How People Estimate Effect Sizes: The Role of Means and Standard Deviations

Motoyuki Saito (m-saito@kwansei.ac.jp)
Department of Psychological Science, Kwansei Gakuin University
Hyogo 662-8501, JAPAN

Abstract

Many studies of causal judgments have dealt with the relation between the presence and the absence of a cause and an effect. However, little is known about causal learning with a continuous outcome. The present study adopted Cohen's *d* as an objective standard for effect size in situations where a binary cause influenced a continuous effect and investigated how people use means and standard deviations in the estimation of effect sizes. The experimental task was to read a scenario where the performance of two groups was compared and to infer the causal effect. Whereas means were manipulated while holding standard deviations constant in the mean difference group, standard deviations were varied with holding means constant in the standard deviation difference group. The results demonstrate that participants could respond appropriately to the difference in two means, and that they gave a higher estimate of effect size in large standard deviation situations than in small standard deviation situations. Judgments about standard deviations are in contrast to Cohen's *d*, indicating disproportionate attention to different kinds of data samples.

Keywords: causal learning; causal reasoning; intuitive statistics; effect size; continuous variable.

Introduction

Knowledge of causality is essential to explain past events, to control the present environment, and to predict future outcomes. Decision making based on causal knowledge enables us to achieve desired outcomes and avoid undesired consequences. In order to acquire precise knowledge of causal relations, we need to consider not only whether the causal relation exists, but also how much influence the cause has. When a teacher develops a new instruction method, for example, he or she has to examine whether the new instruction method has more educational effect than the previous one and how much improvement occurs. Scientists are accomplished at designing experiments and performing statistical analysis of the results. However, how do nonscientists estimate the influence of a cause on its effects, especially when outcomes are continuous values? The present study sheds light on this question.

The problem has been extensively investigated in the causal learning literature (Gopnik & Schulz, 2007; Shanks, Holyoak, & Medin, 1996; see also Holyoak & Cheng, 2011 for a review). Hume (1739/2000) argued that causal relations are not observable, and therefore must be induced from observable events; indeed, covariation among events serves as a fundamental cue for learning causal relations. Covariation is formally represented as a joint probability distribution for continuous variables and is specifically explained as the combination of presence and absence for binary variables. In a typical experimental situation, participants are asked to observe the states of the cause and its effect and then to judge the strength of the causal relation (e.g., Buehner, Cheng, & Clifford, 2003). It has been shown that both children and adults are quite sensitive to covariation information (e.g., Shultz & Mendelson, 1975; Wasserman, Elek, Chatlosh, & Baker, 1993). A classical and representative model is the ΔP rule (Jenkins & Ward, 1965). The ΔP rule is defined by subtracting the probability of the effect occurring when the cause is absent from the probability of the effect occurring when the cause is present (i.e., $\Delta P = P(\text{effect}|\text{cause}) - P(\text{effect}|\neg\text{cause})$). Positive ΔP values indicate a generative causal relation; negative ΔP values indicate a preventive causal relation. However, several studies have pointed out that covariation does not imply causation (Cheng, 1997) and that causal judgments do not always correspond to ΔP (e.g., Buehner et al., 2003; Shanks, 1985). Many models have been proposed, focusing on how people extract causal strength estimates from combinations of the presence and absence of cause and effect (Hattori & Oaksford, 2007; Perales & Shanks, 2007).

In contrast to the many empirical and theoretical studies using binary variables, little is known about causal learning with continuous variables (e.g., White, 2001; Young & Cole, 2012). Early studies have shown the difference in judgments between binary and multilevel variables. For instance, White (2001) examined causal learning from three level variables and revealed that causal judgment differed from correlational judgment in terms of sensitivity to confounding. In addition, White (2013) reported that cause-absent information carried greater weight than cause-present information, indicating that causal judgments about multilevel variables differed from those about binary variables. It has also been demonstrated that the interpretation of ambiguous values of a variable was depended on participants' hypothesis about the causal relation (Marsh & Ahn, 2009). Only a few studies have focused on causal learning from continuous variables. For example, Young and Cole (2012) showed participants were sensitive to the strength of the causal relation between two continuous variables in a video game task. Furthermore, Rashid and Buehner (2013) investigated whether people consider the base rate of the effect in situations where a binary cause produced a magnitude change on a continuous outcome. The results demonstrated that participants took the base rate of the effect into account in preventive scenarios, but not in generative scenarios. These findings differ from those obtained in previous studies using binary variables.

A crucial difference between binary variables and continuous variables is the distribution of data samples. In

order to estimate the effect of a binary cause on a continuous outcome precisely, one needs to consider not only the difference in the means, but also their distributions. One basic statistical measure appropriate for use in this situation is Cohen's *d* (Cohen, 1962, 1988), where one of the means from the two distributions is subtracted from the other and the result is divided by the pooled standard deviation for the variables:

$$d = \frac{M_E - M_C}{SD_{pooled}} \quad (1)$$

In this equation, M_E and M_C are experimental (E) and control (C) means and *SD* refers to the pooled standard deviation. According to this index, the effect size becomes larger as the difference in two means become large and as the standard deviations became small. Cohen's *d* is widely used in the psychology literature (Cumming, 2014; Fritz, Morris, & Richler, 2012).

The purpose of the present study was to investigate how people use means and standard deviations in the estimation of effect sizes. The experimental task was to read a scenario comparing the performance of two groups, and to infer the causal effect size. Means and standard deviations were systematically manipulated. If participants evaluate effect sizes in a manner consistent with calculating Cohen's *d*, their estimations should increase as the difference in means becomes large and as the standard deviations become small.

Method

Participants and design

A total of 42 undergraduates in an introductory psychology class participated in the experiment and received course credit. They were randomly assigned to either the mean difference group or the standard deviation difference group. In the mean difference group, means were manipulated while holding standard deviations constant. In contrast, standard deviations were varied while holding means constant in the standard deviation difference group. In addition, the effect sizes (Cohen's d = 0.5, 1.0, 2.0, 4.0) were manipulated within-participants. Each participant completed four causal learning tasks with different data sets.

Procedure

The participants' task was to respond to questions in a 10-page booklet written in Japanese. The first page outlined the experiment and asked for age and gender. On the second page, participants received the following instructions:

> Imagine that you are a teacher who tries to find a better way of teaching Japanese (English, Math, or Science) in a school. In order to improve students' academic performance, you developed a new instruction method that was different from a previous method, and investigated its effect. Students were divided into two homogeneous classes according to their academic ability. Whereas students in one class took lessons with the previous method, students in the other class took lessons with the new instruction method. Your task is to estimate the effect of the new instruction method on students' academic performance.

The academic subjects (e.g., science) were designed to distinguish each effect size condition.

Following the instructions, participants were informed about the detailed results of students' academic achievement. The information consisted of 40 exam scores. Half of the scores were obtained from students experiencing the previous instruction method (i.e., the control group in the cover story), and the other half of the scores were from students who had experienced the new instruction method (i.e., the experimental group in the cover story). Examination scores could vary from 0 to 100. Participants were required to look at the listed results thinking whether the new instruction method had an influence on the improvement in academic performance.

Table 1 depicts the data sets in each effect size condition. The different conditions were d = 0.5, d = 1.0, d = 2.0, and d = 4.0, resulting from the manipulation of means and standard deviations. As shown in Equation 1, Cohen's *d* is calculated by dividing the difference between two means by the pooled standard deviation. In the mean difference group, differences between means of the two groups were manipulated ($M_{\text{experimental}} - M_{\text{control}}$ = 5, 10, 20, 40) while keeping the standard deviations constant (all SD = 10). The larger the differences between two means were, the larger the effect sizes were. In the standard deviation difference group, in contrast, standard deviations of the two groups

Table 1
Data sets for each effect size condition by group

		Effect size condition							
		d = 0.5		d = 1.0		d = 2.0		d = 4.0	
Group		Control	Experimental	Control	Experimental	Control	Experimental	Control	Experimental
Mean difference	*M*	40	45	40	50	40	60	40	80
	SD	10	10	10	10	10	10	10	10
Standard deviation difference	*M*	40	50	40	50	40	50	40	50
	SD	20	20	10	10	5	5	2.5	2.5

were manipulated ($SD_{experimental}$ = $SD_{control}$ = 20, 10, 5, 2.5) while holding differences between the two means constant ($M_{experimental}$ − $M_{control}$ = 10). The smaller the standard deviations were, the larger the effect sizes were. The examination scores in each condition were designed to be normally distributed. Due to the constraints of natural numbers, sample size, and normal distribution, the calculated effect sizes in the standard deviation difference group were slightly different from those of the mean difference group (e.g., d = 1.99, 4.05 in the standard deviation difference group; d = 2.00, 4.00 in the mean difference group). These effect size conditions enabled me to investigate whether people were sensitive to means and standard deviations in estimating effect size.

On the third page of the task booklet, participants were asked to infer the effect of the new instruction method on students' academic performance. Specifically, the question was "To what extent does the new instruction method have an influence on the improvement of the academic performance in Japanese (English, Math, or Science)?" A rating was made on a scale from 0 (the new instruction method does not cause an improvement at all) to 100 (the new instruction method causes a great improvement). In addition to estimating effect size, participants reported their confidence in the judgment with a scale ranging from 0 (not confident at all) to 100 (extremely confident). Then, participants completed the next effect size condition in a similar procedure. The order of the effect size conditions was counterbalanced across participants using a Graeco-Latin square design.

The last page of the booklet consisted of questions about statistical knowledge. In particular, the instructions stated statistical terms and asked participants to choose one of four options about each term: (1) don't know it, (2) have heard of it, (3) have learned it, and (4) can calculate it. The statistical terms included mean, variance, standard deviation, and effect size. These questions were added for exploratory reasons.

Results

Responses to questions about statistical knowledge revealed that none of the participants could calculate effect size. Figure 1 (left panel) shows the mean ratings of effect size in each condition. In the mean difference group, higher estimations were obtained as the effect size became large. In contrast, there was a small reduction in the estimations of the standard deviation difference group. A two-way mixed ANOVA with type of statistic (mean difference, standard deviation difference) as a between-participants factor and effect size condition (0.5, 1.0, 2.0, 4.0) as a within-participants factor yielded significant main effects of type of statistic, $F(1, 40) = 6.44$, $MSE = 1035.96$, $p = .015$, $\eta^2_G = .091$, and effect size condition, $F(3, 120) = 3.14$, $MSE = 208.52$, $p = .028$, $\eta^2_G = .029$. The interaction between type of statistic and effect size condition was also significant, $F(3, 120) = 15.21$, $MSE = 208.52$, $p < .001$, $\eta^2_G = .125$. Subsequent tests of the simple main effects of effect size condition were significant for both the mean difference, $F(3, 60) = 17.47$, $MSE = 164.48$, $p < .001$, $\eta^2_G = .242$, and standard deviation difference groups, $F(3, 60) = 3.77$, $MSE = 252.55$, $p = .015$, $\eta^2_G = .068$. In the mean difference group, individual comparisons showed ratings in the d = 0.5 condition to be significantly greater than those in the d = 4.0 condition ($p < .001$ with a Shaffer correction). However, the opposite pattern occurred in the standard deviation

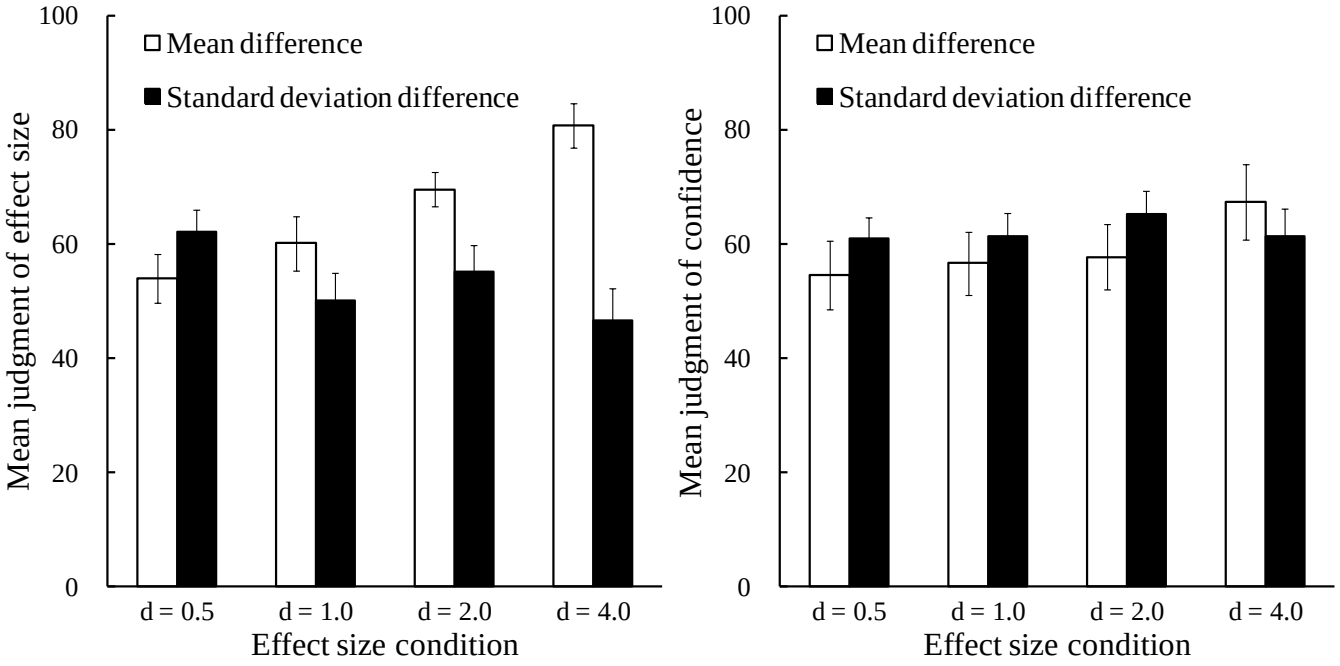

Figure 1. Mean judgment of effect size (left panel) and confidence (right panel) in each effect size condition. The error bars represent standard errors of the mean.

difference group (p = .035 with a Shaffer correction). Participants estimated higher effect sizes as the difference between two means became larger, but not as the standard deviations became smaller.

The mean confidence ratings for each condition are presented in the right panel of Figure 1. There was a small increase in confidence as a function of effect size in the mean difference group. A 2 (type of statistic) × 4 (effect size condition) mixed ANOVA was performed, with type of statistic as a between-participants factor and effect size condition as a within-participants factor. A main effect of effect size condition was found, $F(3, 120) = 2.96$, $MSE = 122.68$, $p = .035$, $\eta^2_G = .012$. The interaction between type of statistic and effect size condition was also significant, $F(3, 120) = 3.29$, $MSE = 122.68$, $p = .023$, $\eta^2_G = .013$. Subsequent tests for the simple main effects of effect size condition were significant within the mean difference group, $F(3, 60) = 5.78$, $MSE = 117.96$, $p = .002$, $\eta^2_G = .033$, but not in the standard deviation difference group, $F(3, 60) = 0.66$, $MSE = 127.41$, $p = .579$, $\eta^2_G = .009$, showing that participants were more sensitive to the means than to the standard deviations.

In summary, participants could respond appropriately to the difference in two means. However, they gave a higher estimate of effect size in large standard deviation situations than in small standard deviation situations. The results of the confidence ratings revealed that participants were confident in the evaluation of effect size when two means differed greatly from each other.

Discussion

The present study adopted Cohen's d as an objective standard for effect size in situations where a binary cause influenced a continuous effect, and investigated how people use means and standard deviations in the estimation of effect sizes. The results show that participants judged greater effect sizes as the difference between two means became large, and smaller effect sizes as the standard deviations became small. Although the rank order of the estimates corresponded to the order of Cohen's d in the mean difference group, the opposite pattern was obtained in the standard deviation difference group. One possible explanation for this pattern of the results is differential weighting for each data sample. That is, people pay more attention to information consistent with their hypothesis and less attention to disconfirming evidence. When standard deviations of two groups are large, some data samples in one group are much higher than the average of the other group. Judgments based on these samples result in a higher estimation of effect size. When both groups have a small standard deviation, in contrast, data samples are not overlapping and are spaced closely within the group. Therefore, each sample would be weighted similarly. Indeed, many studies about causal learning with binary variables have pointed out that each type of information has differential weighting in judgments (e.g., Kao & Wasserman, 1993; Mandel & Vartanian, 2009). For example, Kao and Wasserman (1993) reported that information about the presence of both cause and effect is given more weight than other type of information.

The present study adopted continuous variables that were normally distributed, and investigated the effect of mean difference on the estimation of effect sizes. One methodological limitation of the current experiment is that we cannot discriminate whether participants responded to mean, median, or mode. This is because these indices are equal in normal distributions. Since people are limited in their memory capacity, they need to summarize data samples in some way. An intriguing question for future study is to investigate how people summarize continuous values. Using skewed distributions would enable us to differentiate mean, median, and mode. Peterson and Miller (1964) demonstrated that participants were sensitive to median and mode, but not to mean, when probability distributions were highly skewed.

Another key question to be addressed is the role of sample size and experimental design. Effect sizes are indices unaffected by sample size. However, evidence from studies about causal learning with binary variables indicates that causal judgments become greater as the number of samples increases (Clément, Mercier, & Pastò, 2002; Shanks, 1985). Furthermore, studies of correlation judgments suggest that the detection of correlation becomes easier as sample size increases (Anderson & Doherty, 2007), except for very specific situations (Kareev, 1995). These extensions will provide converging evidence to investigate the effect of sample size on the estimation of effect sizes in situations where a binary cause influences a continuous effect. In addition, it is also valuable to examine the difference between within-subjects and between-subjects designs. Since previous studies about causal learning have relied heavily on the situations with independent samples, little is known about causal judgments for repeated measurements (cf. Rottman & Keil, 2012). These investigations will shed more light on the question of how people estimate effect sizes.

References

Anderson, R. B., & Doherty, M. E. (2007). Sample size and the detection of means: A signal detection account. *Memory & Cognition, 35*, 50-58.

Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: A test of the assumption of causal power. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 1119–1140.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*, 367–405.

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

Cohen, J. (1962). The statistical power of abnormal social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–153.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cumming, G. (2014). *The new statistics: why and how. Psychological Science*, *25*, 7–29.

Doherty, M., Anderson, R., Angott, A., & Klopfer, D. (2007). The perception of scatterplots. *Perception & Psychophysics*, *69*, 1261–1272.

Doherty, M. E., Anderson, R. B., Kelley, A. M., & Albert, J. H. (2009). Probabilistically Valid inference of covariation from a single x, y observation when univariate characteristics are known. *Cognitive Science*, *33*, 183–205.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology. General*, *141*, 2–18.

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics*. Lawrence Erlbaum Associates, Inc.

Gopnik, A., & Schulz, L. E. (2007). *Causal learning: Psychology, philosophy, and computation*. New York: Oxford University Press.

Hattori, M., & Oaksford, M. (2007). Adaptive non-interventional heuristics for covariation detection in causal induction: Model comparison and rational analysis. *Cognitive Science*, *31*, 765–814.

Hume, D. (1739/2000). *A treatise of human nature*. Oxford: Oxford University Press.

Holyoak, K, J., & Cheng, P, W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual Review of Psychology, 62*, 135-163.

Jenkins, H., & Ward, W. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs, 7*, 1-17.

Kao, S.-F., & Wasserman, E. A. (1993). Assessment of an information integration account of contingency judgment with examination of subjective cell importance and method of information presentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 1363–1386.

Kareev, Y. (1995). Through a narrow window: Working memory capacity and the detection of covariation. *Cognition*, *56*, 263–269.

Mandel, D. R., & Vartanian, O. (2009). Weighting of contingency information in causal judgement: evidence of hypothesis dependence and use of a positive-test strategy. *Quarterly Journal of Experimental Psychology*, *62*, 2388–2408.

Marsh, J. K., & Ahn, W.-K. (2009). Spontaneous assimilation of continuous values and temporal information in causal induction. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*, 334–352.

Perales, J. C., & Shanks, D. R. (2007). Models of covariation-based causal judgment: A review and synthesis. *Psychonomic Bulletin & Review*, *14*, 577–596.

Peterson, C., & Beach, L. (1967). Man as an intuitive statistician. *Psychological Bulletin*, *68*, 29–46.

Peterson, C., & Miller, A. (1964). Mode, median, and mean as optimal strategies. *Journal of Experimental Psychology*, *68*, 363–367.

Rashid, Ahmad, Azad, Ab, & Buehner, M. J. (2013). Causal reasoning with continuous outcomes. In M. Knauff, M., Pauen, N., Sebanz, & I. Wachsmuth (Eds.) *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 115-120). Austin TX: Cognitive Science Society.

Rottman, B. M., & Keil, F. C. (2012). Causal structure learning over time: Observations and interventions. *Cognitive Psychology*, *64*, 93–125.

Shanks, D. R. (1985). Continuous monitoring of human contingency judgment across trials. *Memory & Cognition*, *13*, 158–167.

Shanks, D. R., Holyoak, K. J., & Medin, D. L. (1996). Causal learning: The psychology of learning and motivation (Vol. 34). San Diego: Academic.

Shultz, T., & Mendelson, R. (1975). The use of covariation as a principle of causal analysis. *Child Development*, *46*, 394-399

Spellman, B. A. (1996). Acting as intuitive scientists: Contingency judgments are made while controlling for alternative potential causes. *Psychological Science*, *7*, 337-342.

Wasserman, E. A., Elek, S. M., Chatlosh, D. L., & Baker, A. G. (1993). Rating causal relations: Role of probability in judgments of response-outcome contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 174–188.

White, P. A. (2001). Causal judgments about relations between multilevel variables. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 499–513.

White, P. A. (2013). Causal judgement from information about outcome magnitude. *Quarterly Journal of Experimental Psychology*, *66*, 2268–2288.

Young, M. E., & Cole, J. J. (2012). Human sensitivity to the magnitude and probability of a continuous causal relation in a video game. *Journal of Experimental Psychology: Animal Behavior Processes*, *38*, 11–22.