UNDERSTANDING CONFIDENCE INTERVALS WITH VISUAL REPRESENTATIONS

Bilgin NAVRUZ [*]
Erhan DELEN [**]

ABSTRACT
In the present paper, we showed how confidence intervals (CIs) are valuable and useful in research studies when they are used in the correct form with correct interpretations. The sixth edition of the APA (2010) *Publication Manual* strongly recommended reporting CIs in research studies, and it was described as "the best reporting strategy" (p. 34). Misconceptions and correct interpretations of CIs were presented from several textbooks. In addition, limitations of the null hypothesis statistical significance test (NHSST) were discussed, and using CIs was discussed as an alternative to the NHSST. Finally, the calculation and the visual representation of CIs for mean and effect size were illustrated to help readers comprehend the concept of CIs.
Keywords: confidence intervals, misconception, correct interpretation, confidence intervals for effect sizes.

GÖRSEL SUNUM İLE GÜVEN ARALIKLARI KAVRAMINI ANLAMA

ÖZET
Bu çalışmada doğru kullanım ve yorumlama ile güven aralıklarının araştırmalarda sunulmasının önemi üzerinde durulmuştur. American Psychological Association (APA) (2010) yayım kılavuzu, güven aralıkları değerlerine çok önem vermekte ve çalışmalarda rapor edilmesi gerektiğini belirtmektedir. Araştırmacılar tarafından eksik ve yanlış da yorumlanabilen güven aralıkları bu çalışmada ayrıntılı olarak ele alınmış, kitaplardan da örnekler verilerek doğru ve yanlış tanımlar değerlendirilmiş ve görsel sunum ile örneklendirilerek okuyucuların güven aralıkları konusunu daha kolay anlamaları amaçlanmıştır.
Anahtar sözcükler: güven aralıkları, kavram yanılgısı, doğru yorumlama, etki büyüklüğü için güvenaralıkları.

[*] PhD Student, Texas A&M University, e-mail: bilgin@neo.tamu.edu
[*] Assistant Professor, Giresun University, e-mail: erhan.delen@giresun.edu.tr

1. INTRODUCTION

In educational research studies, we usually use statistics in order to estimate the parameter of the population. Because it is often impossible to access all data of the population, we use special calculations to make the best estimations. Estimating the confidence intervals (CIs) is a method of telling something about the population by using data from the sample. In addition, sample statistics are used for two types of estimation: point and interval. Point estimation is used to estimate the population parameter when the value of sample statistics is obtained. Even though we would like to know a fixed population parameter by using a sample statistic (e.g., using M to estimate μ), it is not an easy step because sample statistic and population parameter will most likely be unequal. At that point, interval estimation is a way to characterize the uncertainty of our estimate (Kline, 2004). Cumming and Finch (2001) described the CIs as follows:

> CIs provide a mechanism for making statistical inferences that give information in units with practical meaning for both the researcher and the reader. They give a best point estimate of the population parameter of interest and an interval about that to reflect likely error—the precision of the estimate. (p. 533)

Cumming and Finch (2001) listed four main advantages of CIs: (a) CIs are easier to understand and interpret because CIs give point and interval information; (b) null hypothesis statistical significance testing (NHSST) is directly related with CIs; (c) CIs help understand previous studies and support meta-analysis and meta-analytic thinking; (d) CIs could be estimated before and after an experiment, which also provides information about precision.

Reporting CIs in studies has many advantages. For example, comparing the intervals from previous studies gives us a better picture of the population parameters (Wilkinson & APA Task Force on Statistical Inference, 1999). In the sixth edition of APA (2010) *Publication Manual*, reporting the CIs is strongly recommended and it is stated as "the best reporting strategy" (p.34).

Social science researchers do not tend to report CIs in their studies even though its importance is widely accepted (Cumming & Finch, 2001; Finch, Cumming, & Thomason, 2001). They just produce statements, whether the result is significant or not, by using NHSST and erroneously believe NHSST evaluates result replicability (Thompson, 1996). They use *p*-values in order to come to conclusions. For example, Di Stefano (2004) reviewed 45 published papers in a journal, *Forest Ecology and Management*, and found that only five of them had CIs reported. Byrd (2007) suggested using CIs with NHSST as follows: "Reporting exact *p* values is warranted and should be used in combination with CIs. However, considering that statistical significance is influenced by sample size, reporting only exact *p* values is not recommended." (p. 388). Cumming and Finch (2001) also pointed out the importance of using CIs and suggested to report them for effect size measures as well.

Byrd (2007) reviewed the quantitative studies, which were published from 1997 to 2006 by *Educational Administration Quarterly* (EAQ), and the author found that CIs were not reported and interpreted properly between these years even though EAQ is one of the journals, which requires authors to follow the rules of the APA *Publication Manual* in its studies. According to Thompson (2002), CIs are not reported in research studies because "It is conceivable that some researchers may not fully understand statistical methods that they (a) rarely read in the literature and (b) infrequently use in their own work" (p. 26). Later studies supported Thompson's view on this issue. For example, Cumming, Williams, and Fidler (2004) and Belia, Fidler, Williams, and Cumming (2005) indicated that CIs have not been truly understood by many researchers in psychology, behavioral neuroscience, and medicine; and results showed that authors have had serious misconceptions about CIs.

2. MISCONCEPTIONS ABOUT CIs

The estimation of CIs is very straightforward. On the other hand, there are many misconceptions about interpreting CIs. According to the APA Task Force on Statistical Inference (Wilkinson & APA Task Force on Statistical Inference, 1999), the common misunderstanding is "assuming a parameter is contained in a confidence interval" (p. 599).

The common mistake when interpreting an estimated CI in a given study is the consideration of only that study itself. However, it is critical that the interpretation of the CI in that study should be considered with CIs from previous related studies together (Fidler & Thompson, 2001; Thompson, 2002, 2006a). Thompson (2006a) emphasized that "the certainty level involved in constructing a given sample CI applies to constructing infinitely many CIs drawn from a population, and not to the single CI constructed in a single sample" (p. 204). We can also draw this conclusion if we think of estimating the CIs. In the estimation, we use the sampling distribution (computing of standard error) in order to apply the procedure properly. We know that the sampling distribution consists of statistics of infinite samples. Therefore, when we interpret the CIs, we should consider that the result is not coming from a single sample. The correct interpretation of a 95% CIs was given by Thompson (2007) as "if we drew infinitely many random samples from the population, exactly 95% of the CIs would capture the parameter, and exactly 5% would not" (p. 427). It is important to note that using CIs also helps us to understand the results across prior studies, and the results of prior studies versus our current study (Fidler & Thompson, 2001; Thompson, 2002, 2006a).

Thompson (2006a) criticized some researchers for interpreting large confidence levels (95% or 99%) as "100% certainty" (p. 203). Because 100 does not equal 99 or 95, we should always keep in mind that if our levels are lower than 100%, there is always another probability against certainty. For example, we have 19 blue balls and 1 red ball in a box. If I randomly pick a ball from this box, the probability of grabbing a blue ball is 95%. I cannot say that the probability of getting a blue ball is 100% even though this 95% percentage is close to 100%.

Another common misconception is the view that CIs do not do more than the NHSST (e.g., Hagen, 1997; Knap & Sawilowsky, 2001). The application of this view is that if

our CI subsumes zero, we are not able to reject the null hypothesis. On the other hand, if our CI does not subsume zero, we are able to get the same statistically significant result from the NHSST. Thompson (2001) stated that this aspect of CI use and NHSST use could be clearly distinguished because NHSST cannot be conducted unless there is a null hypothesis; on the other hand, CIs can be drawn without a null hypothesis. Also, Schmidt (1996) indicated that researchers do not have to interpret CIs regardless of whether they subsume zero to do statistical significance testing.

Another good answer to researchers who believe that CIs do nothing more than NHSST is that in the NHSST, there is only one hypothesis to test, but we can test multiple hypotheses with a given CI; values captured by intervals are more likely than values outside, and values close to the point estimate are more likely than values at the ends (Fidler & Loftus, 2009). Instead of interpreting CIs by looking at whether or not they subsume hypothesized parameters, researchers should interpret them by comparing them with CIs from previous related studies to find true population parameters via meta-analytic thinking (Thompson, 1998, 2002, 2006a).

Some researchers wrongly interpret their 95% CIs as "I can be 95% certain that my 1-(α=.05) CI captured the true population parameter" (Thompson, 2006b, p. 592). We cannot interpret a given CI in this way because "our confidence interval comes from an infinite sequence of potential confidence intervals" (Cumming & Finch, 2005, p. 171). Similarly, saying there is 95% probability that our 95% CI captures the true population parameter is misleading because the probability of capturing the true population parameter by a single CI is "1 or 0" (Cumming & Fidler, 2009, p. 5). Also, Thompson (2002) indicated that "a given interval either does or does not capture the parameter. This is a binary outcome with only these discrete possibilities, just as one can only be pregnant or not pregnant, but cannot be 95% pregnant" (p. 27). Thus, researchers should avoid using probability statements for CIs. If we say that there is 95% probability that a 95% CI captures the true population parameter, it might be understood that the population parameter is a variable, but in fact, population parameters are fixed values (Cumming, 2011; Cumming & Finch, 2005). When we interpret CIs, we should never forget Thompson's inequality: $1 \neq \infty$ (Thompson, 2006b). According to Thompson (2007), we cannot say we are X% confident with our drawn X% CI capturing the true population parameter. Instead, if we drew infinitely many CIs in a given level (X), X% of these CIs would capture the true population parameter, and (100-X)% would not capture the true population parameter (Thompson, 2006a).

In the previous paragraphs, the common misconceptions and true interpretations of them were explained. Now, we will examine some textbook definitions and interpretations for CIs to make our discussion more concrete.

2.1. Correct and Incorrect Definitions for CIs from Textbooks

Keller and Warrack (2003) presented the correct definition of CIs, and they explained CIs with a few examples. An X % CI was described as if the same sample size was repeatedly chosen from a population and if X% CIs were drawn for these samples, X% of these CIs would include true population parameter, and (1-X) would not, and also it was stated that the probability of a given CI that includes the true population parameter is 1 or 0.

Good and Hardin (2003) also gave a useful definition and interpretation for CIs:

> A common error is to misinterpret the confidence interval as a statement about the unknown parameter. It is not true that the probability that a parameter is included in a 95% confidence interval is 95%. What is true is that if we derive a large number of 95% confidence intervals, we can expect the true value of the parameter to be included in the computed intervals 95% of the time. (p. 101)

This is a correct definition of CI and another good explanation of why some of the definitions below have been criticized because of their statements about probability.

Ott and Longnecker (2010) defined CIs for mean as "When using the formula $M \pm 1.96\ \sigma/\sqrt{n}$; that is, 95% of the time in repeated sampling, intervals calculating using the formula $M \pm 1.96\ \sigma/\sqrt{n}$ will contain the mean μ" (p. 226). (The value 1.96 corresponds to $z_{.95}$ in the z distribution table for two-tailed test). This is another correct interpretation for CIs.

Oakes (1986) stated that CI is a "plausible range of values for the unknown population parameter" (p. 52). Even though the definition is correct, much more should be said about CIs.

Lomax (2001) also gave an interpretation of CI:

> If we form 68% confidence intervals for 100 sample means, then 68 of those 100 intervals would contain or include the population mean. Because the applied researcher typically only has one sample mean and does not know the population mean, he or she has no way of knowing if this one confidence interval actually contains the population mean or not. (pp. 87-88)

This interpretation is true but not clear enough. Again, we will refer to Thompson's inequality ($1 \neq \infty$) (Thompson, 2006b) because in the present definition, the author seems to be confident that 68 of 100 CIs totally capture the true parameter. However, it cannot be said that 68 out of 100 include the true population parameter because these all 100 CIs may or may not include the true population parameter, or the number of CIs that include the true population parameter might be any number between 0 and 100. It is true that *on average* 68 of 100 CIs capture the true population parameter for repeated samples, but we cannot certainly say 68 of 100 CIs capture true population parameter.

Lynch (2007) explained CIs as such: "a 95% CI is that 95% of the CIs that could be drawn from the sampling distribution for sample mean would capture the population μ" (p. 342). Even though his explanation is correct, in his book he suggested CIs as an alternative of statistical inference, but as we explained above, the meaning of CIs is more than NHSST.

Pagano (2009) defined a CI as "a range of values that probably contains the population value" (p. 331), and he gave the interpretation for 95% CI as "an interval such that the probability is 0.95 that the interval contains the population value" (p. 331). This is a

deficient definition for CIs because we cannot talk about probability such as 0.90 or 0.95 for a given CI because the probability of a given CI regardless of whether it includes the true population parameter is "1 or 0" (Cumming & Fidler, 2009, p. 5).

An additional erroneous definition for CIs includes the misconception that "An interval, with limits at either end, [has] a specified probability of including the parameter being estimated" (Howell, 2008, p. 296). This is an incorrect statement for defining CIs because we should never interpret a given confidence level as a probability statement for a given CI. Instead, as it is stated above, the probability of a given CI with some confidence level is "1 or 0", not 95% or 99% (Cumming & Fidler, 2009, p. 5).

As seen from the above CI definitions, authors may have misconceptions about the probability of a single CI interval. They state confidence level as a probability for a given single CI, but we can only say for a given CI whether or not it includes the true population parameter is "1 or 0" (Cumming & Fidler, 2009, p. 5). And, the other misconception is that they are not aware of $1 \neq \infty$ (Thompson's inequality). We should not forget that our single 95% CI does not capture the population parameter with a 95% chance. Instead, if we drew infinitely many CIs in the same sample size from the same population, 95% of the CIs would include the true population parameter and 5% of them would not include the true population parameter (Thompson, 2007).

3. VISUAL REPRESENTATIONS

3.1. Understanding CIs based on Cumming's (2011) Book and Software

So far, we have explained how to understand CIs, some common misconceptions, and true interpretations of these misconceptions. Now, more interpretations and explanations for CIs will be given by using graphical representations created through the acclaimed software, the Exploratory Software for Confidence Intervals (ESCI) developed by Geoff Cumming (see Cumming, 2011; Cumming & Finch, 2001, 2005). These interpretations will be mainly based on Cumming's book (2011). The software works under Excel 2007 and 2010 and could be downloaded at the following link: http://www.latrobe.edu.au/psy/research/projects/esci

Correct Interpretation of CIs and Graphing Them by ESCI

We can never say that our single 95% CI captures the true population parameter with 95% chance (Thompson, 2006a). Instead, the true interpretation is that if infinitely many CIs were drawn in a given sample size, 95% of these CIs would capture true population parameter, and 5% would not capture the true parameter (Thompson, 2007). The APA Task Force on Statistical Inference (Wilkinson & the APA Task Force on Statistical Inference, 1999) and the APA *Publication Manuals* (2001, 2010) strongly recommended the use of graphical representations whenever possible; and the ESCI program gives perfect graphical representations for the above statement. Figure 1 somewhat illustrates "Thompson's inequality" (Thompson, 2007, p. 427) ($1 \neq \infty$) by drawing CIs by using the ESCI.

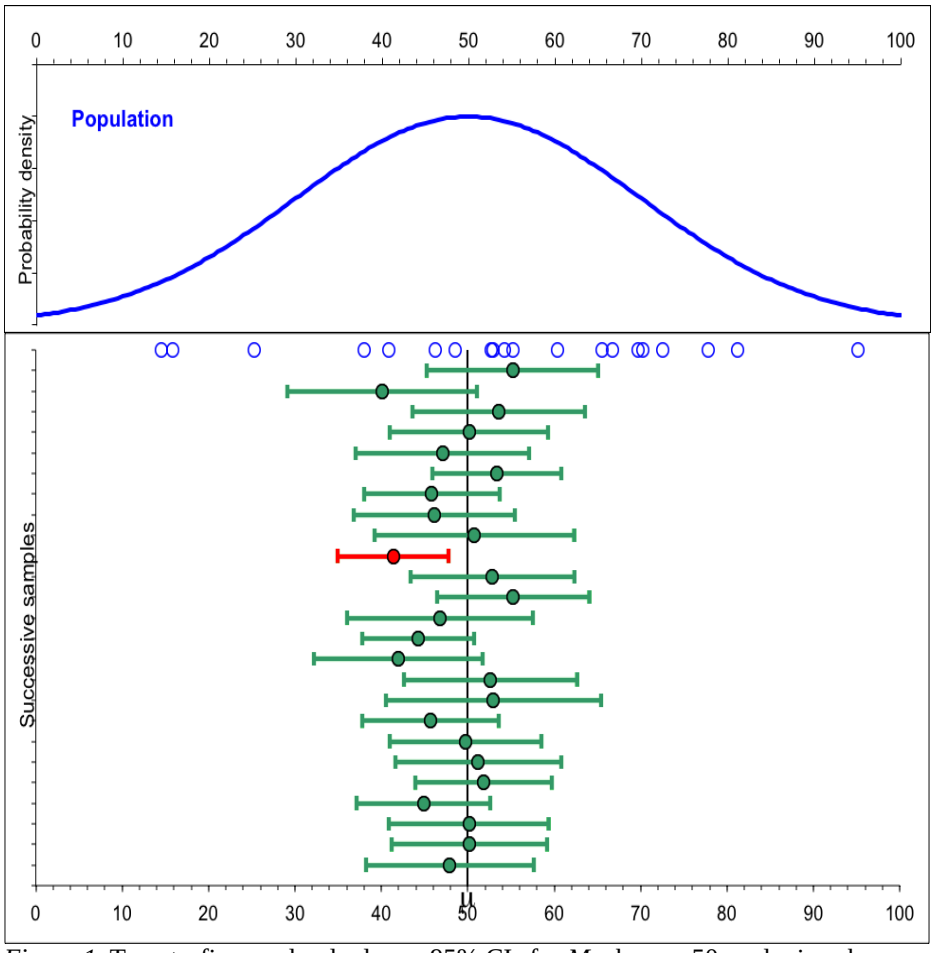

Figure 1. Twenty-five randomly drawn 95% CIs for *M* when μ=50, and σ is unknown

In the present example, a population of normally distributed scores has been created. The population standard deviation (σ) is unknown and the population mean (μ) is 50. This population is shown in the top of Figure 1. Every randomly drawn sample includes 20 scores, and the first sample scores are shown with small empty circles. The 95% CI for the mean is presented below the small 20 circles representing the scores in the first sample. The M_x for the first sample was about 55 and 95% CI for the mean of 55 for the first sample ranged from ~45 to ~65.

The second part of Figure 1 shows 25 95% CIs for the mean. Not all widths of CIs for the mean are equal because the population standard deviation is an unknown parameter. If we knew the population standard deviation, then all widths of CIs for the mean would be equal. The width of the CI is specified by the standard error and the confidence level for mean, so if the standard error is bigger, the width of the intervals in the same confidence level will be bigger. We would like to find small standard errors to be more confident about our point estimate. Thus, CIs for mean give us information about point estimate and precision of this point estimate because we can see the standard error via

the width of the intervals (Cumming, 2011; Thompson, 2006b). Even though an estimated interval width is narrower than other intervals, it might not capture the true population parameter. For example, in Figure 1, the tenth CI is the narrowest one, and it does not capture the true population mean. If we did not know the true population mean (μ =50), we would think that the tenth interval would give us more precision for our point estimate. However, our precise point estimate would be wrong even though the CI length is narrow (Thompson, 2006a).

Figure 1 presents 25 CIs and one of them does not capture the true population parameter. It might be logical to say *in average, 95% of the drawn 95% CIs will capture the true population parameter*, but we should never forget that if we drew infinitely many 95% CIs, 95% of them would capture the true population parameter (Thompson, 2006b). In the present example, 96% (1/25) of 95% CIs capture the true population parameter; it is close to 95%, but this percentage could be 100% or 0% for these 25 95% CIs for mean.

Plausibility of Estimations Based on Cat's Eye Picture

When we are estimating a population parameter, CIs provide valuable information for this parameter, because the values in the range of CIs are plausible population parameters, but we should never forget that our parameter might be outside of a particular CI (Cumming, 2011). For instance, if we found our 95% CI as [2.50, 7.50], it plausible to say that the population parameter would be between this interval, but we should remember that the parameter might be outside of this interval because of Thompson's inequality (see Thompson, 2007). The values that are close to our sample mean (M=5) are more plausible for μ than the values at the end of the intervals: also the values in the range of intervals are more plausible for μ than values beyond intervals, but the values which are outside of the intervals are not totally implausible for μ (Cumming, 2011; Cumming & Finch, 2005). The "Cat's Eye" Confidence Interval in Figure 2, which has been drawn by using ESCI, will illustrate this information.

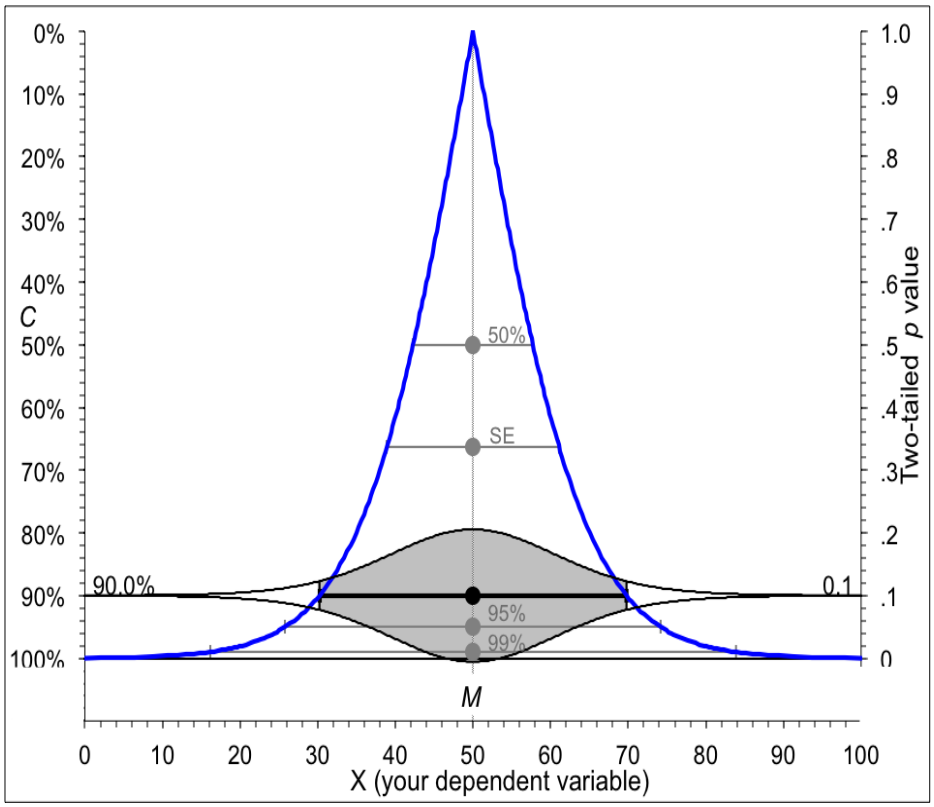

Figure 2. Cat's-Eye Confidence Interval.

In Figure 2, the shadow area around the 90% CI is called the "Cat's Eye" Confidence Interval, and it gives us information about plausibility (Cumming, 2011). It can be drawn for every confidence level such as 50%, 90%, 95%, 99%, etc. "The cat's-eye picture describes how the plausibility, or relative likelihood, that a value is μ is greatest at *M*, in the center of the CI, then decreases smoothly to either end of the CI, then drops further beyond the interval" (Cumming, 2011, p. 100). In Figure 2, M=50, SD=40, and n=13, and based on this information 90% CI for mean is [30.23, 69.77].

As it is seen in Figure 2, the shadow area presents the plausibility of the values in the interval for the population mean. Values close to our sample mean are more likely to be the true population mean. Also, it is important to note that the true population could be beyond the intervals, as we mentioned above.

4. LIMITATIONS OF NHSST

One of the ways to evaluate sample statistics is to use NHSST by using a linkage between the sample and the population by assuming the sample was drawn from that population. This procedure is called inferential statistics and is evaluated by using estimated probability (i.e., $p_{CALCULATED}$). In NHSST, the p value provides the probability of a sample statistic (e.g., mean, standard deviation, kurtosis, correlation

coefficient, effect size, etc.) supposing the sample is coming from the population that is defined by the null hypothesis and given a specified sample size (Thompson, 2006a). *p* value calculation is totally a mathematical procedure, so it does not say anything about practical importance or replicability (Thompson, 1996, 2006a).

Effect size and sample size are two factors that affect $p_{CALCULATED}$ value. That means, in the case of non-zero effect size, we eventually will get statistically significant results at some sample size (Thomson, 2006a). The only time that we cannot obtain statistically significant results is the case of zero effect size. If we have a zero effect size, $p_{CALCULATED}$ will always be 1. Zero effect size means that sample statistic does not diverge from the null hypothesis at all, so the probability of our sample statistics coming from the population, which is defined in the null hypothesis, is 1. Thompson (1999) explained the relation of *p* values between effect size and sample size as:

> Because *p* values are confounded indices, in theory 100 studies with varying sample sizes and 100 different effect sizes could each have the same single $p_{CALCULATED}$, and 100 studies with the same single effect size could each have 100 different values for $p_{CALCULATED}$. (p.169-170)

Because of the limitations on and criticism of NHSST, the APA Task Force on Statistical Inference (Wilkinson & the APA Task Force on Statistical Inference, 1999) discussed banning the use of NHSST but decided not to ban NHSST for the studies and instead indicated "Always provide some effect size estimate when reporting a *p* value" (p. 399).

4.1. An Alternative to NHSST

CIs include information on both location and the precision of your location. Also, we indicated previously that even though CIs can tell us whether our results are statistically significant or not, they do more than that. If we used CIs only as a tool to test whether our results are statistically significant or not, CIs would not be as useful as they are. Instead of using CIs to conclude whether our results are statistically significant or not, we should use them meta-analytically.

CIs provide estimation and meta-analytic thinking rather than dichotomous thinking (Cumming, 2011). Estimation and meta-analytic thinking give more valuable information than dichotomous thinking for populations because dichotomous thinking is basically the decision, based on the NHSST, for the null hypothesis. The only knowledge based on this decision is whether a single parameter is equal to a fixed value or not.

On the other hand, estimation thinking provides information about parameter estimates and their uncertainty based on the interval. (Cumming, 2011). Cumming (2011) explained estimation thinking as "Estimation thinking focuses on how much, by focusing on point and interval estimates" (p. 9). Also, Cumming (2011) gave the definition for meta-analytic thinking as "thinking that considers any results in the context of past and potential future results on the same question. It focuses on the cumulation of evidence over studies" (p. 9). As stated before, we can only find the true parameter by focusing all related previous studies by comparing CIs (Thompson, 1998,

2002, 2006a). Figure 3 illustrates the difference of dichotomous thinking and estimation thinking clearly.

Figure 3 illustrates a heuristic example in which two studies have evaluated the effectiveness of new curriculum for mathematics achievement. Study 1 used two independent groups each of size n=16 (Total=32), and Study 2 used two groups each with n=18 (Total=36). For each study, Figure 1 reports the difference and meta-analysis result between the means for the new and current curriculum, with the 95% CIs on that difference.

Researchers who advocate that CIs function no more effectively than NHSST in these studies may be classified as dichotomous thinkers. Thus, they will not find statistically significant results in both studies. In the Study 1: M(difference)=3.8, SD(difference)=7.92, SD(pooled)=5.6, and p=0.0645. In the Study 2: M(difference)=2.33, SD(difference) = 7.071, SD(pooled)=5, and p=0.1712. Based on these p values separately, both researchers fail to reject the null hypothesis at alpha (α) equals to 0.05. However, if researchers are estimating and are meta-analytic thinkers, they will clearly see that the new curriculum shows a statistically significant advantage over the current curriculum (p=0.0211, the null hypothesis of no difference is rejected).

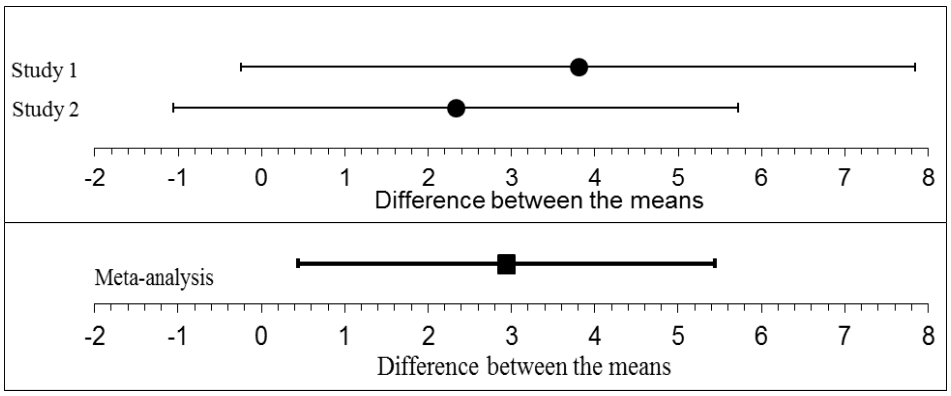

Figure 3. Difference between the means for mathematics achievement in the Study 1 and 2, with 95% confidence intervals; and difference between the means for mathematics achievement, with 95% confidence intervals, from a meta-analysis of two studies that compared a new curriculum with the current curriculum.

Also, in the meta-analytic study, the width of the CI is narrower; it gives us more precise information about point estimates. Also, as stated before, if a CI is drawn for a study and compared with CIs of previous studies, the true parameters will be finally found even if initial expectations are wildly wrong (Schmidt, 1996).

5. CIs FOR EFFECT SIZES

The APA Task Force on Statistical Inference (Wilkinson & the APA Task Force on Statistical Inference, 1999) strongly encouraged researchers to report effect size estimates with their p values. Also, CIs have been stated as "the best reporting strategy"

(APA, 2010, p.34). Thus, estimating CIs for effect sizes is unavoidable. However, there are some difficulties in estimating CIs for effect sizes.

Calculating CIs for commonly used statistics, such as mean, is very straightforward. We can use a specific formula to estimate CIs around the mean:

$$[M-t_v\, SD/\sqrt{n},\ M+ t_v\, SD/\sqrt{n}] \quad (1)$$

In the equation, t_v is the 95% critical value for t, and $v=n-1$ is degrees of freedom. The CI for mean is symmetric around the mean. Also, if our sample size is bigger than $n=30$, we can use z critical values instead of t critical values. In this situation, the formula will be:

$$[M-z_{critical}\, SD/\sqrt{n},\ M+z_{critical}\, SD/\sqrt{n}] \quad (2)$$

However, there is no formula to estimate CIs for effect sizes. Instead, we use an iteration that is a computer intensive statistical procedure. When we calculate CIs for mean, we use central t distribution, so both upper and lower intervals are equal. On the other hand, when we estimate CIs for non-zero effect sizes, the distribution is not central, so the upper and lower intervals are not equal. Moreover, iteration is needed to estimate lower and upper CIs separately. For detailed information about estimating CIs, readers can see Cumming and Finch (2001).

There are several computer programs that can iteratively estimate CIs, such as SPSS (Smithson, 2001), SAS (Algina & Keselman, 2003), R (Kelley, 2007), and EXCEL (Cumming, 2011; Cumming & Finch, 2001). In this paper, we used ESCI software to estimate CIs for Cohen's *d* standardized effect size. ESCI is also very user friendly for estimating CIs for effect sizes. We used Study 1 statistics, which was our initial example, and CIs were drawn for mean difference in Figure 3.

5.1. A Heuristic Example to draw CIs for Cohen's d by using ESCI

Study 1 had two independent groups, each of sample size 16 (n=32). As explained previously, a new curriculum was compared with the existing curriculum in terms of mathematics achievement. The mean difference (Mean of New Curriculum – Mean of Existing Group) was 3.8 with pooled SD=5.6. We did not provide the formula for pooling SDs, but it is commonly found in many statistic books. The mean difference and pooled SD are enough to estimate Cohen's d standardized effect size estimate.

$$d= M \text{ (difference)} / SD \text{ (Pooled)} \quad (3)$$

Based on the third equation (3), Cohen's d will be (3.8/5.6) 0.68. Now we can use ESCI to estimate the CI for our effect size point estimate. What we need in ESCI software is the estimated Cohen's d and sample sizes for two groups. Figure 4 shows 95% CI for the effect size d. 95% CI is estimated as [-0.039, 1.388].

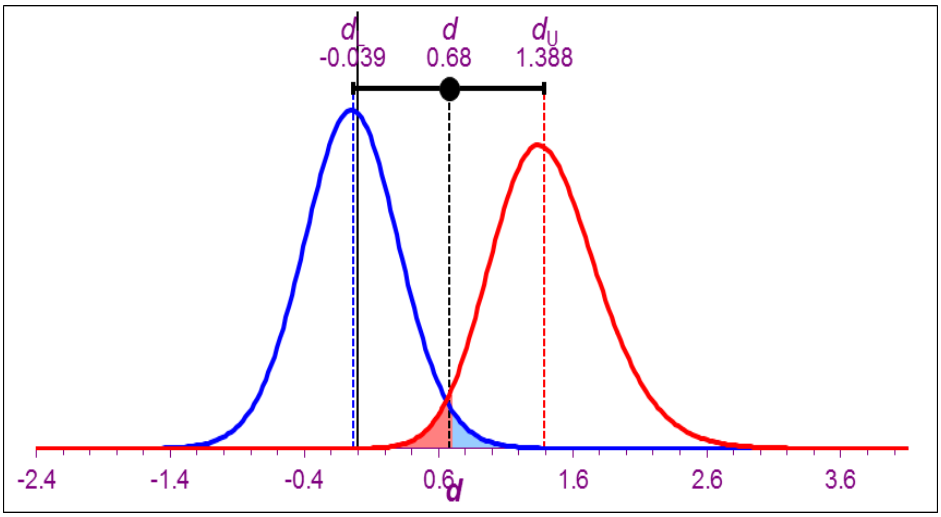

Figure 4. Lower and Upper CIs for Cohen's *d* effect size.

6. IMPLICATIONS FOR TEACHING

In statistics education, students may become confused when trying to learn some concepts such as CIs. Teaching these concepts with mathematical formulas and verbal instruction may not be efficient. In these kinds of situations, teaching with more visual techniques such as graphical representations would be quite effective. In this paper, we summarized the concept of CIs and common misconceptions about them. Then, we used visual representation for CIs to make them more concrete for learners. We utilized Cumming's (2011) book and his software called ESCI.

We discussed limitations of NHSST by providing an alternative to NHSST. We definitely teach students what NHSST does, and what information it provides us about our population. Students and also researchers should understand that we would like to be as precise as possible about our population estimates by using our sample statistics. We do not say that NHSST does not provide any information about our population estimates, but NHSST provides less information than CIs. We can only increase our information about population estimates by using CIs meta analytically.

We also provided information about CIs for effect sizes, specifically Cohen's d. In theory, estimating CIs for any effect size is not an easy procedure because there are no mathematical CI formulas for any effect sizes. Instead, there are computer intensive methods to estimate CIs for effect sizes. Even though the CIs' estimation for effect sizes can be done by using a statistical software, such as SPSS, we demonstrated an easier program (ESCI) to estimate CIs for Cohen's d effect size. Further information about CIs and practice with them in ESCI can be found in Cumming's (2011) book.

REFERENCES

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Algina, J., Keselman, H. J. (2003). Approximate confidence intervals for effect sizes. *Educational and Psychological Measurement, 63*, 537-553. doi: 10.1177/0013164403256358

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10*, 389-396. doi: 10.1037/1082-989X.10.4.389

Byrd, J. K. (2007). A call for statistical reform in EAQ. *Educational Administration Quarterly, 43*, 381-391. doi: 10.1177/0013161X06297137

Cumming, G. (2011). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Cumming, G., & Fidler, F. (2009). Confidence intervals: Better answers to better questions. *Journal of Psychology, 217*, 15-26. doi:10.1027/0044-3409.217.1.15

Cumming, G., & Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*, 532-575. doi: 10.1177/0013164401614002

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals, and how to read pictures of data. *American Psychologist, 60*, 170-180. doi: 10.1037/0003-066X.60.2.170

Cumming, G., Williams, J. & Fidler, F. (2004). Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding statistics, 3*, 299-311. doi: 10.1207/s15328031us0304_5

Di Stefano, J. (2004). A confidence interval approach to data analysis. *Forest Ecology and Management, 187*, 173–183. doi:10.1016/S0378-1127(03)00331-1

Fidler, F., & Loftus, G. (2009). Why figures with error bars should replace *p* values: Some conceptual arguments and empirical demonstrations. *Journal of Psychology, 217*, 27-37. doi: 10.1027/0044-3409.217.1.27

Fidler, F., & Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed- and random-effects effect sizes. *Educational and Psychological Measurement, 61*, 575–604. doi: 10.1177/0013164401614003

Finch, S., Cumming, G., & Thomason, N. (2001). Reporting of statistical inference in the *Journal of Applied Psychology*: Little evidence of reform. *Educational and Psychological Measurement, 61*, 181-210. doi: 10.1177/00131640121971167

Good, P. I., & Hardin, J. W. (2003). *Common errors in statistics (and how to avoid them)*. New York: Wiley.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52*, 15-24. doi: 10.1037/0003-066X.52.1.15

Howell, D. C. (2008). *Fundamental statistics for the behavioral sciences* (6th ed.). Belmond, California: Thomson.

Keller, G., & Warrack, B. (2003). *Statistics for management and economics* (6th ed.). Pacific Grove, CA: Thomson Learning.

Kelley, K. (2007). Confidence intervals for standardizes effect sizes: Theory, application, and implementation. *Journal of Statistical Software, 20*, 1-24.

Kline, R. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Knapp, T. R. & Sawilowsky, S. S. (2001). Constructive criticisms of methodological and editorial practices. *The Journal of Experimental Education, 70,* 65-79. doi: 10.1080/00220970109599498

Lomax, R. G. (2001). *An introduction to statistical concepts for education and behavioral sciences.* Mahwah, NJ: Lawrence Erlbaum.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists.* New York: Springer.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioral sciences*. New York: Wiley.

Ott, R. L., & Longnecker, M. (2010). *An introduction to statistical methods and data analysis* (6th ed.). Belmond, CA: Brooks/Cole.

Pagano, R. R. (2009). *Understanding statistics in the behavioral sciences.* (9th ed.). Belmont, CA: Wadsworth.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods, 1,* 115-129. doi: 2048/10.1037/1082-989X.1.2.115

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and paramaters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement, 61,* 605-632. doi: 10.1177/00131640121971392

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25*(2), 26-30. doi: 10.3102/0013189X025002026

Thompson, B. (1998). In praise of brilliance: Where that praise really belongs. *American Psychologist, 53,* 799-800. doi: 2048/10.1037/0003-066X.53.7.799

Thompson, B. (1999). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory & Psychology, 9,* 167-183. doi: 10.1177/095935439992006

Thompson, B. (2001). Significance, effect sizes, stepwise methods, and other issues: Strong arguments move the field. *Journal of Experimental Education, 70*, 80-93. doi: 10.1080/00220970109599499

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher, 31*(3), 24-31. doi: 10.3102/0013189X031003025

Thompson, B. (2006a). *Foundations of behavioral statistics: An insight-based approach.* New York: Guilford.

Thompson, B. (2006b). Research synthesis: Effect sizes. In J. Green, G. Camilli, & P.B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 583-603). Washington, DC: American Educational Research Association.

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in Schools, 44,* 423-432. doi: 10.1002/pits.20234

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594-604. doi: 10.1037/0003-066X.54.8.594