size, confidence, and contingency judgement. *Canadian Journal of Experimental Psychology, 56*(2), 128-137. doi:10.1037/h0087391

Cleveland, W. S., & McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science, 229*(4716), 828-833. doi:10.1126/science.229.4716.828

de Langhe, B., Fernbach, P. M., & Lichtenstein, D. R. (2016). Navigating by the stars: Investigating the actual and perceived validity of online user ratings. *The Journal of Consumer Research, 42*(6), 817-833. doi:10.1093/jcr/ucv047

delMas, R. C., Garfield, J., & Chance, B. L. (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. *Journal of Statistics Education: An International Journal on the Teaching and Learning of Statistics, 7*(3). Retrieved from http://ww2.amstat.org/publications/jse/secure/v7n3/delmas.cfm

Fouriezos, G., Rubenfeld, S., & Capstick, G. (2008). Visual statistical decisions. *Perception & Psychophysics, 70*(3), 456-464. doi:10.3758/PP.70.3.456

Gal, I. (1995). Statistical tools and statistical literacy: The case of the average. *Teaching Statistics, 17*(3), 97-99. doi:10.1111/j.1467-9639.1995.tb00720.x

Gleicher, M., Correll, M., Nothelfer, C., & Franconeri, S. (2013). Perception of average value in multiclass scatterplots. *IEEE Transactions on Visualization and Computer Graphics, 19*(12), 2316-2325. doi:10.1109/TVCG.2013.183

Makar, K., & Confrey, J. (2004). Secondary teachers' statistical reasoning in comparing two groups. In D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking* (pp. 353-373). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Masnick, A. M., & Morris, B. J. (2008). Investigating the development of data evaluation: The role of data characteristics. *Child Development, 79*(4), 1032-1048. doi:10.1111/j.1467-8624.2008.01174.x

Meletiou, M. M. (2000). *Developing students' conceptions of variation: An untapped well in statistical reasoning* (Unpublished doctoral dissertation). University of Texas, Austin, TX, USA.

Morris, B. J., & Masnick, A. M. (2015). Comparing data sets: Implicit summaries of the statistical properties of number sets. *Cognitive Science, 39*(1), 156-170. doi:10.1111/cogs.12141

Newman, G. E., & Scholl, B. J. (2012). Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. *Psychonomic Bulletin & Review, 19*(4), 601-607. doi:10.3758/s13423-012-0247-5

Obrecht, N. A., Chapman, G. B., & Gelman, R. (2007). Intuitive t tests: Lay use of statistical information. *Psychonomic Bulletin & Review, 14*(6), 1147-1152. doi:10.3758/BF03193104

Obrecht, N. A., Chapman, G. B., & Suárez, M. T. (2010). Laypeople do use sample variance: The effect of embedding data in a variance-implying story. *Thinking & Reasoning, 16*(1), 26-44. doi:10.1080/13546780903416775

Okan, Y., Galesic, M., & Garcia-Retamero, R. (2016). How people with low and high graph literacy process health graphs: Evidence from eye-tracking. *Journal of Behavioral Decision Making, 29*, 271-294. doi:10.1002/bdm.1891

Okan, Y., Garcia-Retamero, R., Galesic, M., & Cokely, E. T. (2012). When higher bars are not larger quantities: On individual differences in the use of spatial information in graph comprehension. *Spatial Cognition and Computation, 12*(2-3), 195-218. doi:10.1080/13875868.2012.659302

Pfannkuch, M. (2006). Comparing box plot distributions: A teacher's reasoning. *Statistics Education Research Journal, 5*(2), 27-45.

Pfannkuch, M., Arnold, P., & Wild, C. J. (2015). What I see is not quite the way it really is: Students' emergent reasoning about sampling variability. *Educational Studies in Mathematics, 88*(3), 343-360. doi:10.1007/s10649-014-9539-1

Reading, C., & Reid, J. (2005). Consideration of variation: A model for curriculum development. In G. Burrill & M. Camden (Eds.), *Curricular development in statistics education* (pp. 36-53). Voorburg, The Netherlands: International Statistical Institute.

Reading, C., & Reid, J. (2006). An emerging hierarchy of reasoning about distribution: From a variation perspective. *Statistics Education Research Journal, 5*(2), 46-68.

Saito, M. (2015). How people estimate effect sizes: The role of means and standard deviations. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 2075-2079). Austin, TX, USA: Cognitive Science Society.

Shah, P., & Carpenter, P. A. (1995). Conceptual limitations in comprehending line graphs. *Journal of Experimental Psychology: General, 124*(1), 43-61. doi:10.1037/0096-3445.124.1.43

Shah, P., & Freedman, E. G. (2011). Bar and line graph comprehension: An interaction of topdown and bottomup processes. *Topics in Cognitive Science, 3*, 560-578. doi:10.1111/j.1756-8765.2009.01066.x

Trumpower, D. L. (2015). Aspects of first year statistics students' reasoning when performing intuitive analysis of variance: Effects of within- and between-group variability. *Educational Studies in Mathematics, 88*(1), 115-136. doi:10.1007/s10649-014-9574-y

Trumpower, D. L., & Fellus, O. (2008). Naïve statistics: Intuitive analysis of variance. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 499-503). Austin, TX, USA: Cognitive Science Society.

Trumpower, D. L., Hachey, K., & Mewaldt, S. (2009). Persistence of naïve statistical reasoning concerning analysis of variance. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 3157-3162). Austin, TX, USA: Cognitive Science Society.

Watson, J. M., & Moritz, J. B. (1998). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics, 37*, 145-168. doi:10.1023/A:1003594832397