Chen, L. T., & Liu, L. (2019). Content analysis of statistical power in educational technology research: Sample size matters. *International Journal of Technology in Teaching and Learning, 15(1)*, 49 -75.

Content Analysis of Statistical Power in Educational Technology Research: Sample Size Matters

Li-Ting Chen & Leping Liu
University of Nevada, Reno

In educational technology research, most studies are conducted to explore the effectiveness of using technology to improve teaching and learning. Priori power analysis enables researchers to determine sufficient sample size for achieving adequate statistical power during research planning. Observed power analysis is carried out on completed studies to estimate statistical power. While priori power analysis is recommended for sample size estimation, observed power analysis has been criticized for being incorrect and misleading. To understand current practices of power analysis in the field, we conducted a content analysis on five years' publications in *Educational Technology Research and Development* from 2014 to 2018, a total of 178 articles. Our findings showed that only two articles (1.1%) reported a priori power analysis and seven articles (4.0%) reported observed power although it is not recommended. To facilitate sample size determination during research planning, we generated sample size tables for various *t* tests and ANOVAs from G*Power. Best practice recommendations to conduct and report educational technology research are provided.

Keywords: research planning, ANOVA, statistical power, effect size, sample size, educational technology research

INTRODUCTION

Statistical power is defined as the probability of rejecting a false null hypothesis. Therefore, the value of statistical power ranges from 0 to 1. Why should researchers care about statistical power? Why should researchers perform power analysis to plan sample size? Statistical power depends on three parameters: significance level (α level), effect size, and sample size. Given an effect size value and a fixed α level, recruiting more participants in a study increases statistical power and the accuracy of the result. In his book of *Statistical Power Analysis for the Behavioral Sciences*, Cohen (1988) wrote "Since statistical significance is so earnestly sought and devoutly wished for by behavioral scientists, one would think that the *a priori* probability of its accomplishment would be routinely determined and well understood. Quite surprisingly, this is not the case" (p. 1).

Li-Ting Chen is Assistant Professor of Educational Measurement and Statistics, Leping Liu is Professor of Information Technology and Statistics. Both are at Counseling and Educational Psychology, College of Education, University of Nevada, Reno. Li-Ting Chen can be reached at litingc@unr.edu

Priori and observed power analyses are two types of power analyses that have been identified in published articles (Peng, Long, & Abaci, 2012). While a priori power analysis is used to estimate the minimum sample size required for a given power, population effect size, and α level, an observed power analysis is carried out to estimate the power, given the sample size, sample effect size, and α level (Peng et al., 2012). Yuan and Maxwell (2005) conducted a Monte Carlo study to examine what information can be provided by observed power. Yuan and Maxwell (2005) concluded "the observed power is almost always a biased estimator of the true power" (p. 162). Likewise, Cumming and Calin-Jageman (2017) wrote "post hoc [observed] power is useless. Avoid post hoc [observed] power–simply never use it" (p. 284). Indeed, a priori power analysis is the power analysis that is recommend by the sixth edition of *Publication Manual of the American Psychological Association* (2010) and the newly released journal article reporting standards in the APA Publications and Communications Board Task Force Report (Appelbaum et al., 2018). In literature, priori power analysis, planned power analysis, and prospective power analysis have been used interchangeably (Peng et al., 2012).

Peng et al. (2012) analyzed 1,357 articles published in 12 education related journals from 2005 to 2010 for examining power analysis conducted by researchers. Their findings revealed that priori power analysis was conducted in 24 articles (2%). Observed power was reported in 47 articles (3%), although it is not recommended (Cumming & Calin-Jageman, 2017; Hoenig & Heisey, 2001; Levine & Ensom, 2001; Yuan & Maxwell, 2005). In addition, there were 192 articles (14%) in which authors mentioned power and sample size but did not actually compute or estimate power.

In the field of educational technology research, most studies have been conducted to examine the effectiveness of using or integrating information technology to improve teaching and learning (Hwang, Lai, & Wang, 2015; Levin & Wadmany, 2008; Liu & Chen, 2018). To preserve the statistical validity of data analysis, researchers are recommended to conduct a priori power analysis to determine the minimum sample size that is sufficient for their studies (Aberson, 2019a, 2019b; Dwork et al., 2015; Liu & Maddux, 2008). However, little is known about whether educational technology researchers estimate sample size during their research planning. Therefore, we conducted this study to (a) review and summarize current practices of power analysis in the field of educational technology research from a content analysis on five years' publications by one of the leading journals in this field, and (b) provide sample sizes required for popular statistical tests used in the field to facilitate research planning.

LITERATURE REVIEW

When a researcher makes a statistical conclusion for the null hypothesis, there are four possibilities (see Table 1). Given the null hypothesis (H_0) is true, the researcher may correctly fail to reject the true null hypothesis or incorrectly reject the true null hypothesis. The probability of correctly failing to reject the true null hypothesis is $1 - \alpha$, whereas the probability of incorrectly rejecting the true null hypothesis is α (also called Type I error rate or false positive rate). Typically, a researcher uses $\alpha = .05$ for statistical significance. When $\alpha = .05$ is used and the null hypothesis is true, there is one chance out of 20 that the true null hypothesis will be falsely rejected. Given the null hypothesis (H_0) is false, the researcher may incorrectly fail to reject the false null hypothesis or correctly reject the false null hypothesis. The probability of incorrectly failing to reject the false null hypothesis is β (also called Type II error or false negative rate), whereas the probability of correctly rejecting the false null hypothesis is $1-\beta$ (also called statistical power). Given the statistical

power is .80, there is one chance out of five that a researcher may falsely fail to reject the false null hypothesis. Cohen (1988) suggested that when there is no other basis for setting the value of desired statistical power, .80 can be used. The rationale is that typically, Type I errors (.05) are four times as serious as Type II errors (.20).

Table 1. *The two by two table to illustrate the Type I error, Type II error, and statistical power*

		True Situation	
		H_0 True (Game-based curriculum does not improve learning)	H_0 False → H_1 True (Game-based curriculum improves learning)
Researcher's Decision	Fail to reject H_0	Correct decision Probability = $1 - \alpha$	Type II error Probability = β
	Reject H_0	Type I error Probability = α	Correct decision Probability = $1 - \beta$ = Statistical power
	Total Probability	1.00	1.00

Suppose that an educational technology researcher designs a game-based curriculum for learning fourth grade math, the new curriculum may in fact yield the same math performance for children who learn based on this new curriculum and who learn from the traditional curriculum. Without knowing the effectiveness of the new curriculum, the researcher may recruit a group of fourth graders. Half of the fourth graders are assigned to receive the game-based curriculum and half of the fourth graders are assigned to receive the traditional curriculum. After completion the curriculums, the researcher can test the null hypothesis of equivalent math performance for these two groups. Given the data, the researcher may fail to reject the true null hypothesis and therefore make a correct decision. On the other hand, the researcher may incorrectly reject the true null hypothesis and thus make a Type I error. With a different scenario in which the new curriculum in fact improves math performance of fourth graders, the researcher may fail to reject the false null hypothesis and thus make a Type II error. On the other hand, the researcher may reject the null hypothesis and make a correct decision.

FACTORS ASSOCIATED WITH STATISTICAL POWER

We have defined Type I error (α), Type II error (β), and statistical power ($1-\beta$). Below we use graphs to illustrate the relationships among the statistical power, α, effect size, and sample size. Figure 1 shows the H_0 distribution and the H_1 distribution. Using the game-based example above, the H_0 distribution represents the sampling distribution of the mean difference when the null hypothesis is true. The null hypothesis is that mean math performance for children who receive game-based curriculum is equal to or lower than children who receive traditional curriculum or simply $\mu_{game\_based} \leq \mu_{traditional}$. The H_1 distribution represents the sampling distribution of the mean difference when the null hypothesis is false (or $\mu_{game\_based} > \mu_{traditional}$). As shown in Figure 1 and Figure 2, when α increases and other things remain the same, it simultaneously decreases β and increases statistical power. The effect size can be expressed as the separation between H_0 distribution and H_1 distribution. When the effect size increases, it increases the distance between H_0 distribution and H_1 distribution. For the game-based example, larger effect size means

larger mean difference between the two groups. As shown in Figure 1 and Figure 3, when the effect size increases and other things remain the same, the area of 1– β increases and hence increases the statistical power. Lastly, the variance of the sampling distribution of the mean difference decreases as the sample size increases. It is because the variance of the sampling distribution of the mean difference is defined as Equation 1:

$$\sigma^2_{\bar{X}_{\text{game\_based}}-\bar{X}_{\text{traditional}}} = \frac{\sigma^2_{\text{game\_based}}}{n_{\text{game\_based}}} + \frac{\sigma^2_{\text{traditional}}}{n_{\text{traditional}}} \quad (1)$$

When $\sigma^2_{\bar{X}_{\text{game\_based}}-\bar{X}_{\text{traditional}}}$ decreases with other things being equal, the overlap between H_0 distribution and H_1 distribution reduces and hence increases the statistical power (see Figure 1 and Figure 4). In sum, statistical power increases as α increases, effect size increases, and sample size increases.

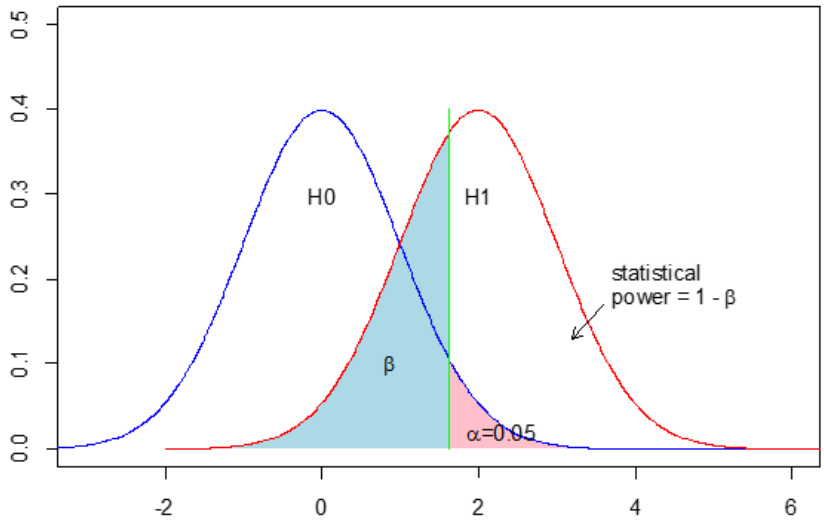

Figure 1. The probability of making Type I error, Type II error, and statistical power

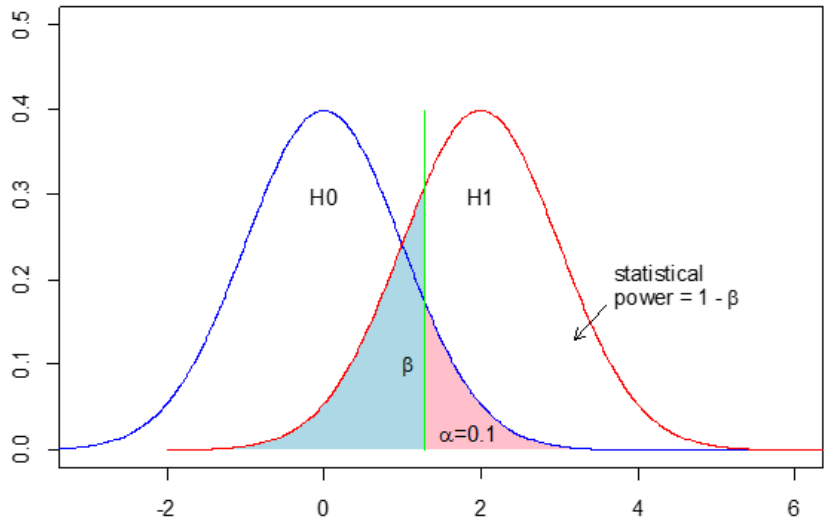

Figure 2. The probability of making Type I error, Type II error, and statistical power, when everything remains the same but α increases

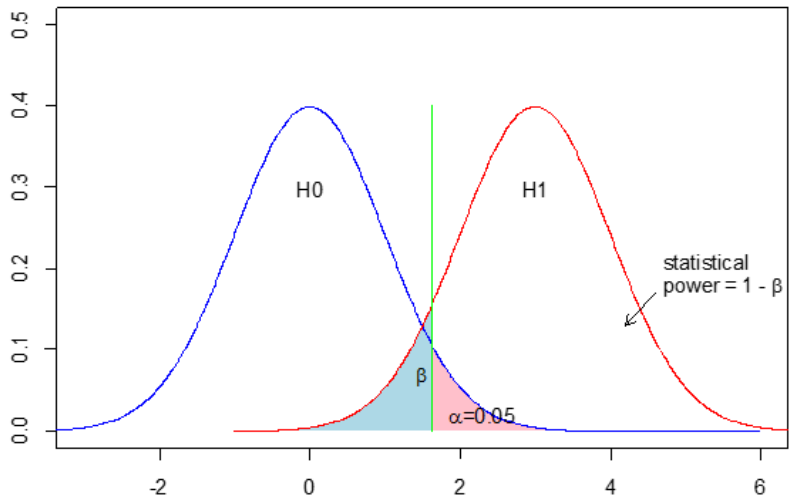

Figure 3. The probability of making Type I error, Type II error, and statistical power, when everything remains the same but effect size increases

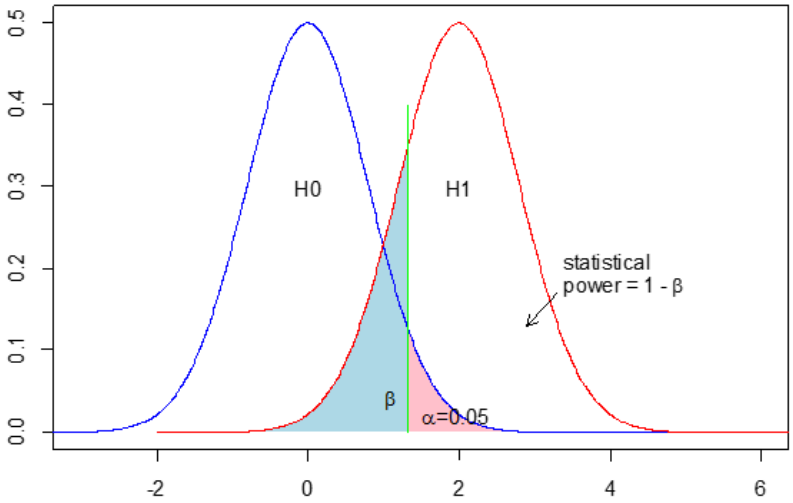

Figure 4. The probability of making Type I error, Type II error, and statistical power, when everything remains the same but sample size increases

Because the relationships among statistical power, α, effect size, and sample size, when three are known or estimated, the fourth parameter can be calculated. When statistical power is calculated on completed studies, this is called an observed power analysis. When required sample size is estimated based on a desired statistical power, an α, and a certain effect size, this is called a priori power analysis. In some circumstances, researchers may find it is useful to calculate the detectable effect size for the desired power, α, and available number of participants (Cohen, 1988). This is called to determine the sensitivity of studies (Murphy & Myors, 1998). Sensitivity analyses can be conducted before executing the study to enhance researchers' understanding of what size of effect could be reasonably detected given a particular sample size, a desired power, and α (Faul, Erdfelder, Lang, & Buchner, 2007; Murphy & Myors, 1998; Perugini, Gallucci, & Costantini, 2018). If the results show that the study can detect only a large effect but a small effect is expected to occur, the

researcher may postpone the study until resources needed to achieve the desired power are available (Murphy & Myors, 1998).

Low statistical power means that the probability to reject a false null hypothesis is low. To avoid a failure to reject a false null hypothesis due to low statistical power, researchers may conduct a priori power analysis for estimating the required sample size. Researchers can also use alternative methods to increase statistical power. For instance, the appropriate use of covariates increases statistical power. In the game-based example, the researcher may use scores from IQ tests as a covariate. Because effect size is also related to statistical power, researchers may use a stronger treatment/intervention/manipulation and standardized procedures to increase statistical power (Perugini et al., 2018). A six week game-based curriculum with one hour a day, five days a week is a stronger treatment than a one week game-based curriculum with the same length of learning per day. In addition, research has shown that the reliability of the dependent and covariate measures have an effect on statistical power (De Schryver, Hughes, Rosseel, & Houwer, 2016; Kanyongo, Brook, Kyei-Blankson, & Gocmen, 2007; Shadish, Cook, & Campbell, 2002). Shadish et al. (2002) summarizes different ways to increase statistical power (see Table 2.3 of their book).

ESTIMATING SAMPLE SIZE FOR PRECISION

Using a priori power analysis to determine minimum sample size based on statistical significance is just one way for sample size planning. When a researcher's goal is to achieve a certain degree of accuracy for estimating the parameter, sample size can be planed for precision. Regarding the previous example of game-based curriculum, the parameter is the mean difference on math performance between children who learn based on the game-based curriculum and on the traditional curriculum. Given the collected data, a range of values can be estimated for the parameter. The width of the range of values refers to the precision of estimation. The precision increases as the width decreases. Yet higher precision requires larger sample size. This alternative approach for sample size planning is called accuracy in parameter estimation (Cumming & Calin-Jageman, 2017; Kelley & Rausch, 2006). The R package "MBESS" (Kelley, 2019) was developed for researchers aiming to plan sample size for achieving an acceptable level of accuracy in estimating the parameter.

SAMPLE SIZE ESTIMATION AND COMPUTING TOOLS FOR POWER ANALYSIS

Scholars have constructed required sample size tables for certain statistical tests in published articles (e.g., Cohen, 1992; Kanyongo et al., 2007) and textbooks (e.g., Kirk, 2008, 2013). These tables are usually constructed for standard statistical tests (e.g., *t* tests) with the effect size benchmarks of small, medium, and large (Cohen, 1988) under two or three levels of α (e.g., .01, .05). Due to the limited conditions presented in the tables, these tables may not be readily applicable for educational technology researchers. Kirk (2013) presented the degrees of freedom curves as a function of statistical power and effect size for ANOVAs, which may still require researchers to approximate the total sample size required.

Multiple computing tools are available for conducting a priori power analysis, such as SAS/STAT, R stats package, and PASS Power Analysis and Sample Size Software. Based on their review of eight programs/packages for conducting power analysis, Peng et al. (2012) recommended two stand-alone/specialized programs: SPSS/SamplePower and G*Power. However, SPSS/SamplePower is no longer available. G*Power can be run on Windows and Mac and it is free downloadable software. Practical guidelines for using G*Power can be found in Faul et al. (2007), Faul, Erdfelder, Buchner, and Lang (2009)

and Perugini et al. (2018). Several free computing tools were recently developed for power analysis, such as R packages "pwr" (Champely et al., 2018) and "pwr2ppl" (Aberson, 2019a) and PS: Power and Sample Size Calculation (Dupont & Plummer, 2018).

PURPOSES AND RESEARCH QUESTIONS OF THE RESEARCH

Again, little is known about the application of power analysis in educational technology research. If power analysis is conducted by educational technology researchers, do they conduct a priori power analysis as recommended in literature? Although sample size tables are presented in some articles and textbooks, these sample size tables are either limited to certain conditions (e.g. for small, medium, and large effects based on Cohen, 1988) or required users to approximate total sample sizes (e.g., Kirk, 2013). To understand current practices in conducting power analysis and to facilitate sample size estimation for achieving acceptable statistical power, we have two aims of this paper: (a) to review power analyses conducted by educational technology researchers, and (b) to provide sample size tables for popular statistical tests in the field of educational technology. To fulfill the first aim, articles that were published in *Educational Technology Research and Development* (ETRD) over the five-year period from 2014 to 2018 were analyzed. To fulfill the second aim, sample size tables for popular statistical methods were generated using G*Power 3.1.9.4. Baydas, Kucuk, Yilmaz, Aydemir, and Goktas (2015) reviewed articles published in ETRD and *British Journal of Educational Technology* from 2002 to 2014. They identified ANOVA/ANCOVA and *t* tests were the top two statistical techniques used by educational technology researchers (Baydas et al., 2015). Hence, we constructed multiple sample size tables for ANOVAs and *t* tests. We also illustrated how to estimate sample size required in ANCOVAs. Specifically, we were interested in answering the following research questions:

1. What types of power analyses were conducted by educational technology researchers?
2. How did educational technology researchers rationalize sample size used in their studies?
3. What were the tools used by educational technology researchers to conduct power analyses?
4. What were the minimum sample sizes required for popular statistical tests in the field of educational technology? Statistical tests considered in this paper include:
    a. independent-sample *t* tests with equal or unequal sample sizes;
    b. dependent-sample *t* tests;
    c. one-way between-subjects ANOVAs with three to five groups;
    d. interaction effects of 2 × 2, 2 × 3, and 3 × 3 two-way between-subjects ANOVAs; and
    e. interaction effects of 2 × 2 mixed ANOVAs.
5. Did educational technology researchers recruit sufficient number of participants for their studies?

METHOD

SAMPLE

One hundred seventy-eight articles published in the journal titled, *Educational Technology Research and Development* (ETRD), from 2014 to 2018 were the study samples. Articles published in the most recent five years were chosen because research has suggested a stable trend in usage of statistical techniques within five years (Goodwin & Goodwin, 1985). ETRD is a bi-monthly publication of the Association for Educational

Communications & Technology. On the journal's homepage, it indicates ETRD is "the only scholarly journal for the field focusing entirely on research and development in educational technology." According to Journal Citation Reports, the 2018 impact factor for ETRD was 2.115. ETRD was placed ninth in Google Scholar ranking of top publications in educational technology (Google Scholar, 2019). Articles published in ETRD has been reviewed in other studies (Baydas et al., 2015; Hsu, Hung, & Ching, 2013; Reeves & Oh, 2017; Shih, Feng, & Tsai, 2008). In the content analysis, we excluded qualitative research articles, quantitative articles that only reported descriptive statistics, theoretical articles, narrative review articles, reflections, introduction to special issues, and errata.

CODING AND ANALYSIS

We coded each article in terms of whether power analysis was reported by the authors. When power analysis was reported, we then coded the type of power analysis conducted by the authors. Tools for conducting power analysis was also coded. When power analysis was not reported, we examined the article in terms of two aspects: (a) if the authors provided references to support sample size used, and (b) if the authors discussed or mentioned insufficient power.

All electronic copies of the articles published in ETRD from 2014 to 2018 were first downloaded to a folder by a graduate assistant. The authors then developed the coding sheet based on Peng et al. (2012), and all the articles were coded.

Descriptive statistics were reported that demonstrated the current practices in conducting power analysis.

*USING G*POWER TO GENERATE SAMPLE SIZE TABLES*

In this section we present the steps of using G*Power 3.1.9.4 to generate the sample sizes required for popular statistical tests in the field of educational technology. Five sample size tables were constructed with these steps (See Tables 2, 3, 4, 5, and 6 in the Appendix section). For all the tables, we used $\alpha = .05$. Researchers who wish to adopt a different level of α can follow the steps we provide below but change the value of α accordingly.

Independent-Sample t Test. When estimating the sample size required for *t* tests, we used two-tailed tests only (e.g., $\mu_{game\_based} = \mu_{traditional}$). Researchers who wish to use a one-tailed test (e.g., $\mu_{game\_based} \leq \mu_{traditional}$) can follow the steps we provide below but change the Tail(s) for the test from "Two" to "One".

Figure 5 (see next page) illustrates the six steps on generating the sample size table for independent-sample *t* tests. Step 1: Selected "t tests" from the "Test family" drop-down menu. Step 2: Selected "Means: Difference between two independent means (two groups)" from the "Statistical test" drop-down menu. Step 3: Selected "A priori: Compute required sample size–given α, power, and effect size" from the "Type of power analysis" drop-down menu. Step 4: Selected "two" from the "Input Parameters: Tail(s)" drop-down menu. Step 5: Entered the desired Effect size d (= Cohen's *d*), α err prob (=.05), Power (1–β err prob), and Allocation ratio N2/N1 (= n_2/n_1). Cohen's *d* is the mean difference expressed in the unit of standard deviation. We varied Effect size d from 0.2 to 1.0 in steps of 0.1, Power as .80, .90, and .95, and Allocation ratio N2/N1 from 1 to 2 in steps of 0.5. Step 6: Clicked the button "Calculate."

Dependent-Sample t Test. We used six steps for generating the sample size table for dependent-sample *t* tests. Step 1: Selected "t tests" from the "Test family" drop-down menu. Step 2: Selected "Means: Difference between two dependent means (matched pairs)" from the "Statistical test" drop-down menu. Step 3: Selected "A priori: Compute required sample size–given α, power, and effect size" from the "Type of power analysis"

drop-down menu. Step 4: Selected "two" from the "Input Parameters: Tail(s)" drop-down menu. Step 5: Entered the Effect size dz, α err prob (=.05), and Power (1–β err prob). Effect size dz is computed as:

$$d_z = \frac{\text{Cohen's } d}{\sqrt{2\times(1-\rho)}} \quad (2)$$

where Cohen's *d* is the effect size if independent samples are used, and ρ is the expected correlation between two measures. We varied Cohen's *d* (Effect size d in G*Power) from 0.2 to 1.0 in steps of 0.1 and ρ from .6 to .8 in steps of .1. Therefore, the entered Effect size dz varied from 0.224 to 1.581. We also used three different values for Power: .80, .90, and .95. Step 6: Clicked "Calculate".

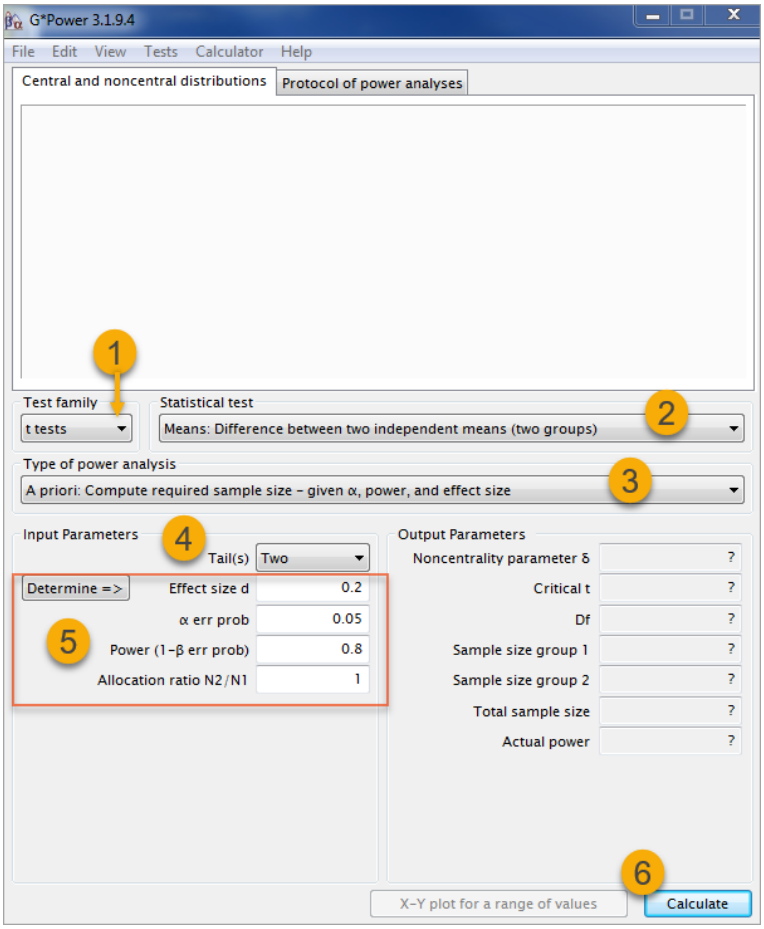

Figure 5. Using G*Power 3.1.9.4 to generate sample size table for independent-sample *t* tests with equal or unequal sample sizes

One-Way Between-Subjects ANOVA. There were nine steps on generating the sample size table for one-way between-subjects ANOVA (Figure 6, see next page). Step 1: Selected "F tests" from the "Test family" drop-down menu. Step 2: Selected "ANOVA: Fixed effects, omnibus, one-way" from the "Statistical test" drop-down menu. Step 3: Selected "A priori: Compute required sample size–given α, power, and effect size" from the "Type of power analysis" drop-down menu. Step 4: Clicked "Determine =>" from Input Parameters. Step 5: Selected "Effect size from variance" from the "Select procedure" drop-down menu. Step 6: Selected "Direct" and typed the estimated population value of Partial η^2. We varied Partial η^2 from .01 to .22 in steps of .01. Partial η^2 in G*Power for one-way ANOVAs is actually η^2 (Perugini et al., 2018). This is because in one-way ANOVAs, there is only the effect of one independent variable. Hence, there is no other effect to partial out.

Partial η^2 is used in factorial designs (e.g., 2×2 ANOVAs). η^2 can be explained as the percentage of variance on the dependent measure that can be explained by the levels of independent variable. When effect size is defined as f, it can be computed from η^2:

$$f = \sqrt{\frac{\eta^2}{1-\eta^2}} \quad (3)$$

Similar to Cohen's *d*, when effect size in ANOVAs is computed from f, it is defined as the difference among means being expressed in units of the within-groups population standard deviation (Kirk, 2013). In G*Power, researchers can either enter the effect size expressed as η^2 and then ask the program to convert η^2 to f or enter f directly. We used the first approach. Step 7: Clicked "Calculate and transfer to main window". Step 8: Entered the desired α err prob (=.05), Power (1–β err prob), and Number of groups. We varied Power as .80, .90, and .95 and Number of groups as 3, 4, and 5. Step 9: Clicked the button "Calculate."

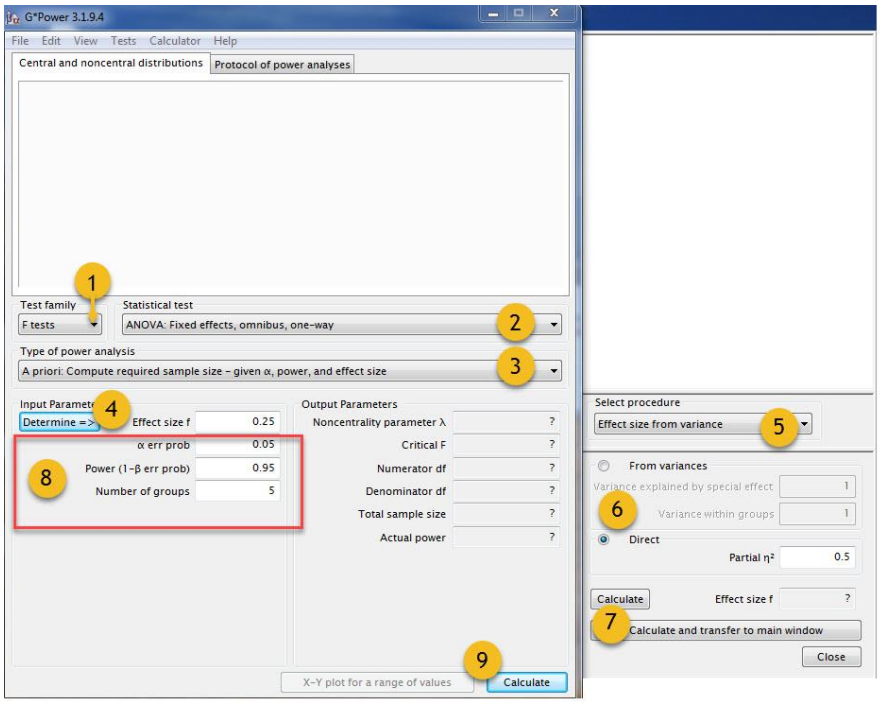

Figure 6. Using G*Power 3.1.9.4 to generate sample size table for one-way between-subjects ANOVAs

Interaction Effects in Two-Way Between-Subjects ANOVA. Eight steps involved in generating the sample size table for testing the interaction effect of 2 × 2, 2 × 3, and 3 × 3 two-way between-subjects ANOVAs. Step 1: Selected "F tests" from the "Test family" drop-down menu. Step 2: Selected "ANOVA: Fixed effects, special, main effects and interactions" from the "Statistical test" drop-down menu. Step 3: Selected "A priori: Compute required sample size–given α, power, and effect size" from the "Type of power analysis" drop-down menu. Step 4: Clicked "Determine =>" from Input Parameters. Step 5: Selected "Direct" and typed the estimated population value of Partial η^2. We varied Partial η^2 from .01 to .22 in steps of .01. The equation of Partial η^2 for the interaction effect is

$$\text{Partial } \eta^2 = \frac{\sigma^2_{\text{interaction}}}{\sigma^2_{\text{interaction}} + \sigma^2} \quad (4)$$

where $\sigma^2_{\text{interaction}}$ is the variance explained by the interaction effect and σ^2 is the population residual variance. As shown in Equation 4, the two main effects are partialed

out. Similar to the effect size for one-way ANOVAs, in G*Power, researchers can either enter the effect size expressed as Partial η^2 and then ask the program to convert Partial η^2 to f or enter f directly. We used the first approach. Equation 3 can be used to convert Partial η^2 to f, with η^2 replaced by Partial η^2. Step 6: Clicked "Calculate and transfer to main window". Step 7: Entered the desired α err prob (=.05), Power (1–β err prob), Numerator df, and Number of groups. For 2 × 2 between-subjects ANOVAs, we typed 1 for "Numerator df" and 4 for "Number of groups." The 1 for "Numerator df" is because of the degrees of freedom for testing the interaction effect = (2 − 1) × (2 − 1) = 1. For 2 × 3 between-subjects ANOVAs, we typed 2 for "Numerator df" and 6 for "Number of groups." The 2 for "Numerator df" is because of the degrees of freedom for testing the interaction effect = (2 − 1) × (3 − 1) = 2. For 3 × 3 between-subjects ANOVAs, we typed 4 for "Numerator df" and 9 for "Number of groups." The 4 for "Numerator df" is because of the degrees of freedom for testing the interaction effect = (3 − 1) × (3 − 1) = 4. We used Power as .80, .90, and .95. Step 8: Clicked the button "Calculate."

Interaction Effects of 2 × 2 Mixed ANOVAs. Eight steps involved in generating the sample size table for the interaction effects of 2 × 2 mixed ANOVAs. Step 1: Selected "F tests" from the "Test family" drop-down menu. Step 2: Selected "ANOVA: Repeated measures, within-between interaction" from the "Statistical test" drop-down menu. Step 3: Selected "A priori: Compute required sample size–given α, power, and effect size" from the "Type of power analysis" drop-down menu. Step 4: Clicked "Determine =>" from Input Parameters. Step 5: Selected "Direct" and typed the estimated population value of Partial η^2. We varied Partial η^2 from .01 to .14 in steps of .01. Step 6: Clicked "Calculate and transfer to main window". Step 7: Entered the desired α err prob (=.05), Power (1–β err prob), Number of groups (=2), Number of measurement (=2), Corr among rep measures, and Nonsphericity correction ε (=1), where *Corr among rep measures* stands for the "correlation among repeated measures." In a 2 × 2 mixed ANOVA, two repeated measures are used. We varied Corr among rep measures from .6 to .8 in steps of .1. It should be noted that when there are only two repeated measures, sphericity is not a concern. Therefore, we entered 1 for the nonsphericity correction. When there are three or more repeated measure and the sphericity assumption is likely to be violated, a correction value should be entered to adjusting the degrees of freedom of the *F* distribution accordingly (Faul et al., 2007). For constructing the sample size table for interaction effects of 2 × 2 Mixed ANOVAs, we used Power of .80, .90, and .95. Step 8: Clicked the button "Calculate."

RESULTS

This section includes the descriptive analysis results. Firstly, we present the current practices in conducting power analysis. Secondly, we summarize the minimum sample sizes, for five popular statistical tests employed in the field of educational technology research, that were generated with G* Power. Finally, we examine the sample size reported in eight ETRD articles on the theme of game-based learning, regarding whether the authors recruited to the sufficient sample sizes for their studies. We organize and present the results by answering each of the five research questions.

CURRENT PRACTICES

Research Question 1. What types of power analyses were conducted by educational technology researchers? Among the 178 reviewed articles, two articles (1.1%) reported that a priori power analysis was conducted to estimate the required sample size. Seven articles (3.9%) reported observed power. Among these seven articles, one article also reported detectable effect size given the sample size. Two articles (1.1%) reported

detectable effect size only. We did not find any article that reported planning sample size for the precision of parameter estimate. Therefore, the majority of educational technology researchers did not conduct priori power analysis to estimate sample size during research planning. Furthermore, when power analysis was conducted, researchers were more likely to conduct an observed power analysis for estimating statistical power than a priori power analysis for estimating the required sample size.

Research Question 2. How did educational technology researchers rationalize sample size used in their studies? In addition to the 11 articles above, authors of seven articles (3.9%) cited references to support sufficient sample size was used. In another 13 articles (7.3%), the authors did not report results from power analysis but mentioned insufficient power. There were a total of 147 (82.6%) articles that the authors provided neither the rationale for sample size used in their studies nor the estimation of statistical power.

Research Question 3. What were the tools used by educational technology researchers to conduct power analyses? For the two articles that used a priori power analysis to determine sample size, one used R package 'pwr' (Champely et al., 2018) and the other used the Optimal Design Software for Multi-Level and Longitudinal Research (Raudenbush et al., 2011). For the seven articles that reported observed power, two used G*Power, two used SPSS, and three did not report the tools used to compute observed power. For the two articles that calculated detectable effect size, one used G*Power and the other did not report the computing tool.

MINIMUM SAMPLE SIZE REQUIRED FOR POPULAR STATISTICAL TECHNIQUES

Research Question 4. What were the sample sizes required for popular statistical tests in the field of educational technology? The following are the minimum sample sizes for five popular statistics tests used in educational technology research.

Table 2 presents *sample sizes required for independent-sample t tests*. To use this table, researchers need to decide (1) the population effect size expressed as Cohen's *d*, (2) the ratio of two sample sizes (n_2/n_1; when $n_1 = n_2$, $n_2/n_1 = 1$), and (3) the desired statistical power. Suppose that a researcher expects the mean math performance of children who receive the game-based curriculum is 0.5 standard deviation higher than those who receive the traditional curriculum (i.e., Cohen's $d = 0.5$) and the researcher wishes to achieve the statistical power of .80 with equal number of children in both groups, the researcher should at least recruit 64 children for each group.

Table 3 presents the *number of pairs required for dependent-sample t tests*. To use this table, researchers need to decide (1) the population effect size expressed as Cohen's *d* or Cohen's d_z, (2) the correlation between two measures (i.e., ρ), and (3) the desired statistical power. Cohen's *d* is the effect size of the study when two independent groups are used. Cohen's d_z is the effect size using two dependent groups. Given the example of game-based curriculum, suppose that the researcher expects the effect of game-based curriculum on two independent groups is 0.5 but the researcher is able to recruit two groups of students matched on their IQ scores with the correlation between the two measures to be .6, 28 pairs of children are required to achieve the statistical power of .8.

Table 4 presents the total *sample sizes required for one-way between-subjects ANOVAs*. To use this table, researchers need to decide (1) the population effect size expressed as η^2 or f, (2) the number of groups to compare, and (3) the desired statistical power. Assuming a researcher is interested in the effect of online learning for fourth grade math, she/he may design similar curriculums for face to face learning, online learning, and hybrid learning. If it is expected that 8% of variance on math performance can be explained by the different delivery methods ($\eta^2 = .08$), 114 children need to be recruited with 38 children in each group to achieve the statistical power of .80.

Table 5 presents the total *sample sizes required for testing the interaction effects of 2 × 2, 2 × 3, and 3 × 3 between-subjects ANOVAs*. To use this table, researchers need to decide (1) the population effect size expressed as partial η^2 or f, (2) the number of levels for the two independent variables, and (3) the desired statistical power. Assuming that a researcher is interested in the interaction effect of learning approach and flexible seating on learning fourth grade math, four settings may be set up for this study: (1) traditional curriculum with assigned seating, (2) traditional curriculum with flexible seating, (3) game-based learning with assigned seating, and (4) game-based learning with flexible seating. In this case, there are two levels of the effect of learning approach and two levels of the effect of seating. If the researcher further assumes the partial η^2 of the interaction is .08 and wishes to achieve the statistical power of .80, 96 children need to be recruited with 24 children in each condition. In fact, Table 4 shows 93 children are required. But, when we computed the number of children needed for each of the four conditions, it yielded 93/4 = 23.25. When the table does not yield a whole number for the sample size for each condition, researchers should round the number up to the nearest whole number for each condition.

Table 6 presents the total *sample sizes required for testing the interaction effects of 2 × 2 mixed ANOVAs*. To use this table, researchers need to decide (1) the population effect size expressed as partial η^2 or f, (2) correlation between the two repeated measures, and (3) the desired statistical power. Assuming that a researcher is interested in the interaction effect of learning approach and time on learning fourth grade math, the researcher may recruit two groups of children and measure their math performance before and after the completion of the different curriculums (e.g., game-based curriculum and traditional curriculum). If the researcher assumes the partial η^2 of the interaction is .08, the correlation between the two measures is .6, and the researcher wishes to achieve the statistical power of .80, 22 children need to be recruited with 11 children being assigned to each group.

Although we did not provide sample size tables for ANCOVAs, the procedures to estimate minimum sample sizes for ANCOVAs are similar to those for ANOVAs. When appropriate covariates are used, the effect size of the independent variable in ANCOVAs should be larger than that in ANOVAs. This is because the population residual variance (σ^2 in Equation 4) is reduced.

MINIMUM SAMPLE SIZE REQUIRED FOR THE REVIEWED ARTICLES

Research Question 5. Did educational technology researchers recruit enough number of participants for their studies? One may argue that no report on priori power analysis doesn't necessary imply inadequate sample size. To gauge whether sufficient number of participants were recruited for educational technology research, we analyzed the reviewed articles on the topic of game-based learning.

Ten out of the 178 reviewed articles focused on game-based learning. We examined the number of participants in those studies. Because we wanted to focus on ANOVA/ANCOVA and *t* tests, we excluded two of the 10 articles that did not report results based on any of these three tests. After identifying these eight articles, we first summarized the eight articles in terms of the themes, subject areas, the age of participants, the independent variables and the dependent variables, and the analysis strategies (Table 7). We then searched for external research studies on game-based learning that performed the same statistical tests with effect size reported or with enough information for us to calculate the effect size.

For all the statistical tests in the eight articles, we located at least three external published articles for a test. We coded all the effect size available from each of the external articles. We then used the medium of the coded effect sizes as the effect size for calculating the minimum sample size. When any of the eight articles included more than one tests (e.g.,

one-way ANOVA and dependent-samples *t* test), we estimated the sample size required for all the tests.

Table 7 presents the results of estimating sample size required for the eight articles. The shaded total sample size showed that not enough participants were recruited for that test. The results revealed that among the eight articles, authors of six articles might not recruit enough number of participants for one or more tests.

DISCUSSION

Low statistical power in empirical articles has been discussed in literature (Button et al., 2013; Cohen, 1992; Maxwell, 2004). Using sufficient sample size in empirical studies can reduce the probability of failing to reject a false null hypothesis and ensure adequate statistical power. In this paper, we examined current practices in power analysis using five years' publications in ETRD from 2014 to 2018. Consistent with findings from Peng et al. (2012), observed power analysis was conducted more often than priori power analysis. However, priori power analysis is recommended not observed power analysis (Cumming & Calin-Jageman, 2017; Hoening & Heisey, 2001; Yuan & Maxwell, 2005). Our analysis of a small pool of the reviewed articles suggested that educational technology researchers might not recruit enough participants for their studies. Because the application of priori power analysis is related to the availability of user friendly tools, we present sample size tables for *t* tests and ANOVAs to facilitate sample size planning for educational technology researchers.

The newly released journal article reporting standards for quantitative research in the APA Publications and Communications Board Task Force Report encourage researchers to describe the intended sample size, the achieved sample size (when it is different from the intended sample size), and strategies for determining sample size (Table 1 of Appelbaum et al., 2018). Specifically, the strategies include (1) a priori power analysis or sample size planning for precision of parameter estimates power, and (2) interim analyses and stopping rules (Appelbaum et al., 2018). The present study focuses on power analysis conducted by educational technology researchers. Interim analyses and stopping rules are used when researchers adopt sequential analyses. Readers who are interested in sequential analyses can refer to Lakens (2014).

Below we discuss difficulties that researchers may encounter when they conduct a priori power analysis. We summarize solutions for these difficulties that have been proposed in literature. We also present recommendations for reporting a priori power analysis and discuss the limitations of the paper.

DETERMINING EFFECT SIZE

Population effect size needs to be determined for a priori power analysis. Because population effect size is oftentimes unknown, how to come up with a reasonable estimate of population effect size may be the largest challenge in sample size planning (Anderson, Kelley, & Maxwell, 2017; Gelman & Carlin, 2014). In practice, a researcher may use a sample effect size obtained from a pilot study or previous studies. Yet pilot studies usually involve small sample sizes, and therefore, may result in variable estimation of the true effect size. When a previous study reported effect size using a biased measure, researchers may need to recalculate the effect size using another measure. Perugini et al. (2018) suggested that when sample effect sizes are obtained from SPSS η^2 or partial η^2, researchers can convert them to ε^2 or partial ε^2, which are less biased. For example, in one-way ANOVAs, the equation is

$$\varepsilon^2 = 1 - (1 - \eta^2) \times (\frac{N-1}{N-K}) \quad (5)$$

where *N* is the total sample size and *K* is the number of groups. In 2 × 2, 2 × 3, and 3 × 3 between-subjects ANOVAs, the equation is

$$\text{Partial } \varepsilon^2 = 1 - (1 - \text{partial } \eta^2) \times (\frac{N-K+df}{N-K}) \quad (6)$$

where df is the degrees of freedom of the effect of interest. Ideally, if researchers can locate multiple prior studies, a meta-analytic approach can be used to estimate the population effect size. This is similar to the procedures we used to estimate minimum sample size required for studies on the topic of game-based learning.

Some researchers discussed that published studies were usually with low statistical power but statistically significant results (Anderson et al., 2017; Maxwell, 2004). This may lead to positive bias of the population effect size (Anderson et al. 2017). Anderson et al. (2017) recommended an alternative approach for sample size planning that adjusts sample effect sizes for publication bias and uncertainty. They also developed the computing tools for this alternative approach.

NOT ENOUGH RESOURCES TO RECRUIT MORE PARTICIPANTS

Due to time and resource constraints, there may be a maximum number of participants that the researcher is able to recruit. As we mentioned earlier, in such a case, researchers may perform sensitivity analysis. When the results show that the study with the maximum number of participants can only detect a large effect but a small effect is expected to occur, the researcher may postpone the study until resources needed to achieve the desired power are available (Murphy & Myors, 1998). Maxwell (2004) suggested an alternative solution. Maxwell (2004) encouraged researchers to consider a collaborative multisite study. He cited Widaman (2000) and wrote "each of 1,000 psychologists [educational technology researchers] would obtain data on 1,000 individuals" (p. 161).

REPORTING A PRIORI POWER ANALYSIS

When reporting a priori power analysis, we agree with Cumming (2012) that researchers need to justify the choices of the research design, population effect size, and α. For example, if a population effect size of f = .25 is used, the researcher needs to explain reasons for using .25. It is not recommended to simply note that the selected effect size value is corresponded to a small, medium, or large effect (Aberson, 2019b). We also recommend researchers to report the tools they use for sample size planning, such as the sample size tables presented in this paper or R package 'pwr'. We quote the priori power analysis reported in one of our reviewed articles (Hegedus, Dalton, & Tapper, 2015) below. Hegedus et al. (2015) examined the effect of replacing traditional algebra 2 curriculum with a dynamic interactive software in two studies. They conducted a cluster-randomized trial in the first study and addressed how sample size was determined:

> A power analysis prior to the study confirmed that the sample size and numbers of classrooms (clusters) necessary for this study. We assumed that all the variability was at the student level for both treatment and control, and that intra-cluster variability was estimated at 0.10 following other classroom-based studies (National Re-search Council 2003). These were calculated in the Optimal Design software. We needed 28 clusters, to achieve power = 0.80 when ρ = 0.10, δ = 0.40 and 0.60, and *n* = 25. This is with a conventional α = 0.05. With a more liberal significance level of α = 0.25, we would require 14 clusters depending on the expected power. (Hegedus et al., 2015, p. 212)

LIMITATIONS

We examined articles published in ETRD from 2014 to 2018 to understand current practices in power analysis. The results may not be generalized to articles published in other educational technology journals or other years of ETRD. Current practices are impacted by editorial policies. Multiple tools were developed to perform power analysis. In this paper, we created sample size tables using G*Power 3.1.9.4. for popular statistical tests used by educational technology researchers. Our goal was not to present all the sample size tables for possible statistical tests but to provide fundamentals to facilitate research planning. Future research may aim to summarize the functionality and the strength and weakness of each computing tool. In addition, there are multiple ways to perform a priori power analysis using G*Power, we did not illustrate the alternatives in this paper. Researchers who are interested in alternative ways for performing a priori power analysis using G*Power can refer to the G*Power manual available at http://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html. Articles written by Faul et al. (2007), Faul et al. (2009) and Perugini et al. (2018) are also helpful.

CONCLUSION

Overall, our study reveals that more attention should be given for sample size determination. In educational technology research, most researchers use null hypothesis testing to examine the effectiveness of using technology for improving teaching and learning. When a null hypothesis is rejected, researchers may claim that the use of technology can enhance student learning. Otherwise, researchers may claim that the data did not provide evidence for the effectiveness of using technology. Yet, when sample size is not sufficient, the probability of correctly rejecting a false null hypothesis is low. A study with low statistical power may lead to an incorrect conclusion. For example, a researcher may incorrectly conclude that the use of a math tutoring system does not improve student self-efficacy. But in fact, the students become more confident on applying what they learned to different context after using the tutoring system. If the researcher had used adequate sample size, the researcher could have higher chance to correctly reject the false null hypothesis. Furthermore, more students would have been benefit from using the math tutoring system. In short, sample size matters! We have provided the fundamentals for sample size planning to promote best research practices in this paper. We are calling for researchers' attention of sample size determination.

REFERENCES

Aberson, C. L. (2019a). pwr2ppl: Power Analysis for Common Designs (Power to the People). R package version 0.1.1.

Aberson, C. L. (2019b). *Applied power analysis for the behavioral sciences* (2nd ed.). New York, NY: Routledge.

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

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science, 28*(11), 1547–1562. doi:10.1177/0956797617723724

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The

APA Publications and Communications Board task force report. *American Psychologist, 74*, 3–25. http://dx.doi.org/10.1037/amp0000191

Baydas, O., Kucuk, S., Yilmaz, R.M., Aydemir, M., & Goktas, Y. (2015). Educational technology research trends from 2002 to 2014. *Scientometrics, 105*, 709–725. https://doi.org/10.1007/s11192-015-1693-4

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 365–376. doi:10.1038/nrn3475

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., & Rosario, H. D. (2019). Basic Function for Power Analysis. R package version 1.2-2.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112* (1), 155–159.

Cumming, G. (2012). *Understanding the new statistics: effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

Cumming, G., & Calin-Jageman, R. (2017). *Introduction to the new statistics: estimation, open science, & beyond*. New York, NY: Routledge.

De Schryver, M., Hughes, S., Rosseel, Y., & De Houwer, J. (2016). Unreliable yet still replicable: A comment on LeBel and Paunonen (2011). *Frontiers in Psychology, 6*, 1–8. https://doi.org/10.3389/ fpsyg.2015.02039

Dupont, W. D., & Plummer, W. D. (2018). PS: Power and Sample Size Calculation. Available at http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSize

Dwork, D., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., & Roth, A. L. (2015). Preserving statistical validity in adaptive data analysis. *Proceedings of the 47th Annual ACM Symposium on Theory of Computing*. (pp. 117-126). Portland, Oregon, USA June 14, 2015. doi:10.1145/2746539.2746580

Faul, F. Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 171–191. https://doi.org/10.3758/BF03193146

Faul, F. Erdfelder, E., Buchner, A. & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science, 9*, 641–651. doi:10.1177/1745691614551642

Goodwin, L. D., & Goodwin, W. L. (1985). Statistical techniques in "AERJ" articles, 1979–1983: The preparation of graduate students to read the educational research literature. *Educational Researcher, 14*(2), 5–11. https://doi.org/10.3102/0013189X014002005

Hegedus, S. J., Dalton, S., & Tapper, J. R. (2015). The impact of technology-enhanced curriculum on learning advanced algebra in US high school classrooms. *Educational Technology Research and Development, 63*, 203–228. doi:10.1007/s11423-015-9371-z

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician, 55* (1), 19–24. doi:10.1198/000313001300339897

Hsu, Y.-C., Hung, J.-L., & Ching, Y.-H. (2013). Trends of educational technology research: more than a decade of international research in six SSCI-indexed refereed

journals. *Educational Technology Research and Development, 61*, 685–705. doi:10.1007/s11423-013-9290-9

Hwang, G. J., Lai, C. L., & Wang, S. Y. (2015). Seamless flipped learning: a mobile technology-enhanced flipped classroom with effective learning strategies. *Journal of Computers in Education, 2*(4), 449-473. https://doi.org/10.1007/s40692-015-0043-0

Kanyongo, G. Y., Brook, G. P., Kyei-Blankson, L., & Gocmen, G. (2007). Reliability and statistical power: How measurement fallibility affects power and required sample sizes for several parametric and nonparametric statistics. *Journal of Modern Applied Statistical Methods, 6*(1), 81–90. doi:10.22237/jmasm/1177992480

Kelley, K. (2019). MBESS: The MBESS R Package. R package version 4.6.0.

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods, 11*(4), 363–385. doi:10.1037/1082-989X.11.4.363

Kirk, R. E. (2008). *Statistics: an introduction* (5th ed.). Belmont, CA: Thomson.

Kirk, R. E. (2013). *Experimental design: Procedures for the behavior sciences* (4th ed.). Thousand Oaks, CA: Sage.

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology, 44*, 701–710. doi:10.1002/ejsp.2023

Levin, M., & Ensom, M. H. (2001). Post hoc power analysis: an idea whose time has passed? *Pharmacotherapy, 21*(4), 405–409. doi:10.1592/phco.21.5.405.34503

Levin, T., & Wadmany, R. (2008). Teachers' views on factors affecting effective integration of information technology in the classroom: Developmental scenery. *Journal of Technology and Teacher Education, 16*(2), 233-263.

Liu, L., & Chen, L. (2018). Conducting synchronous assessment through web videoconference to improve online learning: Case outcomes with nonparametric analysis. *Journal of Educational Technology Development and Exchange, 11*(1), 45-64. Retrieved from https://aquila.usm.edu/jetde/vol11/iss1/4/

Liu, L., & Maddux, C. (2008). Web 2.0 articles: Content analysis and a statistical model to predict recognition of the need for new instructional design strategies. *Computers in the Schools, 25*(3/4), 314-328. https://doi.org/10.1080/07380560802365856

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychological Methods, 9*(2), 147–163. doi:10.1037/1082-989X.9.2.147

Murphy, K. R., & Myors, B. (1998). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests*. Mahwah, NJ: Lawrence Erlbaum Associates.

Peng, C.-Y. J., Long, H., & Abaci, S. (2012). Power analysis software for educational researchers. *The Journal of Experimental Education, 80*(2), 113-136. doi: 10.1080/00220973.2011.647115

Perugini, M., Gallucci, M., & Costantini, G. (2018). A practical primer to power analysis for simple experimental designs. *International Review of Social Psychology, 31*, 1–23. https://doi.org/10.5334/irsp.181

Raudenbush et al. (2011). Optimal Design Software for Multi-level and Longitudinal Research (Version 3.01). Available from www.wtgrantfoundation.org

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized casual inference*. Belmont, CA: Wadsworth, Cengage Learning.

Shih, M., Feng, J., & Tsai, C.-C. (2008). Research and trends in the field of e-learning from 2001 to 2005: A content analysis of cognitive studies in selected journals. *Computers & Education, 51*, 955–967. doi:10.1016/j.compedu.2007.10.004

Widaman, K. F. (2000, October). *Scaling manifest and latent variables to promote a progressive stance in research.* Paper presented at the annual meeting of the Society of Multivariate Experimental Psychology, Saratoga Springs, NY.

Yuan, K.-H., & Maxwell, S. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics*, *30*, 141–167. https://doi.org/10.3102/10769986030002141

APPENDIX

TABLES

Table 2. *Approximate n_1 and n_2 Required for Independent-Sample t Tests with $\alpha = .05$ for Two-Tailed Tests*

		Statistical power (1–β)					
		.80		.90		.95	
Cohen's *d*	n_2/n_1	n_2	n_1	n_2	n_1	n_2	n_1
0.2	1	394	394	527	527	651	651
	1.5	492	328	659	439	814	542
	2	591	295	791	395	976	488
0.3	1	176	176	235	235	290	290
	1.5	220	146	294	196	362	242
	2	264	132	352	176	435	217
0.4	1	100	100	133	133	164	164
	1.5	125	83	166	110	205	137
	2	149	75	199	99	245	123
0.5	1	64	64	86	86	105	105
	1.5	80	54	107	71	132	88
	2	96	48	128	64	157	79
0.6	1	45	45	60	60	74	74
	1.5	56	38	74	50	92	62
	2	68	34	89	45	111	55
0.7	1	34	34	44	44	55	55
	1.5	42	28	55	37	68	46
	2	51	25	67	33	81	41
0.8	1	26	26	34	34	42	42
	1.5	32	22	43	29	53	35
	2	39	19	51	25	63	31
0.9	1	21	21	27	27	34	34
	1.5	26	18	35	23	42	28
	2	31	15	41	21	51	25
1.0	1	17	17	23	23	27	27
	1.5	22	14	28	18	35	23
	2	25	13	33	17	41	21

Table 3. *Approximate Number of Pairs Required for Dependent-Sample t Tests with α = .05 for Two-Tailed Test*

			Statistical power (1–β)		
Cohen's *d*	Cohen's d_z	Correlation btw two measures	.80	.90	.95
0.2	0.224	.6	159	212	261
	0.258	.7	120	160	198
	0.316	.8	81	108	133
0.3	0.335	.6	72	96	118
	0.387	.7	55	73	89
	0.474	.8	37	49	60
0.4	0.447	.6	42	55	67
	0.516	.7	32	42	51
	0.632	.8	22	29	35
0.5	0.559	.6	28	36	44
	0.645	.7	21	28	34
	0.791	.8	15	19	23
0.6	0.671	.6	20	26	31
	0.775	.7	16	20	24
	0.949	.8	11	14	17
0.7	0.783	.6	15	20	24
	0.904	.7	12	15	18
	1.107	.8	9	11	13
0.8	0.894	.6	12	16	19
	1.033	.7	10	12	15
	1.265	.8	8	9	11
0.9	1.006	.6	10	13	15
	1.162	.7	8	10	12
	1.423	.8	7	8	9
1.0	1.118	.6	9	11	13
	1.291	.7	7	9	10
	1.581	.8	6	7	8

Table 4. *Approximate Total Sample Sizes Required for One-Way Between-Subjects ANOVAs with Three to Five Groups and α = .05*

		Statistical power (1–β)								
		.80			.90			.95		
η^2	f	*k* = 3	*k* = 4	*k* = 5	*k* = 3	*k* = 4	*k* = 5	*k* = 3	*k* = 4	*k* = 5
.01	.101	957	1084	1190	1257	1408	1530	1533	1704	1845
.02	.143	477	540	590	624	700	760	762	848	915
.03	.176	315	360	395	414	464	505	504	560	610
.04	.204	237	268	295	309	348	375	375	420	455
.05	.229	189	212	235	246	276	300	297	332	360
.06	.253	156	176	195	204	228	250	246	276	300
.07	.274	132	152	165	174	196	210	210	236	255
.08	.295	114	132	145	150	168	185	183	204	220
.09	.314	102	116	130	132	148	165	162	180	195
.10	.333	90	104	115	117	132	145	144	160	175
.11	.352	84	96	105	108	120	130	129	144	160

.12	.369	75	84	95	96	108	120	117	132	145
.13	.387	69	80	85	90	100	110	108	120	130
.14	.403	63	72	80	81	92	100	99	112	120
.15	.420	60	68	75	75	88	95	93	104	115
.16	.436	54	64	70	72	80	90	87	96	105
.17	.453	51	60	65	66	76	85	81	88	100
.18	.469	48	56	60	63	72	80	75	84	90
.19	.484	45	52	60	60	68	75	69	80	85
.20	.500	42	48	55	54	64	70	66	76	80
.21	.516	42	48	50	51	60	65	63	72	75
.22	.531	39	44	50	51	56	60	60	68	75

Table 5. *Approximate Total Sample Sizes Required for Testing Interaction Effects of 2 × 2, 2 × 3, and 3 × 3 Between-Subjects ANOVAs with α = .05*

		Statistical power (1–β)								
		.80			.90			.95		
Partial η^2	f	2 × 2	2 × 3	3 × 3	2 × 2	2 × 3	3 × 3	2 × 2	2 × 3	3 × 3
.01	.101	779	957	1187	1043	1256	1530	1289	1532	1844
.02	.143	387	476	590	517	624	760	639	760	915
.03	.176	256	315	391	342	413	503	423	503	606
.04	.204	191	235	292	255	307	375	314	374	451
.05	.229	152	187	232	202	244	298	249	297	358
.06	.253	125	155	192	167	202	247	206	246	296
.07	.274	107	132	164	142	172	210	175	209	252
.08	.295	93	114	143	123	149	183	152	181	219
.09	.314	82	101	126	109	132	161	134	160	193
.10	.333	73	90	113	97	118	144	119	143	173
.11	.352	66	82	102	88	106	130	108	129	156
.12	.369	60	74	93	80	96	118	98	117	142
.13	.387	55	68	86	73	88	109	89	107	130
.14	.403	51	63	79	67	81	100	82	99	120
.15	.420	47	58	73	62	75	93	76	91	111
.16	.436	44	54	68	58	70	87	71	85	103
.17	.453	41	51	64	54	66	81	66	79	96
.18	.469	38	48	60	50	61	76	62	74	90
.19	.484	36	45	57	47	58	71	58	70	85
.20	.500	34	42	54	45	54	67	55	66	80
.21	.516	32	40	51	42	51	64	51	62	76
.22	.531	31	38	48	40	49	60	49	59	72

Note. When the table does not show a whole number for each condition, researchers need to round the number up to the nearest whole number for each condition.

Table 6. *Approximate Total Sample Sizes Required for Testing Interaction Effects in 2 × 2 Mixed ANOVAs With α = .05*

			Statistical power (1–β)		
Partial η^2	f	Correlation between two repeated measures	.80	.90	.95
.01	.101	.6	158	210	260
		.7	120	158	196
		.8	80	106	132
.02	.143	.6	80	106	130
		.7	60	80	98
		.8	42	54	66
.03	.176	.6	54	70	88
		.7	42	54	66
		.8	28	38	46
.04	.204	.6	40	54	66
		.7	32	40	50
		.8	22	28	34
.05	.229	.6	32	42	52
		.7	26	32	40
		.8	18	24	28
.06	.253	.6	28	36	44
		.7	22	28	34
		.8	16	20	24
.07	.274	.6	24	30	38
		.7	18	24	28
		.8	14	18	20
.08	.295	.6	22	28	32
		.7	16	22	26
		.8	12	16	18
.09	.314	.6	20	24	30
		.7	16	20	22
		.8	12	14	16
.10	.333	.6	18	22	26
		.7	14	18	20
		.8	10	12	14
.11	.352	.6	16	20	24
		.7	12	16	18
		.8	10	12	14
.12	.369	.6	14	18	22
		.7	12	14	18
		.8	10	12	12
.13	.387	.6	14	18	20
		.7	12	14	16
		.8	8	10	12
.14	.403	.6	12	16	20
		.7	10	12	16
		.8	8	10	12

Note. Number of groups = 2, Number of measurements = 2, and Nonsphericity correction ε = 1.

Table 7. *Estimating Minimum Sample Sizes for the Reviewed Research Studies on Game-based Learning*

Theme (leaning subject)	Participants	Independent variable(s) (IVs) and dependent variables (DVs)	Analysis
Timing of implementing digital game (math)	Sixth grade students	IV: Timing for gameplay (*N* =103, 90 minutes of game play and 90 minutes of instruction) (1) game play before instruction (*n* = 35) (2) game play during instruction (*n* = 41) (3) game play after instruction (*n* = 27) DV: knowledge of algebraic properties and ability to solve linear equations	• One-way ANCOVA ➢ Power analysis: ->Median of the ES: η^2 = .21 with α = .05, power = .90, numerator df = 2, number of groups = 3, number of covariates = 1 ->**Total sample size = 51**
Peer assessment-based game (natural science)	Sixth grade students	IV: Peer assessment-based game (*N* = 167, 10-weeks, 50 minutes per week) (1) peer assessment-based game development approach (*n* = 82) (2) conventional game development approach (*n* = 85) DV: learning achievements, motivations, and problem-solving skills	• One-way ANCOVA ➢ Power analysis (learning outcome): ->Median of the ES: η^2 = .21 with α = .05, power = .90, numerator df = 1, number of groups = 2, number of covariates = 1 ->**Total sample size = 42** ➢ Power analysis (attitude): ->Median of the ES: η^2 = .075 with α = .05, power = .90, numerator df = 1, number of groups = 2, number of covariates = 1 ->**Total sample size = 132**

Animal identification (natural science)	Fifth grade students	IV: Game-based (N = 53, 60 minutes) (1) two-tier test-based educational game (n = 26) (2) conventional e-learning (n = 27) DV: achievement, motivation, ease of use, usefulness	• One-way ANCOVA ➢ Power analysis (learning outcome): ->Median of the ES: η^2 = .21 with α = .05, power = .90, numerator df = 1, number of groups = 2, number of covariates = 1 ->**Total sample size = 42** ➢ Power analysis (attitude): ->Median of the ES: η^2 = .075 with α = .05, power = .90, numerator df = 1, number of groups = 2, number of covariates = 1 ->**Total sample size = 132**
Fraction learning game (math)	Third grade students	IV: The extent to game-like apps (N = 95, 90 minutes on the apps) (1) games (2) worksheets DV: knowledge and enjoyment	• 2 × 2 repeated measures ANOVA (one between-subjects variable and one within-subjects variable) ➢ Power analysis: I. Group effect: ->Median of the ES: partial η^2 = .037 with α = .05, power = .90, number of groups = 2, Number of measurements = 2, Corr among rep measures = 0.6 ->**Total sample size = 222** II. Group × time effect: ->Median of the ES: partial η^2 = .175 with α = .05, power = .90, number of groups = 2, Number of measurements = 2, Corr among rep measures = 0.6, Nonsphericity correction = 1 ->**Total sample size = 14**

Extrinsically or intrinsically integrated content (math)	Vocational secondary education students (14-17 years old)	IV: Types of game-based learning environment (*N* = 58, four 50 minutes course hours) (1) extrinsically integrated (*n* = 30) (2) intrinsically integrated (*n* = 28) DV: knowledge, motivation, usefulness, playfulness	• 2 × 2 repeated measures ANOVA(one between-subjects variable and one within-subjects variable) ➢ Power analysis (knowledge): I. Group effect: ->Median of the ES: partial η^2 = .037 with α = .05, power = .90, number of groups = 2, Number of measurements = 2, Corr among rep measures = 0.6 ->**Total sample size = 222** II.Group x time effect: ->Median of the ES: partial η^2 = .175 with α = .05, power = .90, number of groups = 2, Number of measurements = 2, Corr among rep measures = 0.6, Nonsphericity correction = 1 ->**Total sample size = 14** • One-way ANOVA ➢ Power analysis (attitudes): ->Median of the ES: η^2 = .055 with α = .05, power = .90, number of groups = 2 ->**Total sample size = 184**

Diagnostic mechanism strategy (math)	Second grade students	IV: Digital game-based learning (DGBL) system with or without a diagnostic mechanism (*N* = 53, two 40 minutes lessons for 6 weeks) (1) DGBL with a diagnostic mechanism (*n* = 29) (2) DGBL without a diagnostic mechanism (*n* = 27) DV: learning outcome, anxiety	• Learning outcome -> 2 x 6 mixed ANOVA (6 levels of the within-subjects variable) ➢ Power analysis: I. Group effect: ->Median of the ES: partial η^2 = .037 with α = .05, power = .90, number of groups = 2, Number of measurements = 6, Corr among rep measures = 0.6 ->**Total sample size = 186** II.Group x time effect: ->Median of the ES: partial η^2 = .175 with α = .05, power = .90, number of groups = 2, Number of measurements = 6, Corr among rep measures = 0.6, Nonsphericity correction =1 ->**Total sample size = 8** • Paired sample *t* tests for both groups ➢ Power analysis: ->Median of the ES: d = .45 with α = .05, power = .90 ->**Total sample size = 54 (for each group)**
Online flexible game (math)	12-14 years old	IV: Game-based learning (*N* = 79, 14 weeks) (1) game to solve math problems (*n* = 38) (2) soling problems on paper (*n* = 41) DV: attitudes (value, enjoyment, self-confidence, motivation)	• One-way ANCOVA ➢ Power analysis (attitude): ->Median of the ES: η^2 = .075 with α = .05, power = .90, numerator df = 1, number of groups = 2, number of covariates =1 ->**Total sample size = 132**

Digital board games (English)	High school students	IV: Digital board game language learning (*N* = 96, 50 minutes) (1) ordinary instruction group (*n* = 32) (1) board game language-learning group (*n* = 32) (2) digital board game language-learning group (*n* = 32) DV: learning performance, intrinsic motivation	• One-way ANCOVA ➢ Power analysis (learning performance): -> Median of the ES: η^2 = .21 with α = .05, power = .90, numerator df = 2, number of groups = 3, number of covariates = 1 ->**Total sample size = 51** • One-way ANOVA ➢ Power analysis (attitude): ->Median of the ES: η^2 = .055 with α = .05, power = .90, number of groups = 3 ->**Total sample size = 222**