Assessing Heterogeneity of Effect Size on Sample Size in a Meta- Analysis of Validity Studies

Adeyemo, Emily Oluseyi

Adediwura, Alaba Adeyemi

Department of Educational Foundations and Counselling,
Faculty of Education Obafemi Awolowo University. Ile-Ife
Email seyiadeyemo2007@yahoo.com Email: Yemtoy20002000@yahoo.com

Doi: 10.5901/mjss.2012.03.01.219

Abstract *This study examined the impact of heterogeneity of effect size on the sample size of some validity studies. Thirty (30) empirical studies were selected on the basis of empirical status and relevance; their results both quantitative and qualitative were recorded, coded and analyzed. The findings revealed that the differences in the results of the 30 selected empirical studies were a function of the sample size on which the studies were based. Mean Fisher (Zr =0.393 , WZr=0.42347). Characteristics peculiar to each study did not affect the result of the study. Weighted Mean Fisher by sample size (WZr) =0.398 with associated r=0.375) was an equivalent of the Mean Fisher. The results of the empirical studies were found to be significantly different in terms of their effect sizes (÷2=1444.97 p$^{< 0.05, 29df}$). The heterogeneity of the effect size was a result of the heterogeneity of the sample size. The Weighted Mean Fisher was bigger than the Mean Fisher (WZr=0.489 ,Zr=0.393). The difference in the result was a function of differences in the sample size and not as a result of the study characteristics peculiar to each study.*

Key words: *meta - analysis, effect size, heterogeneity, Sample size.*

1. Introduction

Meta-analysis is a form of research synthesis in which empirical results of previous studies are re-analyzed. It is the statistical analysis of a large collection of analysis from individual studies for the purpose of integrating the findings (Adeyemo,2007). Meta-analysis is a set of statistical procedure designed to accumulate experimental and correlation results across independent studies that address a related set of research questions (Glass 1976. P.3). Meta-analysis uses the summary statistics from individual studies as the data point, unlike the traditional research methods. The key assumption of this analysis is that each study provides a different estimate of the underlying relationship within the group. In the process of accumulating results across studies, one can obtain a more accurate representation of the population relationship that is provided by the individual study estimators.

Some researchers viewed meta-analysis as quantitative literature review (Stanley, 2001) while some asserted that meta-analysis could be used to highlight aspect critical to future development of theory (Rosenthal and Dimatteo, 2001, Golafarb1995). but whatever their views, the goal of meta analysis involves the provision of accurate, impartial and quantitative description of the findings in a population of studies on a particular topic. Although there had been a wide increase in its application, meta-analysis is still surrounded with various methodological difficulties, among which is the impact of the heterogeneity of effect sizes on the results of a meta-analysis study

Heterogeneity in meta-analysis refers to the variation in study outcomes between studies. Typically, meta-analysis has three main goals:

(i) to test whether studies' results are homogenous
(ii) to obtain a global index about the effect magnitude of the studied relation join to a confidence interval and its statistical significance and

(iii) if there is heterogeneity among studies, to identify possible variables or characteristics moderating the results obtained.

The classical measurement heterogeneity is Cochram's Q which is calculated as the weighted sum of squared difference between individual study effect and the pooled effect across studies, with the weight being those used in the pooling methods. Q is distributed as a chi-square with k-1 degree of freedom where k = number of studies .As a test of heterogeneity, Q has low power when the number of studies is small and too much power if the number of studies in large. (Higgins, et al 2003).

The shortcoming of the Q statistic is that it is not powerful to detect true heterogeneity among studies when the meta-analysis include a small number of studies and also has excessive power to detect negligible variable with a high number of studies (Alexander Scozzaro and Borodlain 1989, Cornwell 1993, Cornwell and Ladd 1993, Hardy and Thompson 1998). Alongside, Q statistics does not provide the information about the extent of true heterogeneity but only its statistical significance, but Huggin and Thompson (2002) provided an alternative which is the use of 1^2 indices as a measure of true heterogeneity.

There are two sources of variability that explain the heterogeneity in a set of studies. One of them is the variability due to sampling error also called 'within study variability'. This is always present in meta-analysis because every single study uses different samples. The other source of variability is the" between study variability" which appear in meta-analysis when there is true heterogeneity among the population effect size estimated by the individual studies. The between studies variability is due to the influence of an undetermined number of characteristics that vary among the studies such as those related to the characteristics of the samples ,variation in the treatment, the design quality etc. (Brockwell and Gordon 2001, Hunter and Schmidt 2000, Feild 2003, National Research Council 1992). In meta-analysis, when the difference in results between studies is greater than would be expected by chance, one needs to investigate whether the observed variation in results across studies is associated with methodology features between studies

Assessing heterogeneity in meta-analysis is a crucial issue because of the presence or absence of true heterogeneity. The between studies variability can affect the statistical model that the meta-analyst decide to apply to the meta-analytic database. When studies' results differ by the sampling error (i.e. homogeneity case) a fixed-effect model can be applied to obtain an average effect size. On the other hand, if the study results differ by more than the sampling error (i.e. heterogeneous case), the meta-analyst can assume a random effect model in other to take into account both within and between studies variability or moderator variable can be searched from a fixed effect model. (Field 2001, 2003, Hedges 1994, Hedges and Olkin 1985, Helges & Vevea 1998, Overton 1998).

Heterogeneity is to be expected in a meta-analysis because the multiple studies that are performed by the different researchers in different places with the different methods cannot but end with different estimating underlying parameters. In meta analysis, when the difference in results between studies is greater than the expected by chance one needs to investigate whether the observed variation in results across studies is associated with methodological features between studies. Hence, to identify causes of heterogeneity, to learn about its robustness and to be able to remove it prior to performing meta-analysis, is the need for evaluating its sensitivity.

This study is designed to assess the impact of heterogeneity of the effect size in a meta-analysis of some empirical validity studies

1.1 Research Questions

(i) What factors contribute to the large amount of variance in the strength of previously reported validity studies?

(ii) What is the extent to which study characteristics peculiar to the studies contribute to the large amount of validity studies.?

1.2 Research Hypothesis

(i) The selected studies are not significantly different in terms of their effect sizes

(ii) The heterogeneity of the effect size is not due to heterogeneity of sample size used on the various research studies.

2. Methodology

The study design is a causal-comparative (i.e. ex-post facto). The sample size consisted of 30 empirical studies with document made up of both published and unpolished articles on validity of UME in Nigeria. The 30 validity studies were purposively selected on the basis of empirical status and relevance, using a computer search and hand searching through the reference of collected papers and other relevant books and journal on the subject. The quantitative results were recorded and converted to common effect size while the qualitative results were recorded and coded. The results were analyzed in line with the works of Glass (1981), Rosenthal (1984) and Rosenthal & Roselow (1984).

3. Results

Research Question (1)*.What factor contributes to the large amount of variance in the strength of previously reported validity studies.*

To identify the factors that contribute to the large amount of variance in the strength of the previously reported validity studies, the mean Fisher $\overline{Zr}$ of the selected empirical studies was compared with the Weighted Fisher W$\overline{Zr}$

Weighted Fisher W$\overline{Zr}$ = $\frac{\Sigma(N-3)(Zr)}{\Sigma(N-1)}$
and

Unweighted Mean Fisher $\overline{Zr}$ = $\frac{\Sigma Zr}{K}$

Table 1. Effect of Sample size on the 'effect size' of the selected study.

Study S/N	Sample size	N-3	r	Zr	(N –3)(Zr)
1	250	247	0.39	0.4118	101.7146
2	558	555	0.32	0.3310.6	184.038
3	300	297	0.21	0.232	63.3204
4	121	118	0.47	0.5101	60.1918
5	40	37	0.04	0.04	1.48
6	800	797	0.28	0.2877	229.2969
7	30	27	0.09	0.0902	2.435
8	1800	1797	0.86	1.1155	2004.5535
9	750	747	0.18	0.182	135.954
10	802	799	0.09	0.0902	72.0698
11	100	97	0.57	0.6475	62.8075
12	123	120	0.29	0.2986	35.832
13	1379	1376	0.61	0.7089	975.4464
14	30	27	0.24	0.2448	6.6096
15	40	37	0.04	0.04	1.48

16	120	117	0.70	0.8673	101.4741
17	687	684	0.21	0.2132	145.8288
18	180	177	0.37	0.3884	68.7468
19	54	51	0.30	0.3095	15.7845
20	860	857	0.48	0.533	456.781
21	180	177	0.31	0.3205	56.7285
22	227	224	0.42	0.4477	100.2848
23	6462	6459	0.43	0.4477	2891.694
24	107	104	0.12	0.1206	12.5424
25	78	75	0.36	0.3769	28.2675
26	60	57	0.62	0.725	41.325
27	159	156	0.36	0.3769	58.7964
28	212	209	0.03	0.03	6.27
29	42	39	0.36	0.3769	14.6991
30	222	219	0.78	1.0454	228.9426
		16683	M. Fisher	0.39307	8165.395
			W. Fisher	0.48994	

From the table, the Mean Fisher $\overline{Zr}$ was 0.393037, with the associated r = 0.375. Weighted Mean Fisher $\overline{WZr}$ = 0.434713 with associated r = 0.410 . The Weighted Mean Fisher was greater than the Mean Fisher. Weighting by sample size resulted to bigger estimate of combined effect size than when sample sizes were not used. The difference in the results of the 30 empirical studies was a result of differences in the sample size on which ‘r’ was based. Thus weighted r = 0.410 was a better measure because it corrected for the diversity of sample used by different researchers.

Research Question 2 What is the extent to which study characteristics peculiar to these studies contribute to the large amount of variance in the strength of previously reported validity studies.

From the quantitative results, twenty characteristic features coded of the empirical study were recorded. The summations were used as weights. The” weights” were ascribed as independent variables because to a certain extent, the results recorded by primary researchers were influenced by those study characteristics. The highest weight for a study was 38 while the lowest was 29. The maximum weight for any study was 47 based on the coded characteristics. The bigger the weight, the more representative the indices on which the primary researcher based the calculation of co-efficient ‘r’

Table 2. Coded Characteristics under Researcher Control and Effect Size r

Study	R	Zr	W	(W)(Zr)
1	0.39	0.4118	29	11.9422
2	0.32	0.3310.6	38	12.600
3	0.21	0.232	27	5.7564
4	0.47	0.5101	26	13.2626
5	0.04	0.04	29	1.16
6	0.28	0.2877	35	10.0695
7	0.09	0.0902	26	2.3452
8	0.86	1.1155	27	30.1185
9	0.18	0.182	34	6.188
10	0.09	0.0902	33	2.9766

11	0.57	0.6475	36	23.31
12	0.29	0.2986	29	8.6594
13	0.61	0.7089	39	27.6471
14	0.24	0.2448	35	8.568
15	0.04	0.04	33	1.32
16	0.70	0.8673	36	31.2228
17	0.21	0.2132	31	6.572
18	0.37	0.3884	37	14.370
19	0.30	0.3095	31	9.5945
20	0.48	0.533	37	19.721
21	0.31	0.3205	31	9.9355
22	0.42	0.4477	37	16.5649
23	0.43	0.4477	38	17.0126
24	0.12	0.1206	32	3.8592
25	0.36	0.3769	33	12.4377
26	0.62	0.725	30	21.75
27	0.36	0.3769	38	14.3222
28	0.03	0.03	30	0.9
29	0.36	0.3769	34	12.8146
30	0.78	1.0454	34	35.5436
	Mean Fisher	0.393037	985	392.5449
	Weighted Fisher	0.398		

Using,

$$\overline{Zr} = \frac{\Sigma(weight)(Zr)}{\Sigma\ weight}$$

Substituting the various weight (W) and Fisher (Zr) from the table, Weighted $\overline{Zr}$ = 0.398 with associated r =.375. Although the value is relatively low, yet it is not bigger than the Mean Fisher $\overline{Zr}$ =0.393 with associated r = 0.375.

The implication of this is that weighting by characteristics under the researcher's control did not contribute to any large amount of variance in the strength of previously reported validity studies.

Hypothesis 1: *The empirical validity studies are not significantly different in terms of their effect sizes.*

The works of Cochram (1967, 1980) were used to assess the statistical heterogeneity of the 30 effect sizes.

$$X^2 = \sum_{j=1}^{k} (Nj-3)(Zrj-\overline{Z})^2$$ is distributed for χ^2 with k – 1 df

Table 3. Computation of Chi-squared Using Correlation Coefficient Effect Size 'r'

Study	Sample size	N-3	r	Zr	$Zr-\overline{Zr}$	$(Zr-\overline{Zr})^2$	(N-3)$(Zr-\overline{Zr})^2$
1	250	247	0.39	0.4118	0.018763	0.000352	0.086956392
2	558	555	0.32	0.3310.6	$\overline{0}$ .061437	0.003775	2.094850258
3	300	297	0.21	0.232	$\overline{0}$ .0179837	0.32341	9.605379931
4	121	118	0.47	0.5101	0.117063	0.013704	1.617042024

5	40	37	0.04	0.04	$\overline{0}$ .353037	0.124635	4.611499565
6	800	797	0.28	0.2877	$\overline{0}$ 105337	0.011096	8.843419204
7	30	27	0.09	0.0902	$\overline{0}$ 302837	0.09171	2.476176711
8	1800	1797	0.86	1.1155	0.722463	0.521953	939.9491571
9	750	747	0.18	0.182	$\overline{0}$ .211037	0.044537	33.26885168
10	802	799	0.09	0.0902	$\overline{0}$ .302837	0.09171	73.27648861
11	100	97	0.57	0.6475	0.254463	0.064751	6.280887582
12	123	120	0.29	0.2986	$\overline{0}$ .094437	0.008918	1.070201636
13	1379	1376	0.61	0.7089	0.315863	0.099769	137.2827422
14	30	27	0.24	0.2448	$\overline{0}$ .148237	0.021974	0.593303621
15	40	37	0.04	0.04	$\overline{0}$ .353037	0.124635	4.611499565
16	120	117	0.70	0.8673	0.474263	0.224925	26.316271
17	687	684	0.21	0.2132	$\overline{0}$ .179837	0.032341	22.12148105
18	180	177	0.37	0.3884	$\overline{0}$ .004637	2.1505	0.003805813
19	54	51	0.30	0.3095	$\overline{0}$ .083537	0.006978	0.355899949
20	860	857	0.48	0.533	0.139963	0.01959	16.78832265
21	180	177	0.31	0.3205	$\overline{0}$ .072537	0.005262	0.9313406097
22	227	224	0.42	0.4477	0.054663	0.002988	0.669321759
23	6462	6459	0.43	0.4477	0.054663	0.002988	19.29977341
24	107	104	0.12	0.1206	$\overline{0}$ .272437	0.74222	7.719079573
25	78	75	0.36	0.3769	$\overline{0}$ .016137	0.00026	0.019530208
26	60	57	0.62	0.725	0.331963	0.110199	6.281367702
27	159	156	0.36	0.3769	$\overline{0}$ .016137	0.00026	0.040622832
28	212	209	0.03	0.03	$\overline{0}$ .363037	0.131796	27.54533544
29	42	39	0.36	0.3769	$\overline{0}$ .016137	0.00026	0.010155708
30	222	219	0.78	1.0454	0.652365	0.425577	93.20146895
		6683	M. Fisher	0.393037			1444.972198*
			W. Fisher	0.434713		*P<0.05 (significant)	

From the table, the observed chi-square $\div^{2}$ = 1444.97 while the critical value at 29 df = 42.557. Since $\div^{2}$ observed is greater than $\div^{2}$ critical, the null hypothesis is rejected. This implies that the empirical studies are significantly different in terms of their effect sizes 'r'. This is an indication that there is no statistical linear trend in terms of effect size across this set of study. The heterogeneity of the set of effect sizes referred to fluctuation from the average of the group. The heterogeneity of the effect sizes was indicative of moderator variables operating.

One would need to be cautious about drawing any simple overall conclusion because of the heterogeneity of the 30 studies combined. That the results of the 30 studies differed from each other could be a function of differences in the sample sizes on which r was based; it could also be publication or methodological features. The question now is ‘is the heterogeneity of the effect sizes due to heterogeneity of the sample sizes used for the various studies? To establish the fact that the effect sizes were influenced by the heterogeneity of the sample size, the null hypothesis two was raised.

Hypothesis 2: The heterogeneity of the effect size is not due to heterogeneity of sample size used on the various empirical studies

Table 4. Test of Heterogeneity

Study S/N	Sample size	N-3	r	Zr	(N –3)(Zr)
1	250	247	0.39	0.4118	101.7146
2	558	555	0.32	0.3310.6	184.038
3	300	297	0.21	0.232	63.3204
4	121	118	0.47	0.5101	60.1918
5	40	37	0.04	0.04	1.48
6	800	797	0.28	0.2877	229.2969
7	30	27	0.09	0.0902	2.435
8	1800	1797	0.86	1.1155	2004.5535
9	750	747	0.18	0.182	135.954
10	802	799	0.09	0.0902	72.0698
11	100	97	0.57	0.6475	62.8075
12	123	120	0.29	0.2986	35.832
13	1379	1376	0.61	0.7089	975.4464
14	30	27	0.24	0.2448	6.6096
15	40	37	0.04	0.04	1.48
16	120	117	0.70	0.8673	101.4741
17	687	684	0.21	0.2132	145.8288
18	180	177	0.37	0.3884	68.7468
19	54	51	0.30	0.3095	15.7845
20	860	857	0.48	0.533	456.781
21	180	177	0.31	0.3205	56.7285
22	227	224	0.42	0.4477	100.2848
23	6462	6459	0.43	0.4477	2891.694
24	107	104	0.12	0.1206	12.5424
25	78	75	0.36	0.3769	28.2675
26	60	57	0.62	0.725	41.325
27	159	156	0.36	0.3769	58.7964
28	212	209	0.03	0.03	6.27
29	42	39	0.36	0.3769	14.6991
30	222	219	0.78	1.0454	228.9426
		16683	M. Fisher	0.39307	8165.395
			W. Fisher	0.48994	

Source: From the work of Snedeco and Cohram (1967,1980)

Weighted Mean Fisher W $\bar{Z}r = \frac{\Sigma(Nkj-3)Zrj}{\Sigma(Nj-3)}$ and

Mean Fisher $\bar{Z}r = \sum \frac{zr}{k}$

The Mean Fisher $\overline{Zr}$ = 0.393039 with its associated r = 0.375, Also the Weighted Mean Fisher $\overline{WZr}$ = 0.489944 with the associated r = 0.450. The Weighted Mean Fisher is greater than Mean Fisher. This implied that weighting by sample size led to a bigger estimate of combined effect than when weight was not used. The implication of this was that sample size affected the Mean Fisher (Zr). Hence, the null hypothesis was rejected. The heterogeneity of the effect size of the selected studies was due to heterogeneity of the sample sizes used by the various primary researchers. The results of the 30 empirical studies therefore differed from each other because of differences in the sample sizes used Hence, the weighted Mean Fisher $\overline{WZr}$ = 0.489944 with associated r = 0.450 was a better measure because it corrected for the diversity of sample sizes used by different researcher.

4 Discussion

Assessing heterogeneity in meta-analysis is a crucial issue because of meta-analyst decision to select the statistical model to be applied i.e...fixed -versus random effect model (Huedo-Medima et al,2006). This misspecification has substantial negative consequence on the result of this meta-analysis study. The result of this study cast some lights on the implications of heterogeneity of the sample size on the effect size of a meta-analysis study, thus establishing the findings of Koetse, Fiorax, & De-root (2005) (which is the impact of omitted variables and erroneous effect size measures on the result of a meta-analysis. Also that the Q has low power when the number of studies is small and too much power when the number of studies is high). The Weighted Mean Fisher was greater than the Mean Fisher because of differences in the sample size on which 'r' was based .This established the findings of Field (2003) that variation could be influenced on undetermined numbers of characteristic that vary among the studies as those related to the characteristics of the sample. Meta-analysis sample size is far more effective in reducing meta-estimator than primary study sample size. With relative small increase in meta–analysis sample size, the quality of the outcome of the analysis is substantially improved, even when the effect size heterogeneity is high. The various types of effect size heterogeneity may have substantial detrimental effect from the true underlying effect average out of sample size that are common in practice.

That the studies were significantly different in terms of their effect size was an indication that there is no statistical linear trend in terms of effect size across this set of study. The heterogeneity of the set of effect sizes referred to fluctuation from the average of the group which gave the indication of moderator variables. Weighting by sample size led to a bigger estimate of combined effect thus establishing the works of Higgin & Thompson (2002). Characteristics under researchers' control did not contribute to any large amount of variance in the strength of previously reported validity studies of this meta- analysis as Brockwill et al (2001) assumed as part of the sources of variability in the result of any meta-analysis study. The heterogeneity in a set of the studies was a result of variability which occurs as a result of different sample sizes used for the study samples.

Conclusion

Heterogeneity is to be expected in a meta-analysis. It could be surprising therefore if not impossible if multiple studies performed by different researchers in different places with different methods all ended up establishing the same underlying parameter. For this study the impact of heterogeneity of the sample sizes resulted to difference in the results of the primary researchers. In research work therefore, a standardized sample size is recommended to be assigned to published articles and the unpublished articles should also be within a given range of sample sizes.

References

Adeyemo E.O (2007) “A Meta-Analysis of Empirical Studies on the Validity of University Matriculations Examinations in Nigeria” *Unpublished PhD Thesis, Department of Educational Foundations and Counselling ,Obafemi Awolowo University Ile-Ife*.

Alexander R.A,Scozzaro M.J.&Borokin L.J. (1989) “Statistics and Empirical Examination of the Chi-squre;’] test for homogeneity of correlations in Meta-analysis. *Psycological bulletin;106,329-331.*

Biggerstaff B.J. and Twede R.L. (1997) Incorporating variability estimate of heterogeneity in the random effect model in meta-analysis *Statistics in medicine.16,753*

Birge R.T.(1932) The Calculation of errors by methods of least squre Physical Review.40,207-227

Brockwill S.E & Gordom R.I .(2001) A comparison of Statistical methods for meta-analysis. *Statistics in medicine,20,825-840.*

Cochram W.G. (1954) The combination of Estimates from different experiments Biometrics, 10,101-129.

Cochram W.G.(1952) The chi-squared test of goodness of fit .Annual of Mathematical *Statistics 23,315-345*.

Field A .P.(2001). Meta-Analysis of Correlation Coefficients: A Monte Carlo Comparison of

Fixed- and Random-Effects Methods. *Psychological Methods* 6: 161–180.

Field A.P. (2003) The problem of using fixed effect models of meta analysis world data

Fleishman A.I. (1978) A method for Simulating non normal distributions. *Psychometrical,43,521-531*

Glass G.V.(1976) Primary Secondary and Meta-Analysis of Research “A paper presented at the annual meeting of the American Educational Research Association San-Francisco.

Hardy R.J & Thompson S.G .(1998) Detecting and Describing heterogeneity in Meta-analysis *Statistics in medicine 17,841-856*.

Harwell M (1997) An empirical Study of Hedges homogeneity test. *Psychological methods 2, 219-231.*

Hedges L.V. and Okin I (1985) Statistics Methods for Meta-analysis .New York Academic Press.

Hedges L.V. and Vervea J.L.(1998),Fixed and random effect models in meta-analysis. *Psychological method ,3 486-504*

Higgin J.P.T & Thompson S.G.(2002) Quantifying heterogeneity in a meta-analysis. *Statistics in medicine,21.1539-1558*

Higgin J.P.T ,Thompson S.G, Deek J.J,& Attman D.G (2003) Measuring Inconsistency in Meta-analysis. *British Medical Journal 327;557-560.*

Koetse M J, Florax R J G M, de Groot H L F.(2005). Correcting for Primary Study Misspecifications in Meta-Analysis. Tinbergen Institute Discussion Paper 05-029/3. Tinbergen

National Research Council (1992).Combining Information: *Statisticscal issue andOpportunities for Research*. Washington D.C National Academy Press

Rosenthal R (1984) Meta-analytic procedures for social research Beverly Hills, *Califonia Sage Publication.*

Rosenthal R.and Rosnow (1984) Essentials of behavioral research. Methods for data analysis McGraw-Hi,Inc*Understanding Statistics,2, 77-96*