The Instrumental Approach and Academic Achievement in Mathematics: A Systematic Review, Reliability Evaluation and Meta-analysis

Francisco J. Alvarez-Montero* *, Eneyda Rocha-Ruiz, Maria G. Leyva-Cruz, Flerida Moreno-Alcaraz, Anselmo Alvarez-Arredondo
Facultad de Ciencias de la Educacion, Universidad Autonoma de Sinaloa, Ave. Cedros y Calle Sauces s/n Fracc. Los Fresnos, C.P. 80034. Culiacan, Sinaloa, Mexico
**Corresponding Author: francisco_alvarez_montero@uas.edu.mx*

ABSTRACT

The Instrumental Approach and its related notions such as Instrumental Genesis and Instrumental Orchestrations, form a theoretical framework often cited in research on the teaching and learning of mathematics. Nevertheless, the impact of such interventions on academic achievement has not been investigated. Moreover, the reliability of the statistical results of these studies and the issues concerning the sample size needed to produce reliable and replicable results has not been addressed either. In this sense, this article presents a systematic review, meta-analysis and reliability assessment of studies conducted during the 2001-2017 period, which used the Instrumental Approach as a theoretical framework. Six conclusions can be made from the analysis. First, that there is a very limited set of interventions based on the Instrumental Approach that seek to improve academic performance. Second, that individually most of the studies are statistically unreliable due to wide confidence intervals and low statistical power. Third, that although the average effect size is positive it is below Hattie's zone of desired effects. Fourth, that the prediction interval is very wide showing a high level of heterogeneity and dispersion of effects. Fifth, that that a considerable percentage of future interventions can be potentially harmful. Sixth, that conducting original studies that are reliable and replicable require sample sizes that are way above those commonly found in the literature.

Keywords: Academic Achievement, Instrumental Approach, Instrumental Genesis, Instrumental Orchestrations, Meta-analysis, Precision, Reliability, Confidence Interval, Statistical Power, Sample Size.

1. INTRODUCTION

At the heart of science is an essential balance between two seemingly contradictory at-attitudes--an openness to new ideas, no matter how bizarre or counterintuitive they may be, and the most ruthless skeptical scrutiny of all ideas, old and new. This is how deep truths are winnowed from deep nonsense (Sagan, 1997 p. 287).

However, in the field of education this skeptical and rigorous analysis seems to be lacking. Myth and reality have been difficult to separate in this area (Bloom, 1972; De Bruyckere, Kirschner, and Hulshof, 2015; Holmes, 2016), and the adoption of educational practices and programs has been based on ideologies, fads and marketing, rather than on the scientific evidence available about their effectiveness (Kirschner & van Merriënboer, 2013, Slavin, 2008b). Furthermore, 91.5% of publications in education, social sciences and psychology confirm their hypotheses (Fannelli, 2010, 2011), which produces the illusion that everything seems to work when it comes to improving learning (Hattie, 2009).

Consequently, there has been a call to base educational practice and policy on evidence from rigorous experiments (Kowalski, 2009; Slavin, 2008a, 2017) and to favor those

approaches with strong evidence. This way, what is known and true can be acted on and it can be established what new ideas are worth considering and how they can be tested, while what is superstition, fad, and myth can be discarded (Bloom, 1972).

The issue of strong evidence has recently been the subject of much debate inside and outside the field of education with respect to the replicability of scientific findings (Asendorpf et al., 2013; Begley & Ioannidis, 2015; Goodman, Fanelli, & Ioannidis, 2016; Ioannidis, 2012; Ioannidis et al., 2014; Iqbal et al., 2016; Kepes & McDaniel, 2013; Munafò et al. 2017; Open Science Collaboration, 2015; Peng, 2011; Stodden, 2016). The evidence shows that across different research areas, replicability, namely the obtention of similar results using different samples and the same research design, is compromised due to the lack of reliability and accuracy of the statistical data.

Sample size is at the center of the problem as there are many studies carried out with small samples (Bakker, van Dijk, and Wicherts, 2012; Flint et al., 2015; Lohse, Buchanan & Miller, 2016, Marszalek et al., 2011; Vadillo, Konstantinidis and Shanks, 2016). Small samples negatively impact the strength of the evidence in three different ways. First, the confidence intervals of the estimates are embarrassingly large (Brand & Bradley, 2016; Peters & Crutzen, 2017). Second, effects tend to be inflated or are false positives (Asendorpf et al., 2013; Bakker, van Dijk, & Wicherts, 2012; Eklund, Nichols, & Knutsson, 2016; Fanelli, 2010, 2011; Forstmeier , Wagenmakers, & Parker, 2016; Ioannidis, 2005, 2008, 2012; Ioannidis, Tarone, & McLaughlin, 2011; Simmons, Nelson & Simonsohn, 2011). Third, the achieved statistical power is very low (Asendorpf et al., 2013; Bakker, van Dijk, & Wicherts, 2012; Bogg & Lasecki, 2014; Button et al., 2013; Christley, 2010; Ioannidis, 2008; Keen, Pile, & Hill, 2005; Maxwell, 2004; Lohse, Buchanan & Miller, 2016, Vadillo, Konstantinidis, & Shanks, 2016). Additionally, within the field of education, even when the evidence is statistically sound, the number of studies providing support for a hypothesis is very low. For example, Mayer (2014, p. 19-21) reports that in the field of computer games for learning, only 10% (95% CI [2,18]) of the literature addresses academic performance or the development of cognitive skills.

In this sense, this paper presents a systematic review and meta-analysis (Koretz and Lipman, 2017, Ravindran and Shankar, 2015), as well as a reliability evaluation of statistical results, of studies which have used the Instrumental Approach to improve academic achievement in mathematics. Although this approach and its related notions (i.e., Instrumental Genesis and Instrumental Orchestrations) have been used as a theoretical framework for improving the teaching and learning of mathematics since the late 1990s (Artigue, 2002; Drijvers, 2000; Guin & Trouche, 1998; Hollebrands & Okumuş, 2017; Lagrange, 1999), there are no reported efforts that have sought to examine their impact on academic performance or the accuracy and reliability of their statistical results. Furthermore, using the effects calculated for the meta-analysis, four methods for calculating sample sizes were compared. Two methods are based on statistical power and the other two on confidence interval width.

Therefore, the objectives of this research are the following: 1) to determine if the problems of precision and reliability of statistical results are also present in the field of teaching and learning mathematics; 2) to find out for a simple research design of one control group and one treatment group, and a wide range of effects, which method gives the best results in terms of confidence interval width and statistical power.

The rest of this manuscript is organized as follows. First, the interpretation of effect sizes with respect to academic performance is addressed. Secondly, the software used for statistical calculations is discussed, as well as the criteria to determine the reliability and precision of the results. Third, the search strategy and inclusion criteria are described. Fourth, the results obtained are presented. Finally, the findings obtained from the results are discussed.

2. THRESHOLDS FOR INTERPRETING EFFECT SIZES

Interpretation is essential if researchers are to extract meaning from their results (Ellis, 2009). One of the most used conventions for effect size (ES) interpretation is the one proposed by Cohen (2013). He established the following thresholds for interpreting effect sizes: trivial [0.000-0.199], small [0.200-0.499], medium [0.500-0.799] and large [≥0.800]. However, dissociated from a context of decision and comparative value, these intervals cannot be adequately interpreted (Glass, Smith, & McGaw., 1981).

To contextualize the effects presented in this manuscript, John Hattie's (2009) barometer of influences is used. It is based on more than 800 meta-analyses, comprehending 138 factors or independent variables, all of them related to academic achievement. The barometer is made to convey two things. First, if the same results can be obtained by other well-known factors not included in the analysis. Second, if the ES is above or below average.

The average effect size in Hattie's analysis is d= .400, and it establishes a level at which the effects of an innovation improve performance or academic achievement, in such a way that differences in learning can be clearly perceived. From this perspective, results in the range [0, 0.149] are below average and can be attained merely by growing up. That is, through the physical and emotional process of maturation that comes with age. An ES within this range can be considered potentially harmful and probably should not be implemented (Hattie, 2009).

Effect sizes in the following range [0.150-0.399] are below average and can be replicated by the work of an average teacher. However, this does not imply that simply placing a teacher in front of a class would lead to these results. Rather, that the teaching methods used, the level of expectations for the students and, the quality of the student-teacher relationships reflect that of a typical teacher and not that of the best teachers (Hattie, 2009). An approach with an ES within this interval can be regarded as in need of more consideration.

Hence, interventions seeking to improve AC should be looking to attain effect sizes greater than d = 0.400 to be considered above average, and greater than d = 0.600 to be considered excellent (Hattie, 2009). Figure 1 shows a graphical representation of Hattie's barometer.

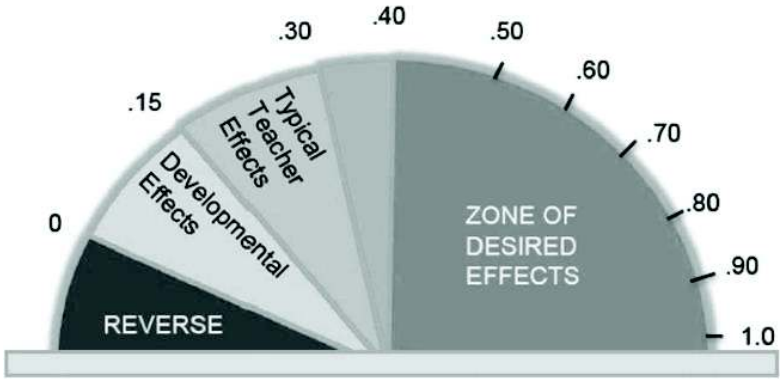

Figure 1. Hattie's barometer of influence w.r.t. academic achievement

The next section presents the software used for the statistical analysis, as well as the criteria for evaluating the accuracy and reliability of the statistical results.

3. Software used for statistical computations and reliability criteria

The Excel sheets developed by Bailey (2009) and Lakens (2013), as well as the companion online calculator for the book Practical meta-analysis (Lipsey, & Wilson,, 2001) available from: http://cebcp.org/practical-meta-analysis-effect-size-calculator/, were used to

calculate the individual effect sizes of the interventions and their confidence intervals. The statistical power of the effects was estimated through G* Power (Faul, Erdfelder, Buchner, & Lang, 2009).

The meta-analysis was performed in R (Gentleman, Huber, & Carey, 2011) using the procedure, packages and functions described by Quintana (2015). The REML (Restricted Maximum Likelihood) method was used to estimate the variability between effects or heterogeneity (i.e., τ^2). In addition, following the recommendations of Chiolero et al. (2012) and IntHout, Ioannidis, Rovers, & Goeman. (2016), a prediction interval was estimated, as well as the probabilities that future studies will obtain effects within Hattie's zone of desired effects and below such area.

A statistical result was considered reliable if and only if (1) the confidence interval does not contain zero, (2) the width of the confidence interval is less than the estimated effect (Brand & Bradley, 2016) and (3) the achieved or observed power is .80 or greater. However, these rules do not always work when the sample is small, and the ES is large. For example, the study by Elgamal, Abas, & Baladoh (2013) in the area of computer programming yields the results presented in table 1.

Table 1. Statistical results of the study by Elgamal, Abas, & Baladoh (2013)

d	*95% CI*	CI width	Power
1.719	[1.025, 2.381]	1.324	0.999

Based on the criteria defined above, this result should be considered reliable. However, because any estimate of an effect is drawn at random from the corresponding sampling distribution (Peters & Crutzen, 2017), an interval width of 1.324 implies a high degree of inaccuracy, and most likely, a replication study will obtain a result very different from the original (Halsey, Curran-Everett, Vowler, & Drummond 2015). Consequently, for a result to be reliable, it must also have an adequate level of accuracy. Since there are no guidelines in the literature regarding the width of the confidence interval, the decision was made that the effects greater than or equal to .400 must have a full confidence interval width of 0.400 (± 0.200 half width) to be considered precise. Effects below 0.400 must have an interval width less than the effect itself.

The evaluation of the different methods for calculating sample size was done using the effect sizes obtained for the meta-analysis. The statistical power and the width of the confidence interval produced for each method were compared to determine which method produced the most stable results in terms of interval width and power, across different effect sizes.

The methods based on statistical power are represented by the *safeguard.d* and *SSR.d*. functions developed in R by Perugini, Gallucci, and Costantini (2014). These functions compute sample sizes using a target ES (the parameter *d* in both functions), the desired statistical power (the parameter *power* in both functions) and the percentage of times that power will be obtained, if the same study would be replicated an infinite number of times (the *conf* parameter in both functions). In this study, the value for both the *conf* and *power* parameters was .80.

The procedures based on confidence interval width correspond to the *pwr.cohensdCI* and *ss.aipe.smd* functions, included in the *userfriendlyscience* (Peters and Crutzen, 2017) and *MBESS* (Kelly, 2007) R packages. The *pwr.cohensdCI* functions expects an ES (the parameter

d), the half width of the confidence interval (the parameter *w*) and the α level (the parameter *conf.level*). The *ss.aipe.smd* function takes an ES (the parameter *delta*), the width of the interval (the parameter *width*), the α level (the parameter *conf.level*) and the type of width (the parameter *which.width*). For this investigation, the values used for the parameters of both functions are the following: *conf.level*=0.95*, w*=0.2*, width*=0.4*, which.width*=”Full” . However, for effects less than 0.300 the values of the parameters *w* and *width* were 0.15 and 0.3 respectively.

The search strategy and inclusion criteria used in the systematic review are described in the next section.

4. SEARCH STRATEGY AND INCLUSION CRITERIA

Google Scholar (GS) was used as the search engine for the literature. It can be argued that GS is not a bibliographic or scientific literature database; however, the results of studies conducted in the last nine years (Chen, 2010, De Winter, Zadpoor, & Dodou, 2014; Haddaway, Collins, Coughlin, & Kirk, 2015; Harzing, 2014; Harzing & Alakangas, 2016; Howland, Wright, Boughan, & Roberts, 2009; Nourbakhsh, Nugent, Wang, Cevik, & Nugent, 2012), indicate that GS is at the same level of databases such as Scopus, Web of Science and PubMed, in terms of coverage and the recovery of relevant literature.

Due to restrictions on the number of characters that searches in GS can support, five searches were done with slightly different search strings. The searches and their strings are presented in table 2.

Table 2. Searches performed and their search strings

Search number	Search string
1	("instrumental genesis" OR "instrumental orchestration*" OR instrumentation OR *orchestrations) AND ("t-test" OR Regression OR ANOVA OR ANOVA OR MANOVA OR MANCOVA) AND (Drijvers OR Artigue Or Trouche OR Guin OR Laborde)
2	("instrumental approach" OR "instrumental genesis" OR "instrumental orchestration") AND ("t-test" OR Regression OR ANOVA OR ANCOVA OR MANOVA) AND ~Math AND (Artigue OR Trouche OR Guin OR Drijvers OR Haspekian OR Laborde) AND

	(~education OR ~teaching OR ~learning OR ~instruction)
3	("instrumental approach" OR "instrumental genesis" OR "instrumental orchestration") AND ("t-test" OR Regression OR ANOVA OR ANCOVA OR MANOVA) AND ~Math AND (Artigue OR Trouche OR Guin OR Drijvers OR Haspekian OR Laborde OR Rabardel) AND (~teaching OR ~learning)
4	("instrumental approach" OR "instrumental genesis" OR "instrumental orchestration") AND ("t-test" OR Regression OR ANOVA OR ANCOVA OR MANOVA OR MANCOVA) AND ~Math AND (Artigue OR Trouche OR Guin OR Drijvers OR Haspekian OR Laborde OR Gueudet OR Rabardel)
5	("instrumental approach" OR "instrumental genesis" OR "instrumental orchestration" OR Instrumentation OR genesis) AND ("t-test" OR Regression OR ANOVA OR ANCOVA OR MANOVA) AND (Artigue OR Trouche OR Guin OR Laborde) AND ~Math AND (~teaching OR ~learning)
6	("instrumental genesis" OR "instrumental orchestration" OR "instrumental approach")

AND (t-test OR Regression OR ANOVA OR ANCOVA OR MANOVA OR "descriptive statistics" OR "summary statistics") AND (geometry OR calculus OR algebra) AND (teaching OR learning)

The previous search strings were used to download potential manuscripts. The search was done year by year covering the 2001-2017 period. Afterwards, a selection of documents was done. In order to be included in the quantitative analysis, a downloaded document had to provide enough information to calculate an effect size, such as sample size, means and standard deviations (Mayer, 2014, p. 43) or any other useful statistic that can be used to calculate an effect size as described in the book Practical Meta-Analysis by Lipsey & Wilson (2001) and its companion website: http://cebcp.org/practical-meta-analysis-effect-size-calculator/. Additionally, each manuscript included in the final analysis was inspected to determine the type of technology, educational level, subject matter and instructional method used in the interventions.

The results of the search and manuscript selection stages are described next.

5. LITERATURE SEARCH AND INCLUSION RESULTS

In total, one hundred and thirteen (113) manuscripts were downloaded. However, ninety-two (93) were excluded with reasons and only twenty (20), that is 18.584% of the documents, met the inclusion criteria (see Fig. 2). The type of publication of each manuscript included in the final analysis is shown in tables 3.

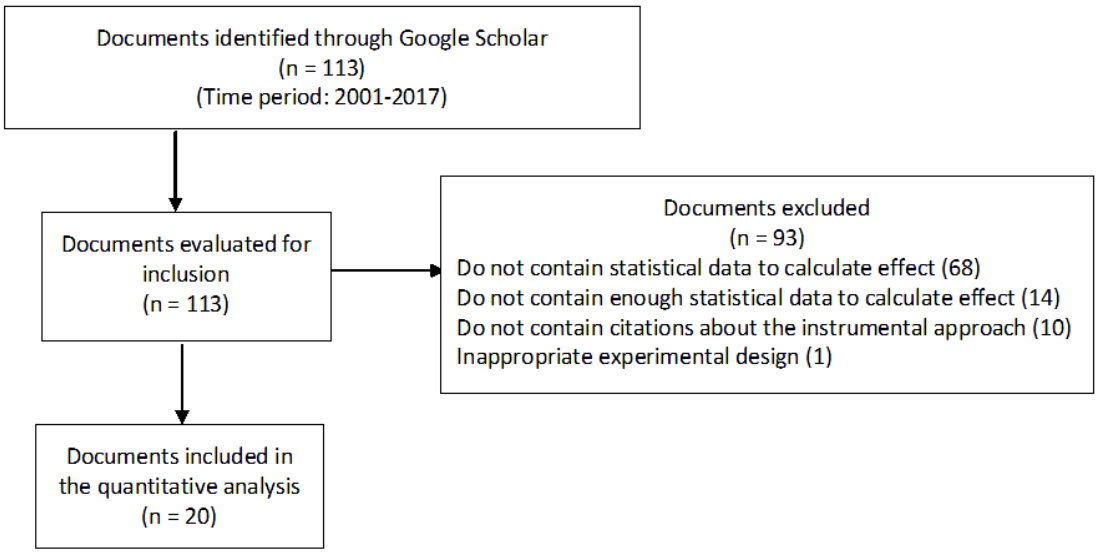

Figure.2. Document search and inclusion flowchart

Table 3. Documents included in the final analysis and their classification

Document	Type of document
Tarmizin & Tajudin (2006)	Symposium
Javed (2008)	Unpublished doctoral dissertation
Velez Caraballo (2008)	Unpublished doctoral dissertation
Gantz (2010)	Unpublished doctoral dissertation
El-Jiryes (2011)	Unpublished doctoral dissertation
Jiang & White (2012)	Congress
Curri (2012)	Unpublished master's dissertation
Rieß y Greefrath (2013)	Congress
DeLoach (2013)	Unpublished doctoral dissertation
Spencer (2013)	Unpublished doctoral dissertation
Drijvers et al. (2014)	Journal
Juan (2015)	Unpublished doctoral dissertation
Jupri, Drijvers, y van den Heuvel-Panhuizen (2015)	Journal
Pelech(2015)	Unpublished doctoral dissertation
Ljajko (2016)	Journal
Rich (2016)	Unpublished doctoral dissertation
Marsh (2016)	Unpublished doctoral dissertation
Kostić et al. (2016)	Journal
Mainali & Heck (2017)	Journal
Ocal (2017)	Journal

The following section presents the results of the categorical and statistical data gathering, the effects obtained, the evaluation of their reliability and precision, the comparison of sample size methods and the final meta-analysis.

5. DATA GATHERING, META-ANALYSIS AND EVALUATION RESULTS

The categorical data obtained from the manuscripts is presented first. Table 4 shows the type of technology, educational level, subject matter and instructional method used in each analyzed document. Figures 3, 4, 5 and 6 present the frequency distributions for technology

type, educational level, subject matter and instructional method. The category "Other" in figures 3 and 6 indicates technology types and instructional methods that appear only once.

Table 4. Type of technology, educational level, subject matter and instructional method used in each intervention

Manuscript	Technology	Educ. Level	Subject	Instruc. method
Tarmizin & Tajudin (2006)	Graphing Calculator (TI-83)	Middle School	Algebra	NA
Javed (2008)	Rigid Tutor	Vocational Education and Training (VET)	Algebra	Blenden learning (Sharma, 2010)
Velez-Caraballo (2008)	Graphing Calculator (TI Nspire)	High School	Algebra	Cooperative learning (Slavin, 2011)
Gantz (2010)	Graphing calculator (TI-Nspire)	Middle School	Algebra	Cooperative learning (Slavin, 2011)
El-Jiryes (2011)	Graphing Calculator (Casio ClassPad 300) Dynamic Geometry Software (GeoGebra)	High School	Algebra	NA
Curri (2012)	SimReal	High School	Trigonometry	NA
Jiang & White (2012)	Dynamic Geometry	Middle School, High School	Geometry	NA

	Sotware (Unknown)			
DeLoach (2013)	Graphing Calculator (Unknown)	High School	Algebra	Non-rule-based instruction (Merriweather & Tharp, 1999)
Rieß y Greefrath (2013)	Graphing Calculator (Casio Classpad)	High School	Algebra	NA
Spencer (2013)	Graphing Calculator (TI-84)	High School	Algebra	NA
Drijvers et al. (2014)	Web-based rigid tutor	Middle School	Algebra	NA
Juan (2015)	Dynamic Geometry Software (GeoGebra)	Middle School	Geometry	University of Chicago School Mathematics Project: Everyday Mathematics (Carroll, 1998, Thompson & Senk, 2001)
Jupri, Drijvers, & van den Heuvel-	Web-based Rigid Tutor	Middle School	Algebra	NA

Panhuizen (2015)				
Pelech (2015)	Graphing Calculator (TI-NSpire)	High School	Algebra	Four principles of teaching (Dick & Burrill, 2009)
Kostić et al. (2016)	Dynamic Geometry Software (GeoGebra)	University	Chemistry	NA
Ljajko (2016)	Dynamic Geometry Software (GeoGebra)	High School	Geometry	NA
Marsh (2016)	Dynamic Geometry Software (GeoGebra)	High School	Algebra	NA
Rich (2016)	Study Island	Elementary School	NA	NA
Mainali & Heck (2017)	Dynamic Geometry Software (GeoGebra)	Middle School	Geometry	NA
Ocal (2017)	Dynamic Geometry Software	University	Calculus	NA

(GoeGebra)

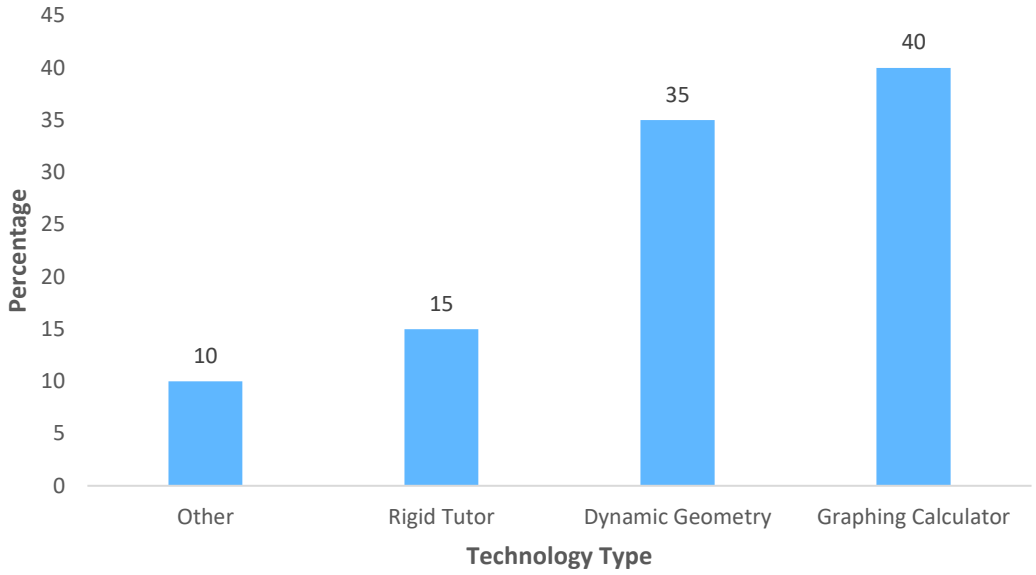

Figure 3. Frequency distribution for technology types expressed as percentages

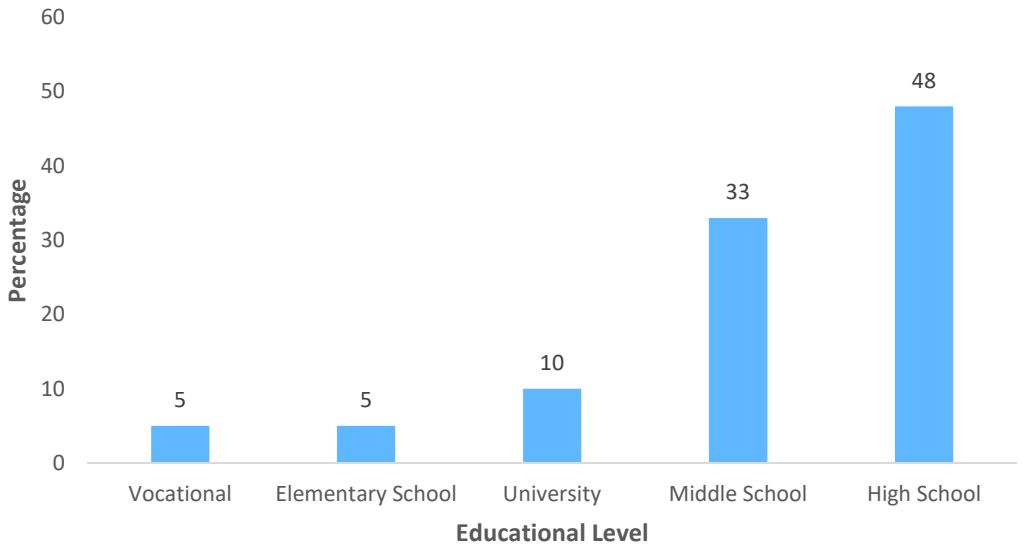

Figure 4. Frequency distribution for educational levels studied expressed as percentages

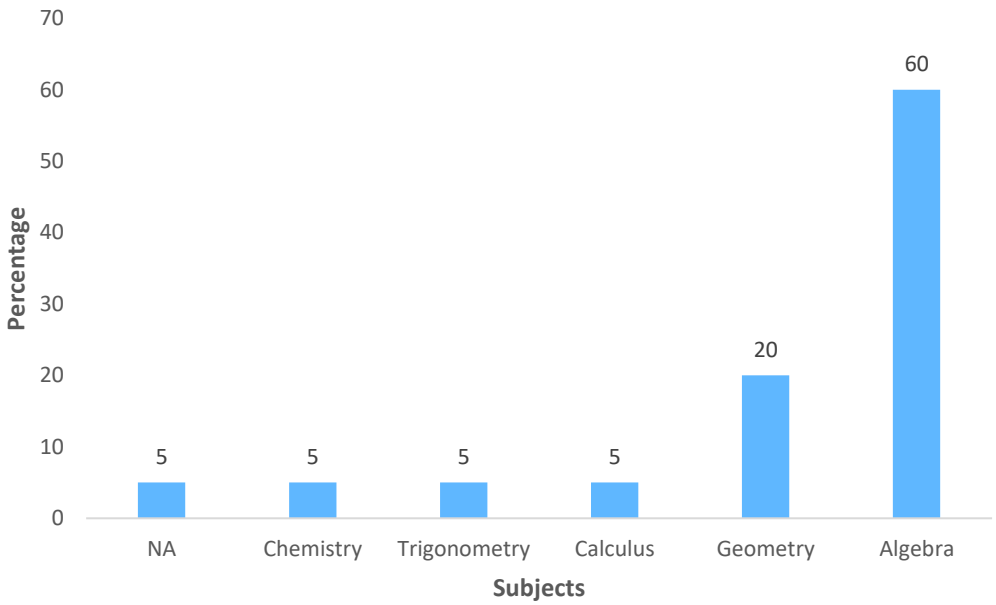

Figure 5. Frequency distribution for subject matters expressed as percentages

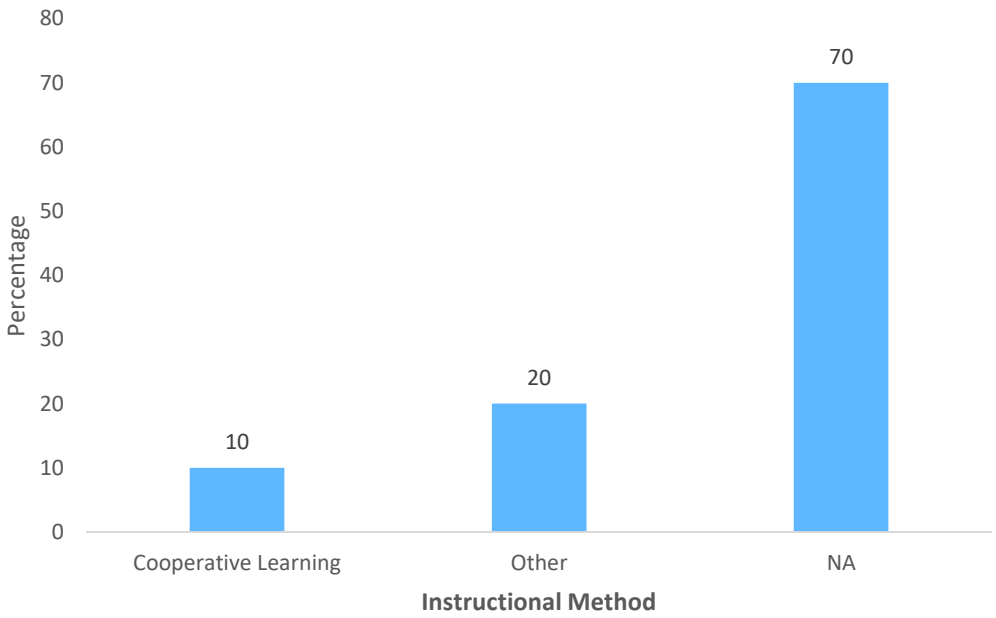

Figure 6. Frequency distribution for instructional methods expressed as percentages

The statistical data from which the effect sizes of the interventions was calculated is presented second. However, the data was not homogenous as some interventions reported sample size, means and standard deviations, and others the results of hypothesis testing

procedures such as t-tests and ANOVAS, etc. This heterogeneity is shown in tables 5 through 8. Some manuscripts reported statistical data from two parts of the same achievement test and some others described more than one intervention with different samples. This kind of data is reflected using alphabetical suffixes after the document in-text citation such as: Jiang & White (2012)a, Jiang & White (2012)b. In this sense, Jiang & White (2012) report three different interventions, Drijvers et al. (2014) the results of two parts of the same achievement test, while Rich (2016) presents studies that were conducted during 6 semesters spanning three years.

Table 5. Statistics for studies with one control group and one treatment group which reported sample sizes, means and standard deviations

Intervention or Evaluation	N	n.Treat	M.Treat	SD.Treat	n.Ctrl	M.Ctrl	SD.Ctrl
Tarmizin & Tajudin (2006)	40	21	59	10.252	19	59.260	21.1890
Javed (2008)	15	8	60.630	22.960	7	69.140	20.040
Velez-Caraballo (2008)	93	46	58.850	14.350	47	55.170	17.200
El-Jiryes (2011)	49	25	12.640	2.480	24	12.540	3.040
Jiang & White (2012)a	508	276	54.190	17.640	232	46.810	15.100
Jiang & White (2012) b	373	210	71.260	16.090	163	69.280	15.500
Jiang & White (2012) c	58	15	88.270	7.01	43	87.070	10.100

Curri (2012)	22	11	79.200	14.200	11	74.900	14.900
Rieß y Greefrath (2013)	242	152	59	13	90	55	16
DeLoach (2013)	53	28	54	20	25	42	13
Spencer (2013)	405	222	655.030	7.870	183	657.740	9.680
Drijvers et al. (2014) a	810	404	6.739	1.744	406	6.978	1.724
Drijvers et al. (2014) b	799	400	6.231	2.006	399	6.386	2.114
Juan (2015)	139	70	84.770	13.761	69	79.360	19.407
Jupri, Drijvers, y van den Heuvel-Panhuizen (2015)	250	131	4.692	2.406	119	3.023	2.446
Rich (2016)a	130	54	803.760	31.220	76	826	38.730
Rich (2016)b	150	76	817.610	32.810	74	812.200	36.460
Rich (2016)c	145	74	828.260	36.190	71	825.700	36.240
Rich (2016)d	134	58	839.900	50.260	76	826	38.730
Rich (2016)e	137	63	811.460	34.680	74	812.200	36.460
Rich (2016)f	140	69	838	30.550	71	825.700	36.240
Marsh (2016)a	39	19	14.950	14.530	20	19.250	19.450

Marsh (2016)b	39	19	27	24.160	20	15.250	12.770
Kostić et al. (2016)	90	45	70.690	20.130	45	48.640	23.140
Mainali & Heck (2017)a	26	13	263.700	57.400	13	191.200	61.400
Mainali & Heck (2017)b	41	21	215.600	48.800	20	163.900	53
Ocal (2017)	55	31	14.730	4.897	24	11.808	4.658

Abbreviations. N, total sample size; n.Treat, sample size of the experimental or treatment group; M.Treat, mean of the experimental or treatment group; SD.Treat, standard deviation of the experimental or treatment group; n.Ctrl, sample size of the control group; M.Ctrl, mean of the control group; SD.Ctrl, standard deviation of the control group.

Table 6. Anova results from Gantz's (2010) study

N	n.Treat	n.Ctrl	F value
32	18	14	1.515

Table 7. Descriptive statistics from Pelech's (2015) with 3 treatment groups and 3 control groups

Statistic	***Treat1***	***Treat2***	***Treat3***	***Ctrl1***	***Ctrl2***	***Ctrl3***
N			116			
n	15	18	23	21	16	23
M	85.880	83.700	81.750	75.870	75	79.710
SD	8.620	11.250	9.230	8.620	8.340	10.150
F Value			3.695			

Table 8. Statistical data from Ljajko's (2016) study

N	Treat.N	Ctrl.N	T-test value
233	139	94	2.710

Abbreviations. Treat1, treatment group 1; Treat2, treatment group 2; Treat3, treatment group 3;Ctrl1, control group 1, Ctrl2, control group 2, Ctrl3, control group 3.

The calculated effects are introduced third. From the data in tables 5, 6, 7 and 8, thirty (30) effect sizes were obtained. They are shown in table 9 along with their confidence intervals, the width of such intervals and the achieved statistical power. The scatter plots of Figures 7, 8 and 9 show the relationships between the calculated effects and their respective sample sizes, confidence interval widths and statistical power, along with a moving average trend line (red dotted line) based on 4 data points.

Table 9. Effect sizes, standard errors, variances, confidence intervals and achieved power

STUDY OR EVALUATION	d	STD.ERR	VAR	95% CI	CI WIDTH	POWER
Tarmizin & Tajudin (2006)	-0.015	0.320	0.102	-0.643, 0.612	1.255	0.050
Javed (2008)	-0.374	0.522	0.272	-1.397,0.650	2.047	0.108
Velez Caraballo (2008)	0.230	0.208	0.043	-0.178,0.638	0.816	0.195
Gantz (2010)	0.382	0.360	0.130	-0.322, 1.087	1.409	0.182
El-Jiryes (2011)	0.036	0.286	0.082	-0.525,0.596	1.121	0.052
Jiang & White (2012)a	0.446	0.090	0.008	0.269,0.623	0.354	0.999
Jiang & White (2012)b	0.125	0.104	0.011	-0.080,0.330	0.410	0.223
Jiang & White (2012)c	0.126	0.300	0.090	-0.463,0.714	1.177	0.070
Curri (2012)	0.284	0.429	0.184	-0.556, 1.124	1.680	0.097

Rieß & Greefrath (2013)	0.281	0.134	0.018	0.019,0.543	0.524	0.557
DeLoach (2013)	0.693	0.283	0.080	0.138,1.248	1.110	0.695
Spencer (2013)	-0.310	0.100	0.010	-0.507, -0.113	0.394	0.872
Drijvers et al. (2014)a	-0.138	0.070	0.005	-0.276, 0.000	0.276	0.624
Drijvers et al. (2014)b	-0.027	0.071	0.005	-0.165,0.112	0.277	0.619
Juan (2015)	0.320	0.171	0.029	-0.014,0.655	0.669	0.465
Jupri, Drijvers, y van den Heuvel-Panhuizen (2015)	0.660	0.130	0.017	0.405, 0.915	0.510	0.999
Pelech(2015)	0.687	0.191	0.036	0.322,1.061	0.739	0.808
Ljajko (2016)	0.361	0.135	0.018	0.097,0 .624	0.527	0.768
Rich (2016)a	-0.617	0.182	0.033	-0.974, -0.261	0.713	0.931
Rich (2016)b	0.155	0.164	0.027	-0.165, 0.476	0.641	0.156
Rich (2016)c	0.070	0.166	0.028	-0.255, 0.396	0.651	0.070
Rich (2016)d	0.314	0.175	0.031	-0.030,0.657	0.687	0.432
Rich (2016)e	-0.022	0.171	0.029	-0.358, 0.314	0.672	0.052
Rich (2016)f	0.363	0.170	0.029	0.029,0.697	0.668	0.569
Marsh (2016)a	-0.280	0.322	0.104	-0.910,0.351	1.261	0.136
Marsh (2016)b	0.600	0.327	0.107	-0.042,1.242	1.284	0.446
Kostić et al. (2016)	1.008	0.224	0.050	0.569, 1.447	0.878	0.999

Mainali & Heck (2017)a	1.181	0.425	0.181	0.348, 2.014	1.666	0.824
Mainali & Heck (2017)b	0.996	0.331	0.110	0.347, 1.645	1.298	0.875
Ocal (2017)	0.586	0.278	0.077	0.042, 1.130	1.088	0.562

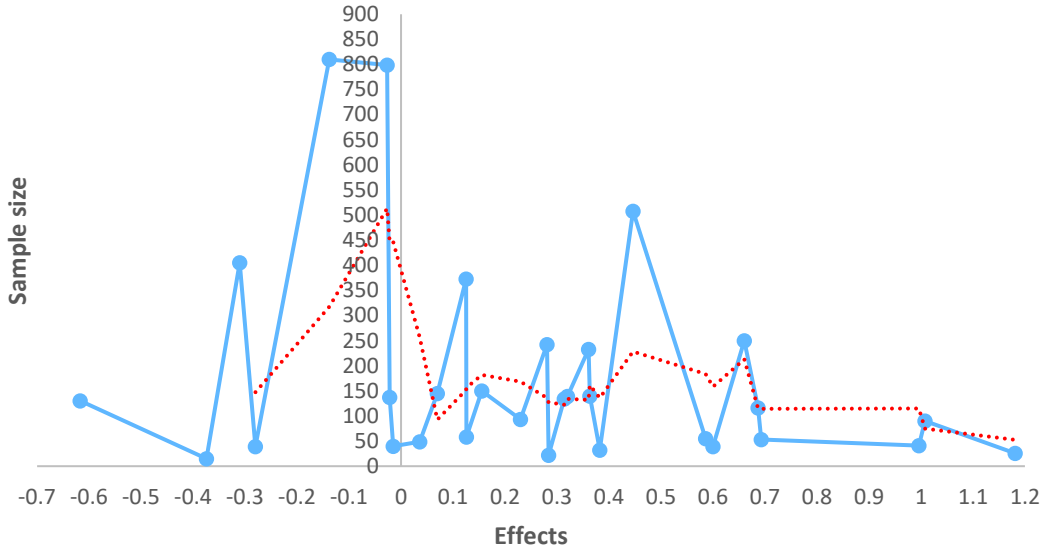

Figure 7. Scatter plot for the effect-sample size data

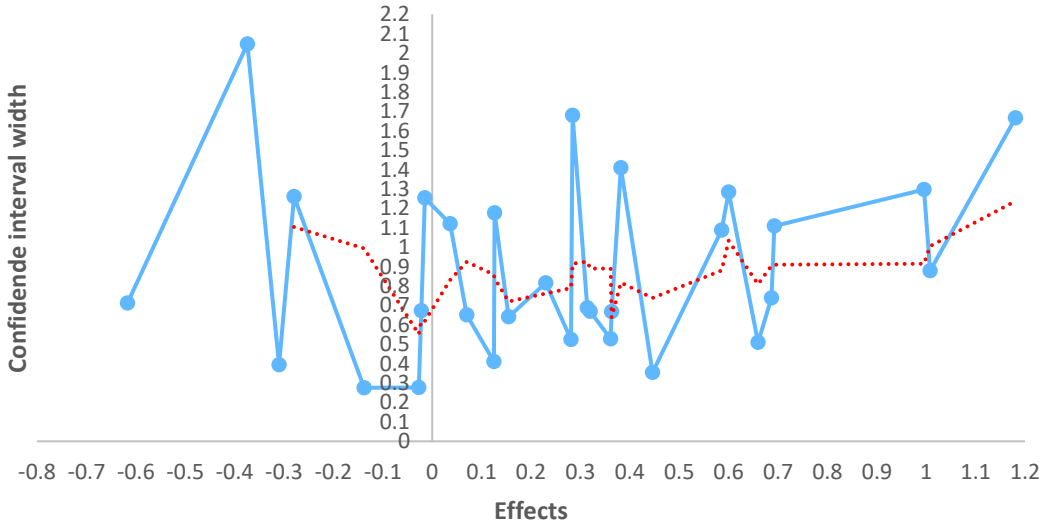

Figure 8. Scatter plot for the effect-confidence interval width data

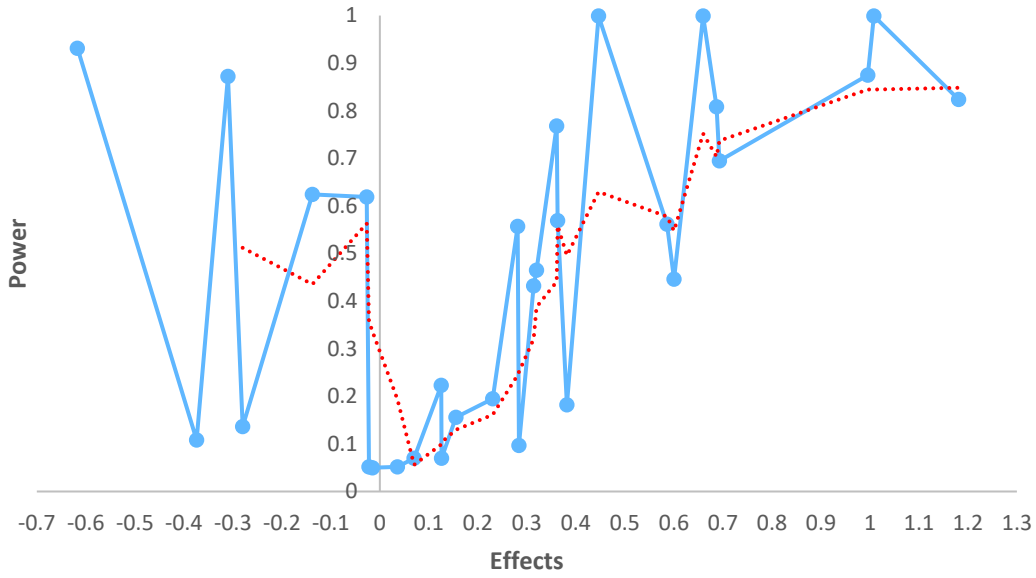

Figure 9. Scatter plot for the effect-power data

The evaluation of the reliability of each effect is presented fourth. Table 10 shows this assessment according to the criteria established previously. Figure 10 summarizes this data showing the percentage of effects that do comply with each unreliability parameter.

Table 9. Reliability evaluation of effect sizes.

STUDY OR EVALUATION	CI.IN.0	CI.WIDTH>d	POWER<.8	IMPRECISE	RELIABLE
Tarmizin & Tajudin (2006)	YES	YES	YES	YES	NO
Javed (2008)	YES	YES	YES	YES	NO
Velez Caraballo (2008)	YES	YES	YES	YES	NO
Gantz (2010)	YES	YES	YES	YES	NO
El-Jiryes (2011)	YES	YES	YES	YES	NO
Jiang & White (2012)a	NO	NO	NO	NO	YES
Jiang & White (2012)b	YES	YES	YES	YES	NO
Jiang & White (2012)c	YES	YES	YES	YES	NO

Curri (2012)	YES	YES	YES	YES	NO
Rieß & Greefrath (2013)	NO	YES	YES	YES	NO
DeLoach (2013)	NO	YES	YES	YES	NO
Spencer (2013)	NO	YES	NO	YES	NO
Drijvers et al. (2014)a	YES	YES	YES	YES	NO
Drijvers et al. (2014)b	YES	YES	YES	YES	NO
Juan (2015)	YES	YES	YES	YES	NO
Jupri, Drijvers, y van den Heuvel-Panhuizen (2015)	NO	NO	NO	YES	NO
Pelech(2015)	NO	YES	NO	YES	NO
Ljajko (2016)	NO	YES	YES	YES	NO
Rich (2016)a	NO	YES	NO	YES	NO
Rich (2016)b	YES	YES	YES	YES	NO
Rich (2016)c	YES	YES	YES	YES	NO
Rich (2016)d	YES	YES	YES	YES	NO
Rich (2016)e	YES	YES	YES	YES	NO
Rich (2016)f	NO	YES	YES	YES	NO
Marsh (2016)a	YES	YES	YES	YES	NO
Marsh (2016)b	YES	YES	YES	YES	NO

Kostić et al. (2016)	NO	NO	NO	YES	NO
Mainali & Heck (2017)a	NO	YES	NO	YES	NO
Mainali & Heck (2017)b	NO	YES	NO	YES	NO
Ocal (2017)	NO	YES	YES	YES	NO

Abbreviations. CI.IN.O, CI includes zero; CI.WIDTH>d, CI width greater than d.

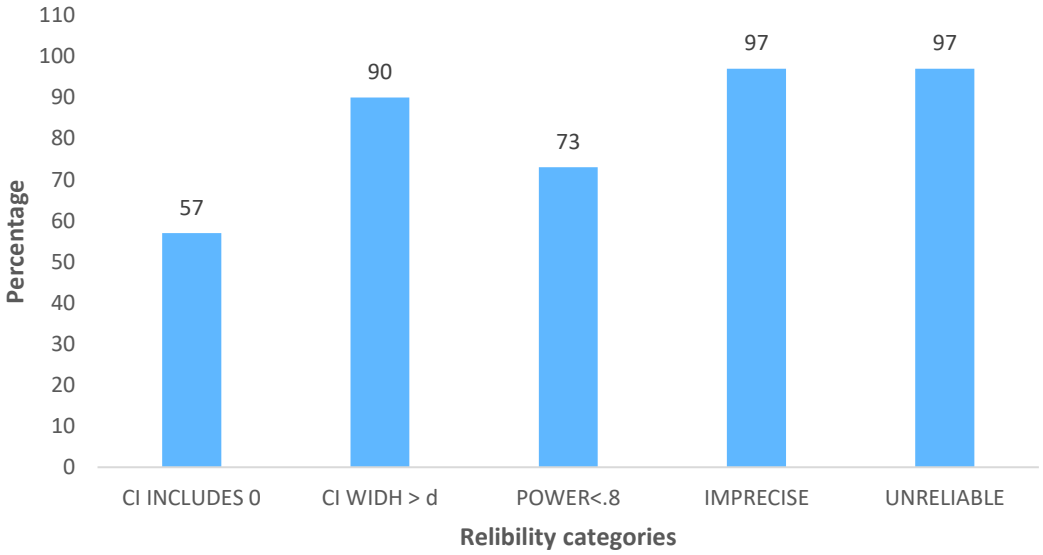

Figure 10. Frequency distribution of effect that do not comply with a reliability parameter

The results of the meta-analysis are presented in Table 12. The values of several statistics are shown along with their confidence intervals. Additionally, for the average effect, its prediction interval, as well as the probability that an effect in a replication study is greater or equal than 0.400 and less than 0.400 are presented. Figures 11 and 12 show the Baujat and forest plots for the meta-analysis. Notice that in figure 12, the red diamond is the average ES and the dotted line represents the prediction interval.

Table 9. Meta-analysis results and other statistics

Statistic	Value	CI.LB	CI.UB	PI.LB	PI.UB	P(d≥0.4)	P(d<.4)
Average effect	0.248	0.108	0.387	-0.427	0.923	32%	68%
τ^2	0.104	0.051	0.240	NA	NA	NA	NA
I^2	81.121	67.786	90.849	NA	NA	NA	NA

Q	143.144**	NA	NA	NA	NA	NA	NA
Average sample	138	88	188	NA	NA	NA	NA
Average CI width	0.893	0.724	1.062	NA	NA	NA	NA
Average power	0.481	0.354	0.608	NA	NA	NA	NA

**p<0.001

Abbreviations. CI.LB, CI lower bound; CI.UB, CI upper bound;PI.LB, prediction interval lower bound; PI.UB; prediction interval upper bound; P(d≥0.4), probability d≥0.4; P(d<0.4), probability d<0.4.

Note. For the average sample size, Iglewicz and Hoaglin's robust test for multiple outliers was used, detecting two outliers, 799 and 810, corresponding to the samples reported by Drijvers et al. (2014) which were excluded.

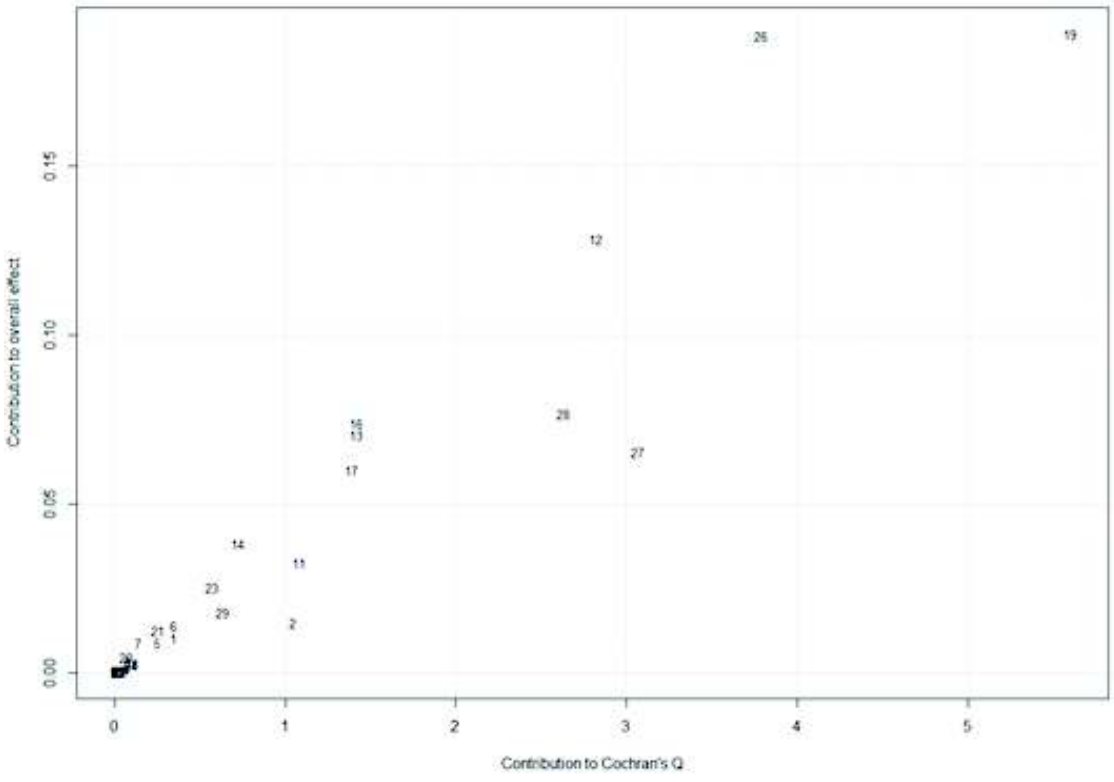

Figure 11. Baujat plot for the meta-analysis

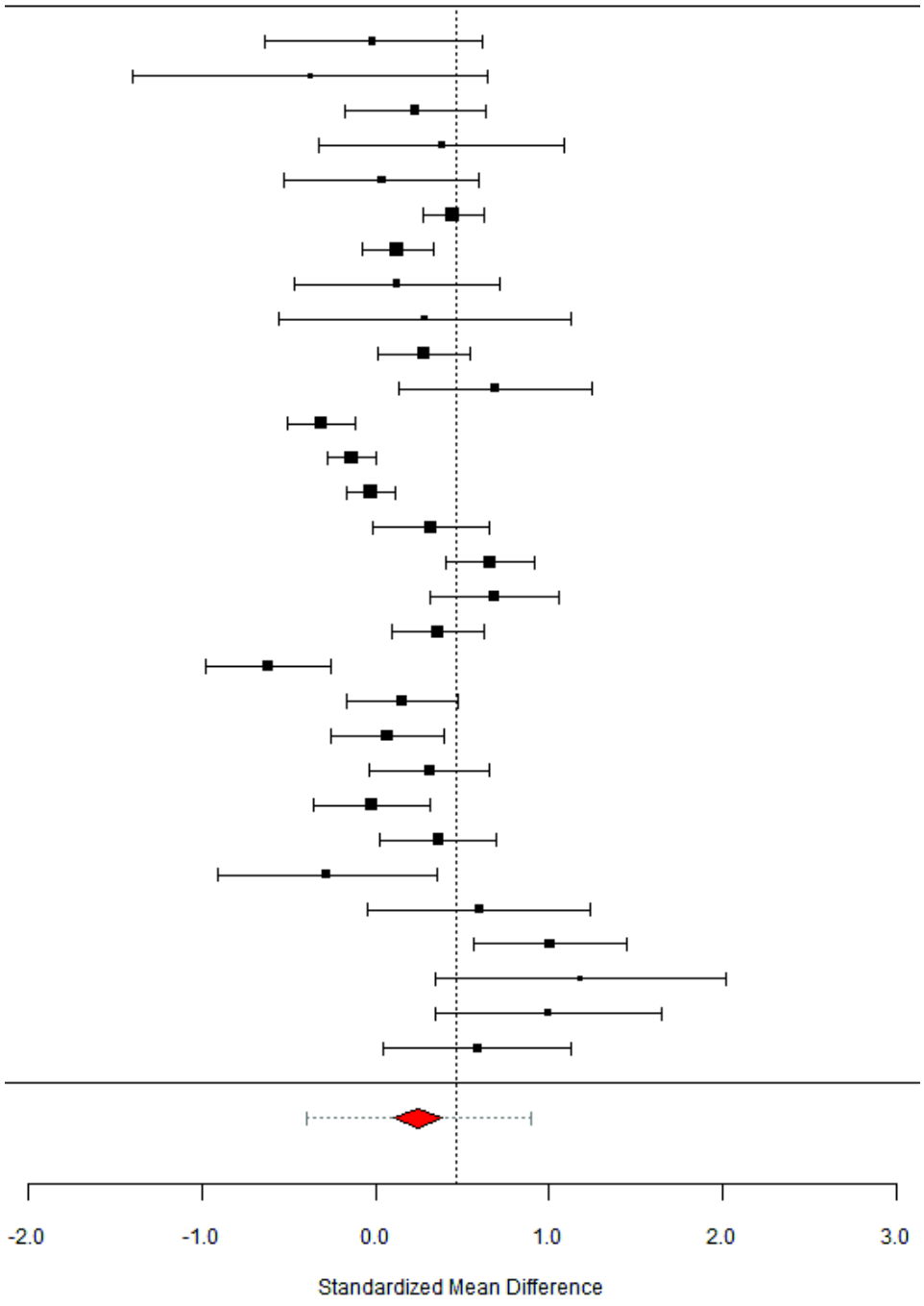

Figure 12. Forest plot for the meta-analysis

The comparison between sample size methods based on statistical power and those based on confidence interval width is shown in figures 13, 14, 15 and 16. These figures depict the impact sample size has on confidence interval width and statistical power. However, since some of the power-based methods produced sample sizes of 80,000 and greater, the maximum sample size was set at 60,000 for graphing purposes. The abbreviations used in the figures are the following: Safeguard.N, sample calculated with the Safeguard method; SSR.N, sample computed with the SSR function; Safeguard CI Width, CI width for the sample obtained with the Safeguard approach; SSR CI Width, CI width for the sample generated through the SSR method; Safeguard Power; statistical power achieved through the sample computed with the Safeguard method; SSR Power, statistical power reached using the sample from the SSR function; USF.N, sample generated through the *userfriendly* method; AIPE.N, sample obtained

using the AIPE approach; USF CI Width, CI width for the sample computed through the *userfriendly* approach; AIPE CI Width, CI width for the sample obtained with the AIPE method; USF Power, statistical power reached with the sample from the *userfriendly* approach; AIPE Power, statistical power achieved with the sample from the AIPE method.

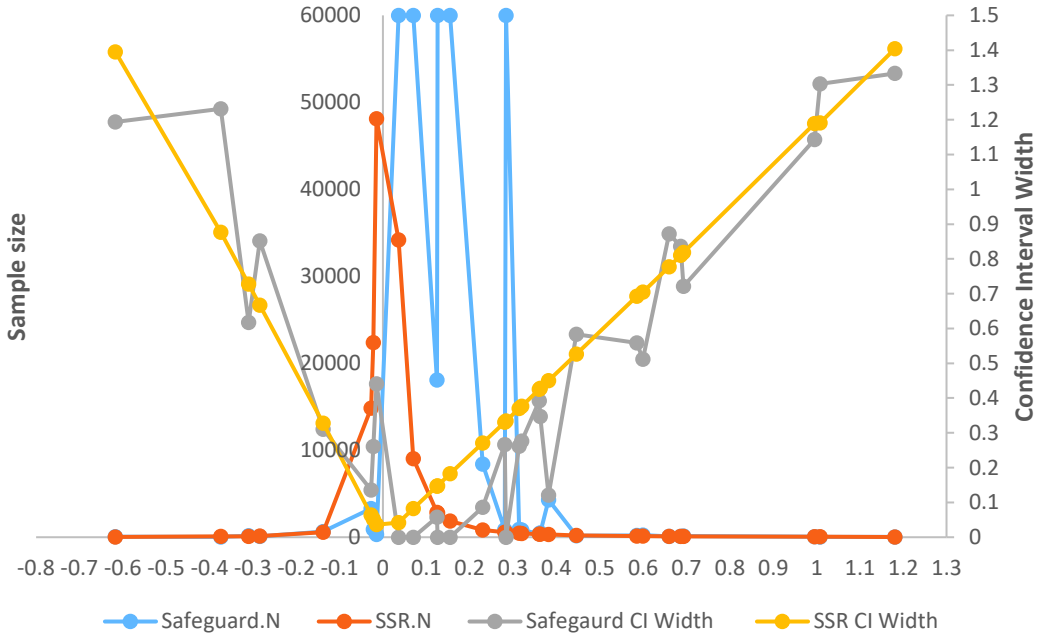

Figure 12. Impact of sample size on interval width for methods based on statistical power

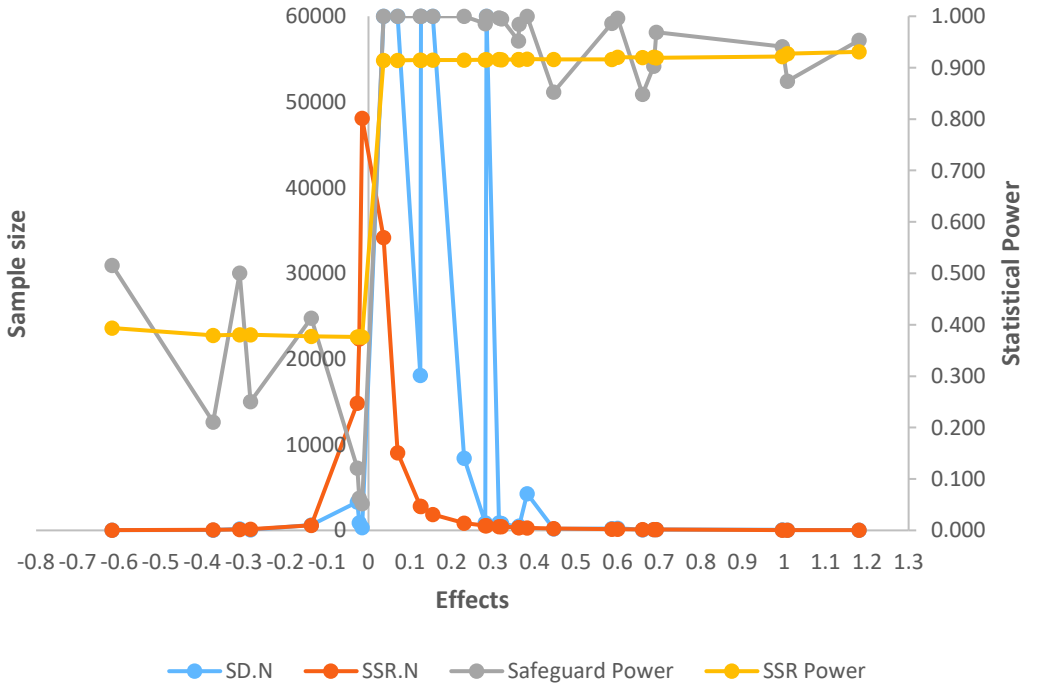

Figure 12. Impact of sample size on statistical power for methods based on statistical power

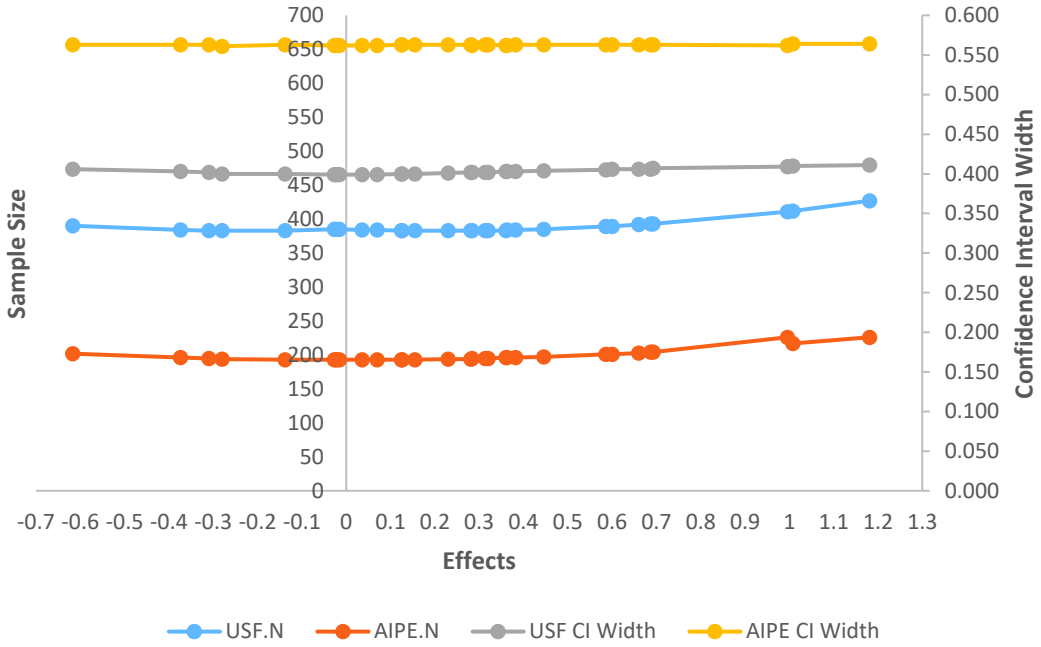

Figure 13. Impact of sample size on interval width for methods based on interval width

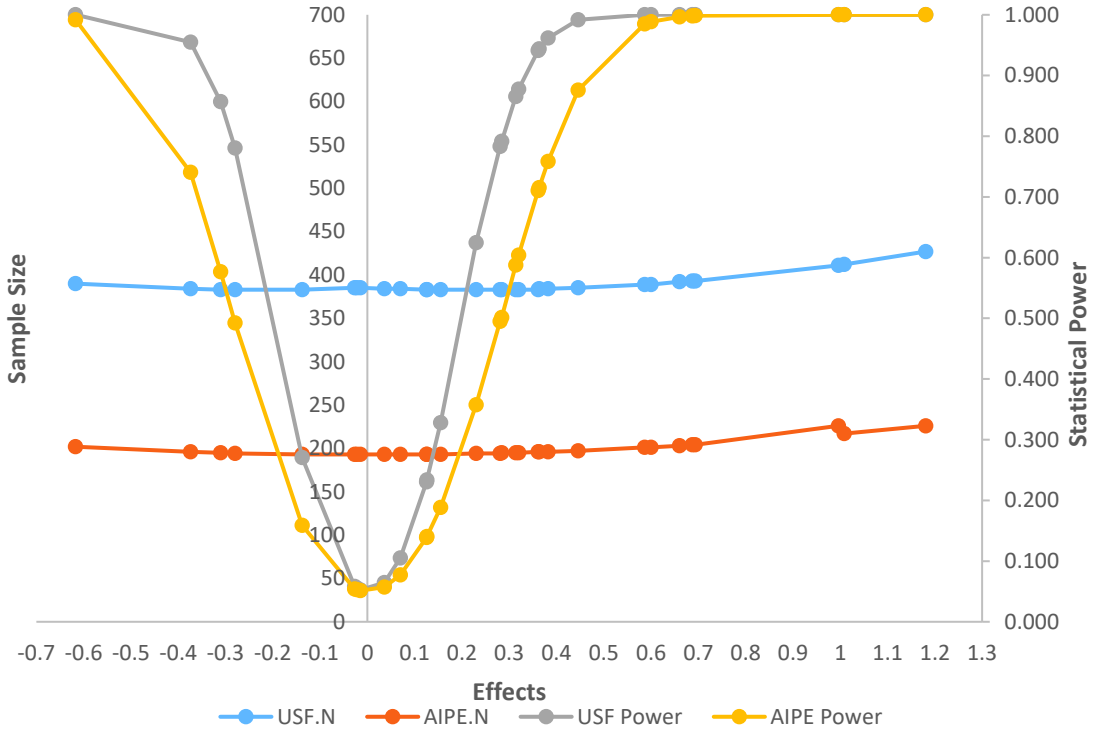

Figure 14. Impact of sample size on statistical power for methods based on interval width.

6. FINDINGS AND DISCUSSION

From the systematic review of the literature and the meta-analysis, four main findings can be reported. First, research that addresses the teaching and learning mathematics using the Instrumental Approach is unfocused and with only a small number of studies that measure the impact of interventions on academic performance. From 113 articles analyzed, only 20 that is 18%, presented enough data to calculate effect sizes with respect to academic performance. This percentage concurs with the data provided by Mayer (2014) in the area of computer games for learning. In particular, the literature downloaded but excluded from the analysis indicates that most of the research citing the instrumental approach is composed of articles reporting: a) phenomenological studies; b) uncontrolled observational studies and c) survey-based studies. Hence, the low number of papers conducting research on academic achievement is not surprising.

Algebra is the preferred research subject (60%) followed by Geometry (20%). This trend is reflected in the type of technology used, where graphing calculators are mainly used (40%), followed by dynamic geometry software (35%). However, there is a considerable percentage of studies (25%) that use other technology types but only once in the analyzed period. A remarkable finding is the fact that 70% of studies do not report any instructional method. Cooperative Learning appears in two studies (10%) and four interventions (20%) report methods that are used once in between 2001 and 2017.

This last point is very important from a psychological perspective. There is convincing evidence that instructional methods are the environmental factors that have the most influence on learning (Clark et al., 2010, Rosenshine, 2009). Also, because decisions about how to teach always reflect an underlying conception of how people learn, even if the teaching strategy, or the learning theory on which it is based, are not explicitly mentioned or described (Mayer, 2009, p.60). Consequently, it is difficult to determine, for a positive effect of sufficient magnitude, if the learning gains are due to the teaching method or to the technology used. It seems that the interventions that use the Instrumental Approach follow a technology-centered approach (Mayer, 2009, p. 10), where the focus is on using technology and force students to adapt themselves to it, rather than adapting technology to fit their needs.

Second, the studies included in the final meta-analysis show serious methodological deficiencies. None of them is a randomized controlled trial. All have a quasi-experimental design. The lack of planning in most of them is notorious, and negatively impacts their accuracy, reliability and replicability. Figure 7 shows that the researchers work with the samples they have available and, that on average these tend to be less than 150 participants (M = 138, SD = 126, 95% CI [88, 188]). This is not surprising since 45% of the effects have samples less than 100. Therefore, the confidence intervals of the effects are embarrassingly long (M = 0.893, SD = 0.453, 95% CI [0.724, 1.062]), making replication highly improbable, and the statistical power is low, indicating that if there were 100 non-null effects to be discovered, the typical study in the area could only discover 48 of them (M = 0.481, SD = 0.341, 95% CI [0.354, 0.608]).

The application of the criteria to evaluate the reliability and precision of each effect confirms the above results (see figure 10). Fifty-five percent (55%)of the confidence intervals contain the null, the lengths of the intervals are greater than the estimated effect 90% of the time and 73% of the effects have a statistical power less than 0.80. Therefore, 97% of the effects are neither precise nor reliable. Only one effect (3%), the one reported in in Jiang & White (2012)a, is reliable because its sample is abnormal with respect to the observed mean: 508 participants.

Third, the results of the meta-analysis corroborate the findings described in points one

and two. The p-value of the Cochran Q test is very low ($p < 0.0001$) and the result for the I^2 statistic of very high (81.121) which indicates a considerable variability in the distribution of effects (Guyatt et al., 2011). The forest plot in Figure 11 supports these results by showing the low level of overlapping between the confidence intervals of the effects (Higgins, 2008), although the Baujat plot (figure 11) indicates that only two effects contribute highly to the level of heterogeneity. The data indicates that the high level of inconsistency is mainly due to: a) the variety of samples sizes between effects (see figure 7); b) the diversity of unknown instructional methods used in the studies (see figure 6) and c) the disparateness of the effects obtained (see table 8 and figure 6). A variable that was not considered in the analysis, and that may impact the level of inconsistency, is the duration of the interventions. However, in view of the instability of the other variables, it is plausible to conclude that it will also be very heterogeneous.

The average effect obtained by the random effects model is $d = 0.248$, $SE = 0.071$, 95% CI [0.108, 0.387] which is considerably below the zone of desired effects. The 95% prediction interval has a full width of 1.350 [-0.427, 0.923] which reinforces the conclusions derived from the Q and I^2 tests and the forest plot. Sixty-eight percent (68%) of future studies will have effects less than 0.400, and additional calculations show that 40% of will be below the teacher effects zone and can be potentially harmful.

Fourth, the analysis of the methods to calculate samples based on statistical power and the length of the confidence interval indicate that the use of power-based methods seems to be generally not recommended (see figures 11 and 12). The procedures tested generated narrow confidence intervals for effects in the interval [0.000, 0.350]. Nonetheless, the sample sizes produced were sometimes extremely large. Furthermore, once the effect was bigger than 0.350, the sample sizes obtained through these methods produced wider and wider confidence intervals reaching 1.404 standard deviations. An unexpected finding was that, although the MBESS package method is based on the same principle as the one in *userfriendly* one, the former yields very different results to the latter, in terms of sample size and confidence interval width. However, it is outside the scope of this work to determine the reason for this difference.

The method provided by the *userfriendlyscience* package (Peters & Crutzen) is the one that behaves the best in terms of sample size, interval length, statistical power and the range of effects it covers. Figures 13 and 14 show, that a sample of $N = 400$ consistently produces intervals with a length of 0.400 standard deviations and a statistical power greater than or equal to 0.800, for effects in the interval [0.280, 1.00]. Effects in the interval [0.160, 0.270] require sample sizes between 1200 and 450 to achieve enough precision and statistical power.

In summary, although research in the teaching and learning of mathematics based on the instrumental approach, is not scarce, the sample of studies for the period 2001-2017 indicates that little progress has been made in terms of improving the learning of mathematics with respect to academic performance. The average effect size is well below the zone of desired effects and almost 70% of future studies will be below it. Since instructional methods are scarcely reported and used, it seems that researchers using the Instrumental Approach believe that technology by itself will lead to better performance. However, as research in Educational Psychology has consistently shown, students, specially novice ones, need a lot of guidance even at the metacognitive level (Harris, Santangelo, & Graham, 2010; Kirschner, Sweller, & Clark, 2006; Rosenshine, 2009) and this guidance must be provided through an instructional method. Furthermore, technology needs to be adapted to support the way people learn and not the other way around (Mayer, 2009; Sweller, 2012).

Finally, the precision and reliability of studies needs to be improved. However, this requires the use of sample sizes larger than those commonly found in the literature (LeBel,

Campbell, and Loving, 2017; Peters & Crutzen, 2017; Tversky and Kahneman, 1971). If the population parameter is unknown, a sample size of N=1200, which is equivalent to calculating a sample that will give a ±0.113 confidence interval with the *userfriendlyscience* method, is sufficient to produce reliable results within the teacher effects zone and beyond the excellence threshold of d=0.600 set by Hattie. It can be argued that the use of samples such as those recommended here is costly, reduces the proportion of new findings as well as the progress in any area (Fiedler and Schwarz, 2016, Finkel et al., 2015). However, as LeBel, Campbell, and Loving (2017) point out, the benefits outweigh the costs because to increase the frequency and proportion of true discoveries (i.e., to be able to distinguish true from false hypotheses) it is necessary to reduce the rate of Type I and II errors. However, it must be underlined that the samples computed in this research were calculated for effects represented by Cohen's *d*, and a two-independent-groups design. It is possible that other effects and designs will require different methods producing distinct sample sizes.

7. REFERENCES

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., ... & Perugini, M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108-119.

Bailey, T.M. (2009). Effect Size Calculator Converter (Version 605) [MS Excel workbook]. Downloadable from URL http://psych.cf.ac.uk/home2/mat/

Begley, C. G., & Ioannidis, J. P. (2015). Reproducibility in science. *Circulation research*, *116*(1), 116-126.

Bloom, B. S. (1972). Innocence in education. *The School Review*, *80*(3), 333-352.

Bogg, T., & Lasecki, L. (2014). Reliable gains? Evidence for substantially underpowered designs in studies of working memory training transfer to fluid intelligence. *Frontiers in psychology*, *5*.

Brand, A., & Bradley, M. T. (2016). The Precision of Effect Size Estimation From Published Psychological Research: Surveying Confidence Intervals. *Psychological Reports*, *118*(1), 154-170.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365-376.

Chen, X. (2010). Google Scholar's dramatic coverage improvement five years after debut. *Serials review*, *36*(4), 221-226.

Chiolero, A., Santschi, V., Burnand, B., Platt, R. W., & Paradis, G. (2012). Meta-analyses: with confidence or prediction intervals? *European journal of epidemiology*, *27*(10), 823-825.

Christley, R. M. (2010). Power and error: increased risk of false positive results in underpowered studies. *The Open Epidemiology Journal*, *3*(1).

Clark, R. E., Yates, K., Early, S., Moulton, K., Silber, K. H., & Foshay, R. (2010). An analysis of the failure of electronic media and discovery-based learning: Evidence for the performance benefits of guided training methods. *Handbook of training and improving workplace performance*, *1*, 263-297.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

Curri, E. (2012). *Using computer technology in teaching and learning mathematics in an Albanian upper secondary school: the implementation of simReal in trigonometry lessons* (Master's thesis, Universitetet i Agder; University of Agder).

De Bruyckere, P., Kirschner, P. A., & Hulshof, C. D. (2015). *Urban Myths about Learning and Education*. Academic Press.

De Winter, J. C., Zadpoor, A. A., & Dodou, D. (2014). The expansion of Google Scholar versus Web of Science: a longitudinal study. *Scientometrics*, *98*(2), 1547-1565.

DeLoach, M. (2013). *The impact of graphing calculators on high school students' performance on a standardized mathematics test* (Doctoral dissertation, University of Phoenix).

Drijvers, P., Doorman, M., Kirschner, P., Hoogveld, B., & Boon, P. (2014). The Effect of Online Tasks for Algebra on Student Achievement in Grade 8. *Technology, Knowledge and Learning*, *19*(1-2), 1-18.

Elgamal, A. F., Abas, H. A., & Baladoh, E. S. (2013). An interactive e-learning system for improving

web programming skills. *Education and Information Technologies*, *18*(1), 29-46.
El-Jiryes, H.A. (2011). *Effect of Technology Integration in Teaching Quadratic Functions on Lebanese Students' Learning, Problem-Solving Abilities, and Attitudes* (Doctoral dissertation, LEBANESE AMERICAN UNIVERSITY).
Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PloS one*, *5*(4), e10068.
Fanelli, D. (2011). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891-904.
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*(4), 1149-1160.
Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, *7*(1), 45-52.
Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2015). Best research practices in psychology: Illustrating epistemological and pragmatic considerations with the case of relationship science. *Journal of personality and social psychology*, *108*(2), 275.
Flint, J., Cuijpers, P., Horder, J., Koole, S. L., & Munafò, M. R. (2015). Is there an excess of significant findings in published studies of psychotherapy for depression? *Psychological medicine*, *45*(2), 439-446.
Forstmeier, W., Wagenmakers, E. J., & Parker, T. H. (2016). Detecting and avoiding likely false-positive findings–a practical guide. *Biological Reviews*
Gantz, L. A. G. (2010). *Handheld computer algebra systems in the pre-algebra classroom*. George Mason University.
Gentleman, R., Huber, W., & Carey, V. J. (2011). R language. In *International Encyclopedia of Statistical Science* (pp. 1159-1161). Springer Berlin Heidelberg.
Glass, G. V., Smith, M. L., & McGaw, B. (1981). *Meta-analysis in social research*. Sage Publications, Incorporated.
Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean? *Science translational medicine*, *8*(341), 341ps12-341ps12.
Guyatt, G. H., Oxman, A. D., Kunz, R., Woodcock, J., Brozek, J., Helfand, M., ... & Norris, S. (2011). GRADE guidelines: 7. Rating the quality of evidence—inconsistency. *Journal of clinical epidemiology*, *64*(12), 1294-1302.
Haddaway, N. R., Collins, A. M., Coughlin, D., & Kirk, S. (2015). The role of Google Scholar in evidence reviews and its applicability to grey literature searching. *PloS one*, *10*(9), e0138237.
Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle P value generates irreproducible results. *Nature methods*, *12*(3), 179-185.
Harris, K. R., Santangelo, T., & Graham, S. (2010). Metacognition and strategies instruction in writing. *Metacognition, strategy use, and instruction*, 226-256.
Harzing, A. W. (2014). A longitudinal study of Google Scholar coverage between 2012 and 2013. *Scientometrics*, *98*(1), 565-575.
Harzing, A. W., & Alakangas, S. (2016). Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison. *Scientometrics*, *106*(2), 787-804.
Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London, UK: Routledge.
Higgins, J. P. (2008). Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. *International journal of epidemiology*, *37*(5), 1158-1160.
Holmes, J. D. (2016). *Great Myths of Education and Learning*. John Wiley & Sons.
Howland, J. L., Wright, T. C., Boughan, R. A., & Roberts, B. C. (2009). How scholarly is google scholar? A comparison to library databases. *College & Research Libraries*, 70(3): 227-234.
IntHout, J., Ioannidis, J. P., Rovers, M. M., & Goeman, J. J. (2016). Plea for routinely presenting prediction intervals in meta-analysis. *BMJ open*, *6*(7), e010247.
Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124.
Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640-648.
Ioannidis, J. P. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, *7*(6), 645-654.
Ioannidis, J. P., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in cognitive sciences*, *18*(5), 235-241.
Ioannidis, J. P., Tarone, R., & McLaughlin, J. K. (2011). The false-positive to false-negative ratio in

epidemiologic studies. *Epidemiology*, *22*(4), 450-456.
Javed, S. H. (2008). *Online facilitated mathematics learning in vocational education: a design-based study* (Doctoral dissertation, Victoria University).
Jiang, Z. & White, A. (2012). An Efficacy Study on the Use of Dynamic Geometry Software. *In the Proceedings of the 12th International Congress on Mathematical Education*.
Juan, K. (2015). *Effects of interactive software on student achievement and engagement in four secondary school geometry classes, compared to two classes with no technology integration* (Doctoral dissertation, University of Florida).
Jupri, A., Drijvers, P., & van den Heuvel-Panhuizen, M. (2015). Improving Grade 7 Students' Achievement in Initial Algebra Through a Technology-Based Intervention. *Digital Experiences in Mathematics Education*, *1*(1), 28-58.
Keen, H. I., Pile, K., & Hill, C. L. (2005). The prevalence of underpowered randomized clinical trials in rheumatology. *The Journal of rheumatology*, *32*(11), 2083-2088.
Kelley, K. (2007). Methods for the behavioral, educational, and social sciences: An R package. *Behavior Research Methods*, *39*(4), 979-984.
Kepes, S., & McDaniel, M. A. (2013). How trustworthy is the scientific literature in industrial and organizational psychology? *Industrial and Organizational Psychology*, *6*(3), 252-268.
Kirschner, P. A., & van Merriënboer, J. J. (2013). Do learners really know best? Urban legends in education. *Educational Psychologist*, *48*(3), 169-183.
Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational psychologist*, *41*(2), 75-86.
Koretz, R. L., & Lipman, T. O. (2017). Understanding systematic reviews and meta-analyses. *Journal of Parenteral and Enteral Nutrition*, *41*(3), 316-323.
Kostić, V. D., Jovanović, V. S., Sekulić, T. M., & Takači, D. B. (2016). Visualization of problem solving related to the quantitative composition of solutions in the dynamic GeoGebra environment. *Chemistry Education Research and Practice*, *17*(1), 120-138.
Kowalski, T. (2009). Need to address evidence-based practice in educational administration. *Educational Administration Quarterly*, *45*(3), 351-374.
Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology*, *4*, 863. doi:10.3389/fpsyg.2013.00863
LeBel, E. P., Campbell, L., & Loving, T. J. (2017). Benefits of open and high-powered research outweigh costs. *Journal of personality and social psychology*, *113*(2), 230.
Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Sage Publications, Inc.
Ljajko, E. (2016). Does the Problem Complexity Impact Students' Achievements in a Computer Aided Mathematics Instruction? *Teaching of Mathematics*, *19*(1).
Lohse, K., Buchanan, T., & Miller, M. (2016). Underpowered and overworked: Problems with data analysis in motor learning studies. *Journal of Motor Learning and Development*, *4*(1), 37-58.
Mainali, B. R., & Heck, A. (2017). Comparison of traditional instruction on reflection and rotation in a Nepalese high school with an ICT-rich, student-centered, investigative approach. *International Journal of Science and Mathematics Education*, *15*(3), 487-507.
Marsh, D. L. (2016). *Using manipulatives to investigate ESOL students' achievement and dispositions in algebra* (Doctoral dissertation). Kennesaw State University, Kennesaw, GA.
Marszalek, J. M., Barber, C., Kohlhart, J., & Cooper, B. H. (2011). Sample size in psychological research over the past 30 years. *Perceptual and motor skills*, *112*(2), 331-348.
Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychological methods*, *9*(2), 147.
Mayer, R. E. (2009). *Multimedia learning* (2nd edt.). New York: Cambridge University Press.
Mayer, R. E. (2014). *Computer games for learning: An evidence-based approach*. MIT Press.
Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., ... & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021.
Nourbakhsh, E., Nugent, R., Wang, H., Cevik, C., & Nugent, K. (2012). Medical literature searches: a comparison of PubMed and Google Scholar. *Health Information & Libraries Journal*, *29*(3), 214-222.
Ocal, M. F. (2017). The Effect of Geogebra on Students' Conceptual and Procedural Knowledge: The Case of Applications of Derivative. *Higher Education Studies*, *7*(2), 67.
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.
Pelech, P. A. (2015). *The Effect of the TI-Nspire on Student Achievement in Common Core Algebra*

(Doctoral dissertation, CONCORDIA UNIVERSITY CHICAGO).
Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*(6060), 1226-1227.
Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science*, *9*(3), 319-332.
Peters, G.-J. Y., & Crutzen, R. (2017). *Knowing Exactly How Effective an Intervention, Treatment, or Manipulation is and Ensuring that a Study Replicates: Accuracy in Parameter Estimation as a Partial Solution to the Replication Crisis*. Available online at: http://osf.io/preprints/psyarxiv/cjsk2
Quintana, D. S. (2015). From pre-registration to publication: a non-technical primer for conducting a meta-analysis to synthesize correlational data. *Frontiers in psychology*, *6*, 1549.
Ravindran, V., & Shankar, S. (2015). Systematic reviews and meta-analysis demystified. *Indian Journal of Rheumatology*, *10*(2), 89-94.
Rich, M. J. (2016). *The relationship between using Study Island supplemental math software and third, fourth and fifth grade students' mathematics achievement*. Liberty University.
Rieß, M., & Greefrath, G. (2013). Results on the function concept of lower achieving students using handheld cas-calculators in a long-term Study. In *Proceedings of the eighth Congress of the European Society for Research in Mathematics Education*.
Rosenshine, B. (2009). The empirical support for direct instruction. In S. Tobias & T. M. Duffy (Eds.), *Constructivist instruction: Success or failure?* (pp. 201-220). New York: Routledge.
Sagan, C. (1997). *The demon-haunted world: Science as a candle in the dark*. Random House Digital, Inc.
Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, *22*(11), 1359-1366.
Slavin, R. E. (2008)a. Cooperative learning, success for all, and evidence-based reform in education. *Éducation et didactique*, *2*(2), 149-157.
Slavin, R. E. (2008)b. Perspectives on evidence-based research in education—What works? Issues in synthesizing educational program evaluations. *Educational researcher*, *37*(1), 5-14.
Slavin, R. E. (2011). Cooperative learning. *Learning and cognition in education*, 160-166.
Slavin, R. E. (2017). Evidence-based reform in education. *Journal of Education for Students Placed at Risk (JESPAR)*, *22*(3), 178-184.
Spencer, M. (2013). *The Influence of Using TI-84 Calculators with Programs on Algebra I High Stakes Examinations*. Delta State University.
Stodden, V., McNutt, M., Bailey, D. H., Deelman, E., Gil, Y., Hanson, B., ... & Taufer, M. (2016). Enhancing reproducibility for computational methods. *Science*, *354*(6317), 1240-1241.
Sweller, J. (2012). Human cognitive architecture: Why some instructional procedures work and others do not. In K. Harris, S. Graham, & T. Urdan (Eds.), *APA Educational Psychology Handbook* (Vol. 1). Washington, DC: American Psychological Association.
Tarmizin, R.A. & Tajudin, N. M. (2006). Using Graphic Calculator in Teaching and Learning Mathematics: Effects on Students' Achievement and Meta-Cognitive Skills. *Proceedings of the International Symposium on Technology and its Integration into Mathematics Education*. Dresden, Germany.
Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological bulletin*, *76*(2), 105.
Vadillo, M. A., Konstantinidis, E., & Shanks, D. R. (2016). Underpowered samples, false negatives, and unconscious learning. *Psychonomic Bulletin & Review*, *23*(1), 87-102.
Velez Caraballo, Y. (2008). *The use of technology and cooperative learning in the achievement of college students in the concept of* functions and their attitude towards mathematics* (Doctoral dissertation, University of Puerto Rico, Rio Piedras (Puerto Rico)).
Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., ... & Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research synthesis methods*, *7*(1), 55-79.