Journal of Social Science and Economics
Vol 3, No. 2, Oktober 2024, pISSN : 2829-209X, eISSN : 2829-3312
pp. 92-118

Assessing the Effect Size and Homogeneity of Selected Standardized Tests in Education

Temitope Babatimehin[1], Oyeronke Christiana Paramole[2], Olufunke Favour Akindahunsi[3], Temitope Sarah Ogungbaigbe[4]

[1] Department of Educational Foundations and Counselling, Obafemi Awolowo University, Ile-Ife, Nigeria
[2] Department of Social Studies Education, Al-Hikmah University, Ilorin, Nigeria
[3] Department of Arts and Social Science Education, Osun State University, Osogbo, Nigeria
[4] Department of Educational Foundations and Counselling, Obafemi Awolowo University, Ile-Ife, Nigeria

Keywords: Effect size, Homogeneity, Test, Standardised, Education

Abstract

The study was conducted to analyse the empirical differences in the studies on procedures for validating standardised tests in Education. It examined the difference in the findings of the selected studies in terms of the sample size and also determined the homogeneity of the effect size of the selected studies. The study adopted an ex-post factor research design. The study population comprised empirical studies on validating standardised tests in Education (1988 to 2017). The sample size consisted of 130 empirical studies on the validation of standardised tests in education (1988-2017), which were selected using a purpose sampling technique. The research instrument comprised both published and unpublished articles, PhD and Masters theses. Data were analysed using counts and percentages, study variance, and statistical transformations to convert t-test and f-test to corresponding 'r' statistics. Also, Q statistics, which is the sum of weighted squares based on Chi-square, I2, Tau2 and Fishers Z, were used for the analysis. Results showed an inverse relationship between sample and effect sizes in the selected studies. Also, there is no significant difference in the effect sizes of studies (Zr=1.14403). Furthermore, the test for homogeneity showed that the studies significantly differ in their effect sizes; Q=24940.437, I2 = 99.48%, and Tau^2 = 0.209>0 Tau=0.457 do not belong to the same population.

Kata kunci:
Ukuran efek, Homogenitas, Uji, Terstandardisasi, Pendidikan

Article history:
Received: 09-03-2025
Revised 19-04-2025
Accepted 20-05-2025

Abstrak

Penelitian ini dilakukan untuk menganalisis perbedaan empiris dalam penelitian tentang prosedur untuk memvalidasi tes standar dalam bidang pendidikan. Studi ini meneliti perbedaan temuan dari studi yang dipilih dalam hal ukuran sampel dan juga menentukan homogenitas ukuran efek dari studi yang dipilih. Penelitian ini mengadopsi desain penelitian faktor ex-post. Populasi penelitian terdiri dari studi empiris tentang validasi tes terstandarisasi di bidang pendidikan (1988 hingga 2017). Ukuran sampel terdiri dari 130 studi empiris tentang validasi tes terstandardisasi di bidang pendidikan (1988-2017), yang dipilih dengan menggunakan teknik pengambilan sampel bertujuan. Instrumen penelitian terdiri dari artikel yang diterbitkan dan tidak diterbitkan, tesis PhD dan Master. Data dianalisis dengan menggunakan jumlah dan persentase, varians studi, dan transformasi statistik untuk mengubah uji-t dan uji-f menjadi statistik 'r' yang sesuai. Selain itu, statistik Q, yang merupakan jumlah kuadrat tertimbang berdasarkan Chi-square, I2, Tau2 dan Fishers Z, juga digunakan untuk analisis. Hasil penelitian menunjukkan hubungan terbalik antara sampel dan ukuran efek dalam studi yang dipilih. Selain itu, tidak ada perbedaan yang signifikan dalam ukuran efek penelitian (Zr = 1,14403). Lebih lanjut, uji homogenitas menunjukkan bahwa penelitian-penelitian tersebut secara signifikan berbeda dalam ukuran efeknya; Q = 24940.437, I2 = 99.48%, dan Tau2 = 0.209>0 Tau = 0.457 tidak berasal dari populasi yang sama.

Corresponding Author:
Temitope Babatimehin
Obafemi Awolowo University, Ile-Ife, Nigeria; tbabatimehin@oauife.edu.ng

INTRODUCTION

One of the most important aspects of teaching-learning is testing. It is used to rate students at the end of the teaching-learning process. Any test given in the same manner to all test takers and graded in the same manner for everyone is a standardised test (Simpson, 2017). Pomroy (2020) defined standardised tests as designed so that the questions, conditions for administering, scoring procedures, and interpretations are consistent. They are administered and scored in a standard manner. Tests are also used to measure human attributes. Tests not only measure cognitive abilities but also non-cognitive abilities. However, standardised tests do not necessarily need high-stakes, time-limited or multiple-choice tests. Instead, tests should indicate the abilities or skills being measured. In order to ensure uniformity of test procedures, the manual that accompanies a standardised test usually includes detailed and clear instructions for administering the test so that the same or similar score will be obtained even when different testers administer the test; human judgment is subjective and fallible (Nahar, 2023). The objective nature of standardised tests is one of the main advantages they have over other methods for assisting researchers to understand human behaviour and make decisions about it, not the least because it reduces errors of judgment relating to personal bias or subjectivity (Nahar, 2023).

Zhao et al. (2023) highlight the importance of standardised tests in scientific measurement, which are quantitatively summarised in scores. These tests allow for more precise and clear descriptions of human behaviour, such as IQ scores, which provide a more fine-grained description of a person's intellectual ability. Standardised tests can be given on various topics, such as driving tests, creativity, personality, professional ethics, interest, achievement, attitude, intelligence, and personality. To be standardised, a test must be prepared by experts, administered to a representative population, have adequate psychological properties, have a norm, and have standard uniform procedures for administration and scoring. Meta-analysis, a method used to summarise the results of multiple empirical research studies, is now widely spread in psychology, education, social sciences, and medicine (Mikolajewicz & Komarova (2019). Meta-analysis focuses on contrasting and combining results from different studies to identify patterns, sources of disagreement, or interesting relationships that may arise in multiple studies. Ovute (2015) defined meta-analysis as a statistical analysis combining the results of several independent studies that the analyst considers combinable. Pigott and Polanin (2020) also mentioned that meta-analysis is a collection of systematic techniques for resolving apparent contradictions in research findings. In the opinion of Pigott and Polanin (2020), educational research often produces contradicting results. It was mentioned that differences among studies in

treatment, settings, measurement instruments and research methods make research findings difficult to compare. Even frequent replication can prove inconclusive, and literature on a topic may be extensive, obscure trends, and provide an overwhelming amount of impartiality and quantities of the description of the findings in a population of studies on a particular topic. Currently, in education, researchers are faced with abundant information, and most often, the problem is finding knowledge in the information. There is a need for an orderly summarization of studies so that knowledge can be extracted from the numerous individual studies. This is important locally and internationally, where little work has been done on the systematic procedures for indexing studies. Soni (2023) suggested finding untapped knowledge in complete research studies. The best minds are needed to integrate the staggering of individual studies; this endeavour deserves a higher priority than adding new experiments or surveys to the pile (Gurung et al., 2023).

Effect size is the degree to which a null hypothesis is false (Perdices, 2018). Also, Mengist et al. (2020) explained that effect size allows for examining the relationship between independent and dependent variables of a given study. The effect size in meta-analysis is the strength and direction of the relationship between variables (Gignac & Szodorai, 2016). The difference between the means of pairs of treatment conditions is divided by the standard deviation of either group (control or experimental group standard deviation). Meta-analysis estimates the actual effect size instead of a less precise effect size derived in a single study under a single set of assumptions and conditions. Meta-analysis methods include the voting, literary approach, combined tests, and effect size methods. According to Cook et al. (2018), effect size indicates the practical significance of research outcomes. It is defined as the differences between the means of the two groups divided by their typical standard deviation and an expression of the increase or decrease in the achievement of the experimental group (the group of students exposed to specific techniques) in standard deviation units. Effect size is defined as the frequency of occurrence of a particular effect in a study or a social phenomenon. So, effect size measures the magnitude of a treatment effect in a study.

Effect size spells out how significant an effect is, disregarding its level of significance and its number. It helps examine the strengths and direction of the relationship between the independent and dependent variables in a research study (Dattalo, 2013). This term may be expressed in different ways for various fields. In medicine, the effect size is expressed as the application effect and is sometimes expressed as the odds ratio, the risk ratio or the risk difference. In social sciences and education, the term ‘effect size’ is used frequently but is

sometimes expressed as the standardised mean difference or relationships (Schäfer & Schwarz, 2019). The most frequently used effect size calculations fall into these categories: (1) proportions, (2) averages and (3) correlation coefficients. There is more than one way to calculate effect size in these categories. Studies testing the effect of an intervention or making various causal inferences (between pre-and post-test or between groups receiving and not receiving treatment) are in the category that uses proportions and averages (Barker et al., 2024). Studies investigating the relationship between variables, besides causal direction inferences, are in the category of correlational meta-analysis (Çoğaltay & Karadağ, 2015). In other words, if the effect size results are numerical, then averages are used; if the results are nominal, then proportions are used; and if the results show a relationship, then correlations are preferred (Nuijten et al., 2020). There are two important differences in the calculations of effect size: dichotomous data and continuous data. Dichotomous variables are based on only two categories and frequently represent the presence or lack of a feature or situation. Pregnancy, high school graduation, and gender are examples of such variables. Continuous variables can have a range of values expressed on a numeric scale. Examples of such variables include the number of pregnancies, the duration of training, and the duration of hospitalisation. Test results such as achievement tests or depression inventories can be considered continuous variables (Voyer & Voyer, 2014).

It is particularly valuable in best practices research because it represents a standard measure of all outcomes. For instance, one can compare the effect sizes of academic outcomes. Although many researchers use the concept of statistical significance to determine whether a particular study had an effect, this is not necessarily a good idea since statistical significance is heavily dependent upon sample size.

Cohen provided measures of effect size for many standard statistical tests.

$$d = [\bar{X}1 - \bar{X}2] / SD \qquad (2.1)$$

The absolute value of the difference between the two groups $[\bar{X}1 - \bar{X}2]$ is divided by the (SD) to obtain the standardised scale invariant estimate of the effect size (d). The standard deviation is either the control group or the pre-test SD, as the group's variance is assumed to be equal. For correlation relationships, the average of correlations between the two features that examined the same research questions across separate research studies is obtained by averaging the raw Pearson correlation coefficient (r) using the formula

$$\bar{r} = \Sigma r / n \qquad (2.2)$$

Where "r" = Pearson correlation from each study and

'n" = number of correlation coefficients to be combined

Tanha et al. (2017) stated that high P-values may not necessarily indicate or imply a significant effect and vice versa. For instance, minimal effects can be statistically significant with a study sample of 10,000 students, while relatively large effects are often not statistically significant with 50 students. Ultimately, what matters most is not statistical significance but whether the size of an effect is meaningful in a practical sense. Effect size is also referred to as the vote-counting method because effect size can reveal how significant an effect is regardless of the level of significance and size of the sample. Effect size can be measured in different ways as follows: a) as the standard difference between two means; b) as the correlation between the independent variable classification and the individual scores on the dependent variable; c) the use of I-Q, which may provide a standard index which is used as an effect size without transformation and difference between the proportion of individuals in treatment and control groups. The development of standardised tests includes both qualitative and quantitative procedures to ensure that such tests are reliable and valid. Variations or shortcomings in the a priori processes of standardised test development may lead to construct measures with doubtful validity since the test development procedure varies across researchers. When tests provide measures that do not adequately show the trait they purport to measure, this will lead to wrong conclusions and inappropriate decisions. This may harm an individual's health, academics, career, personal problems, personality, social relations, and skill development. Hence, it is important to synthesise standardised test procedures, analyse scale statistics, and integrate them, considering variations in the sample size.

RESEARCH METHOD

The study adopted an "*Ex-post-Facto*" research design. The ancestral approach of information retrieval was also used to locate studies. According to Ovute (2015), the ancestry approach ensures that related studies are located by tracking citations from one study to another through bibliographies cited in earlier studies. Also, the cause-effect relationship between the results of previous studies conducted in test validation among the selected studies should be established, as these procedures could have resulted in divergent results. The study population comprised all previous empirical studies on the validation of standardised tests published from 1988 through 2017 in local and international journals whose primary purpose was to validate standardised tests. Only

empirical studies were included in the sample. Empirical methods include collecting and organising data and drawing conclusions about those data. Journals, unpublished master's and doctoral theses locally developed were also considered. Each chosen journal issue was examined and read carefully to identify potential articles for inclusion in this study. The sample consisted of 130 empirical studies previously carried out on validation of standardised tests both locally and internationally to cover enough studies and make the results valid because a sparse number of studies had been conducted in this area locally. Some of the samples identified were not used because, in some cases, some studies were reported more than once; from the 149 studies identified during the literature search, nineteen (19) studies did not contain information that would enable the calculation of effect size and were therefore discarded. The purposive sampling technique was used in collecting relevant studies because the researcher focused on sampling results of primary studies that utilised empirical data and relevant psychometric analysis. The research studies comprised published and unpublished journal articles, unpublished theses and dissertations. The sample was chosen based on the following criteria;

1. The study used quantitative methods to report the analysis of the results of the tests developed.
2. The test statistics are convertible to Pearson-r.
3. The study was carried out or published between 1998 and 2017.
4. The study reported a significant level and sample size of its result.
   The distribution of the studies to be used for the meta-analysis in terms of publication source is as indicated in Table 1

Table 1

The Distribution of 130 Studies that were Used for the Research

Type of Study	**Number of Studies**
Journal (Published) International	73
Journal (Published) Local	35
PhD Theses (Unpublished)	5
Masters Theses (Unpublished)	17
Total	130

The research instrument used in this study was a coding sheet by Cooper and Hedges, modified by Zubar (2014). The coding sheet was modified to suit the research objectives and included necessary variables for each study to explain variance in the procedures of validating standardised tests in education. The selection of variables was based on expert review, with the objectives and hypotheses determining the type of variables coded. Two general categories of information were coded for each empirical study on validation of standardised

tests: study descriptors and effect size information. Study descriptors included methodological variables, study context, subject characteristics, and test development and validation procedures. Effect size information included means, standard deviations, and other statistics. The coding sheet was subjected to a pilot study by coding 20 sample subsets. Results were tested using inter-rater kappa statistics to determine internal consistency. Experts corrected the coding sheet, retaining items 1-6, modifying 7 and 8, and deleting items 16 and 17.

Table 2

Instrument for Data Collection (Coding Sheet).

S/N	Preliminary	Code
1	Type of publication	Articles (Published) "4" PhD Thesis (Unpublished) "3" Masters Thesis (Unpublished) "2" Specified Reports/ Conference Papers "1"
2	Year of study/publication	Within 1-5 years ago (2017-2012), "5" Within 6-10 years ago (2011-2006) "4" Within 11-15 years ago (2005-1995), "3" Within 16-20 years ago(1994-1991) "2" Within 21 years and above (1990-1988),"1."
3	Types of statistics used in computing results	Multivariate "3" Bivariate "2" Univariate "1"
4.	Sample Type	Sample drawn from across state "2" Sample drawn from state local govt. "1"
5.	Sample size	1000 and above "4" 999-500 "3" 499-200 "2" 199 and below "1."
6	Location of the study	Urban, Semi-Urban, Rural areas "3" Any two of the above "2." Any one of the above "1."
	Steps in scale validation	
7	Generation of item	Theory "6" Interview/Focus Group "5" Suggestion by experts "4" Literature review "3" Clinical / Observation "2" Responses to open-ended questions "1"

8	Methods of item deletion	Theory "12" Advice "11" Item total correlation "10" Alpha if item deleted "9." Factor Loading "8" Item, means, standard deviation "7" Item correlation too low "6" Content coverage "5." Loading on the wrong factor "4." Poorly worded "3" Low variance "2" None "1"
9	Forms of Reliability	Internal consistency (Cronbach's Alpha) "6" Internal consistency (Spearman-Brown split half) "5." Internal consistency (Guttman's Split half) "4" Test-retest (test of stability) "3" Inter-rater (Marginal or average) "2" IRT "1"
10	Types of validity	Construct "7" Content "6" Predictive "5" Discriminant "4" Convergent "3" Correlation "2" Criterion "1"
11	Factor Analytic Procedure	Exploratory factor analysis "3" Confirmatory factor analysis "2" EFA and CFA "1"
12	Factor ability of correlation matrix	Absolute sample size "5." Inter–item correlation "4" Participant per item ratio "3." Bartlett's test of sphericity "2" KMO method of sampling adequacy "1"
13	Analysis of factor retention	Loadings "7" Communalities "6" Item analysis "5" Eigen-values "4" Scree plot "3" Minimum portion of variance accounted for by a factor "2" Parallel analysis "1"

14	Factor extraction method	PCA Principal component analysis "5." Common factor analysis "4" Principal axis factoring "3" Maximum likelihood "2" Unweighted least squares "1"
15	Rotation method used	Orthogonal (Varimax) "4" Promax "3" Oblique "2" Oblium "1"

Note: Items 7-15 were grouped and named "study quality."

Previous primary studies on "validation" of standardised tests "scale" were searched using computerised databases such as (PsychLIT, ERIC, PsychINFO database), and library. A manual search was also conducted in the Department of Educational Foundations and Counseling Obafemi Awolowo University for Master's and PhD—theses on test validation. Also, the Ancestry approach (Carroll et al., 2020) was used to retrieve relevant materials by tracking citations from one study to another through bibliographies cited in theses, articles and dissertations. The follow-up yielded several articles and studies. The central computation in meta-analysis is effect size, which measures the strength and direction of the relationship between variables (Littel, Corcoran & Pillai, 2008). Statistical transformations were used to convert the t-test and f-test to corresponding 'r' statistics. Research question one was analysed using Q statistics, the sum of weighted squares, and based on the χ^2 **(**Kulinskaya et al., 2021; Tellinghuisen, 2022), I^2, and Tau^2. Research question two was analysed using study variance; hypothesis 1 was analysed using Fishers Zr, and hypothesis 2 was tested by calculating the chi-square value with the help of Statistical Package for Social Sciences (SPSS).

Table 3

Transformation of Statistical Tests of Studies to Pearson 'r'

S/N	Sample Size	Test Statistics	Effect Size 'r'
47	2166	t= 2.103	0.83
53	150	F=3.32	0.92
98	283	t= 2.59	0.87

RESEARCH RESULTS AND DISCUSSION

Research one: How does sample size affect the findings of the selected studies?

Sample and effect sizes were reported in CMA as study variance to answer the research question. This implies that effect sizes vary as a result of sample size.

Table 4

Relationship between Sample Size and Effect Size

S/N	Sample Size	Effect size (study variance)

1	120	0.00855
2	1200	0.00084
3	1500	0.00067
4	1210	0.00083
5	120	0.00855
6	960	0.00104
7	300	0.00337
8	503	0.00200
9	253	0.00400
10	13,000	0.00008
11	858	0.00117
12	1440	0.00070
13	159	0.00641
14	1090	0.00092
15	428	0.00235
16	1710	0.00059
17	2600	0.00039
18	850	0.00118
19	300	0.00337
20	600	0.01754
21	310	0.00326
22	864	0.00116
23	450	0.00224
24	60	0.01754
25	200	0.00508
26	128	0.00800
27	268	0.00377
28	98	0.01053
30	300	0.00337
31	1000	0.00100
32	680	0.00148
33	443	0.00227
34	506	0.00199
35	400	0.00252
36	1360	0.00074
37	680	0.00148
38	202	0.00503
39	362	0.00279
40	491	0.00205
41	1474	0.00068
42	224	0.00452
43	297	0.00340
44	46	0.02326

45	10,000	0.00010
46	960	0.00104
47	2166	0.00046
48	502	0.00200
49	707	0.00142
50	1030	0.00097
51	100	0.01031
52	70	0.01493
53	150	0.00680
54	212	0.00478
55	428	0.00235
56	50	0.02128
57	250	0.00405
58	500	0.00201
59	88	0.01176
S/N	**Sample Size**	**Effect size (study variance)**
61	1500	0.00067
62	600	0.00168
63	750	0.00134
64	1000	0.00100
65	600	0.00168
66	600	0.00168
67	480	0.00210
68	600	0.00168
69	300	0.00337
70	490	0.00205
71	224	0.00452
72	1800	0.00056
73	465	0.00216
74	300	0.00337
75	328	0.00308
76	885	0.00113
77	372	0.00271
78	147	0.00694
79	370	0.00272
80	1014	0.00099
81	500	0.00201
82	70	0.01493
83	202	0.00503
84	965	0.00104
85	550	0.00183
86	375	0.00269
87	227	0.00446
88	1500	0.00067

89	520	0.00193
90	573	0.00175
S/N	**Sample Size**	**Effect size (study variance)**
92	800	0.00125
93	5	0.50000
94	216	0.00469
95	341	0.00296
96	2063	0.00049
97	350	0.00288
98	283	0.00357
99	21156	0.00005
100	346	0.00292
101	1309	0.00077
102	2156	0.00046
103	997	0.00101
104	698	0.00144
105	346	0.00292
106	484	0.00208
107	3000	0.00033
108	391	0.00258
109	294	0.00344
110	127	0.00806
111	315	0.00321
112	325	0.00311
113	172	0.00592
114	1092	0.00092
115	304	0.00332
116	239	0.00424
117	184	0.00552
118	600	0.00168
119	159	0.00641
120	786	0.00128
121	600	0.00168
122	355	0.00284
123	82	0.01266
124	293	0.00345
125	293	0.00345
126	947	0.00106
127	2412	0.00042
128	400	0.00252
129	203	0.00500
130	3883	0.00026

	0.95596

Table 4 shows the inverse effect of the sample size; a large sample size results in a small effect size, while a small sample size results in a large effect size. For example, study 10, with the largest sample size of 13,000, has the least effect size of 0.00008, while studies 93, 44, 50, and 70, with low sample sizes of 5, 46, 50 and 70, reported large effect sizes of 0.500, 0.2326, 0.2128, and 0.01493. Also, studies 45, 99, and 130 with large sample sizes 10,000, 21,156 and 3883 have low effect sizes of 0.00010, 0.00005 and 0.00026, respectively. Therefore, it was concluded that there was an inverse relationship between sample size and effect size of the studies.

Research Question 2: How homogeneous are the effect sizes of the selected studies?

Table 5 shows Q statistics, I2 and the computation of Tau2, which is the variance of the actual study effect in explaining the homogeneity of sample sizes of the studies selected. Q statistics 24940.437 means a significant variance among studies (heterogeneous). Tau^2 is used to explain the heterogeneity of studies further. Also, $Tau^2 = 0.209 \geq 0$ emphasises that the studies are heterogeneous and do not belong to the same population. I^2 is 99.483, almost 100%, which is close to the dispersion of the effect size. Since Tau^2 is the variance of the actual study effect, Tau is the standard deviation of the actual study effect, which is 0.457.

Table 5

Values and Effect Sizes of Studies

Model	**Effect size and 95% confidence interval**						**Test of null [2-Tail]**		**Heterogeneity**				**Tau-squared**			
Model	No S	PE	SE	Var	LL	UL	Z-v	P-v	Q-v	Df (Q)	P-v	I-s	TS	SE	Var	Tau
Fixed	130	1.323	0.003	0;000	1.318	1.329	468.409	0.000	24940.437	129	0.000	99.483	0.209	0.064	0.004	0.457
Random	130	1.190	0.041	0.002	**1.111**	1.269	29.363	0.000								

Note: No S = Number of Studies, PE= Point Estimate, SE = Standard Error, Var = Variance, LL = Lower Limit, UL = Upper Limit, Z-v = Z-value, P-v = P-value, Q-v=Q-value, TS= Tau Square, I-s= I-square

Hypothesis 1: There is no significant difference in the effect size of the selected studies.

Low, medium, and very high effect sizes were identified to determine whether the 130 studies on validation procedures differed significantly in terms of their effect sizes.

The Overall Effect of the Selected Studies

Table 6

Studies with Large Effect Size (0.80- 1.0)

Study	Sample size	Fisher Zr	Effect size 'r'	Study	Sample size	Fisher Zr	Effect size 'r'
10	13,000	2.29	0.98	82	70	1.33	0.87
41	1474	2.29	0.98	88	1500	1.33	0.87
64	1000	2.09	0.97	98	283	1.37	0.87
52	70	1.83	0.95	108	391	1.33	0.87
81	500	1.83	0.95	119	159	1.33	0.87
86	375	1.83	0.95	127	2412	1.33	0.87
100	346	1.83	0.95	51	100	1.29	0.86
4	1210	1.66	0.93	112	325	1.29	0.86
46	960	1.66	0.93	23	450	1.26	0.85
111	315	1.65	0.93	27	268	1.26	0.85
17	23600	1.59	0.92	35	400	1.26	0.85
53	150	1.59	0.92	60	289	1.26	0.84
55	428	1.58	0.92	5	120	1.22	0.84
66	600	1.59	0.91	38	202	1.22	0.84
2	1200	1.53	0.91	49	707	1.22	0.84
8	503	1.53	0.91	50	1030	1.22	0.84
21	310	1.53	0.91	61	1500	1.22	0.84
25	200	1.52	0.91	78	147	1.22	0.84
32	680	1.52	0.91	91	650	1.22	0.84
39	362	1.52	0.91	97	350	1.22	0.84
45	10000	1.53	0.91	33	443	1.19	0.83
48	502	1.53	0.91	47	2166	1.19	0.83
117	184	1.53	0.91	65	600	1.18	0.83
123	82	1.53	0.91	85	550	1.88	0.83
12	1440	1.42	0.89	99	21156	1.88	0.83
30	300	1.42	0.89	103	997	1.18	0.83
57	250	1.42	0.89	3	1500	1.16	0.82
87	227	1.42	0.89	16	1710	1.16	0.82
89	520	1.42	0.89	20	600	1.16	0.82
118	600	1.42	0.89	22	864	1.16	0.82
122	355	1.42	0.89	68	600	1.16	0.82
59	88	1.37	0.88	101	1309	1.16	0.82
92	800	1.37	0.88	121	600	1.16	0.82
114	1092	1.38	0.88	11	858	1.13	0.81
15	428	1.33	0.87	19	300	1.13	0.81
36	1360	1.33	0.87	54	212	1.13	0.81
62	600	1.33	0.87	69	300	1.12	0.81
67	480	1.33	0.87	80	1014	1.12	0.81
71	224	1.33	0.87	84	965	1.12	0.81

75	328	1.33	0.87	107	3000	1.13	0.81

Table 6 showed that 80 out of the 130 studies had high effect sizes, as classified by Cohen. The effect sizes ranged from r=0.98 (studies 10 and 41) to r=0.81 (studies 11, 19, 54, 69, 80, 84, and 107).

Table 7
Studies with Medium Effect size (.50 - 0.79)

Study	Sample size	Fishers Zr	Effect size 'r'
37	680	1.07	0.79
43	297	1.07	0.79
1	120	1.05	0.78
40	491	1.05	0.78
95	341	1.04	0.78
105	346	1.05	0.78
129	203	1.05	0.78
6	960	1.02	0.77
72	1800	1.02	0.77
76	885	1.02	0.77
125	293	1.02	0.77
73	465	0.99	0.76
83	202	0.99	0.76
90	573	0.99	0.76
7	300	0.97	0.75
29	715	0.97	0.75
56	50	0.97	0.75
58	500	0.97	0.75
109	294	0.97	0.75
24	60	0.95	0.74
63	750	0.95	0.74
102	2156	0.95	0.74
44	46	0.93	0.73
9	253	0.91	0.72
13	159	0.91	0.72
31	1000	0.9	0.72
115	304	0.91	0.72
93	5	0.88	0.71
106	484	0.89	0.71
128	400	0.89	0.71
96	2063	0.83	0.68
18	850	0.81	0.67
28	98	0.81	0.67
34	506	0.81	0.67

79	370	0.77	0.65
126	947	0.78	0.65
26	128	0.76	0.64
94	216	0.76	0.64
104	698	0.75	0.64
14	1090	0.74	0.63
42	224	0.74	0.63

Table 7 revealed that 41 out of the 130 studies had medium effect sizes, as Cohen classified them. This indicated that the average effect size between the variables of interest ranged from r = 0.79 (study 37) to r = 0.63 (study 42).

Table 8
Studies with Small Effect Size (0.01- 0.49)

Study	Sample size	Fishers Zr	Effect size 'r'
110	127	0.68	0.49
77	372	0.66	0.48
70	490	0.61	0.45
74	300	0.61	0.44
124	293	0.61	0.47
113	172	0.59	0.43
116	239	0.58	0.42
130	3883	0.49	0.45
120	786	0.44	0.41

Table 8 shows nine of the 130 studies have small effect sizes, as reported in the Pearson correlation (r). The effect sizes range from r=0.41 (study 120) to 0.49 (study 110), which explains the low relationship between the variables of interest.

Table 9
Effect Size (r) of Selected Studies

Study	**Sample size (N)**	**Effect size (r)**	**N-3(W)**	**Fisher s Zr**	**(W)(Zr)**
1	120	0.78	117	1.05	122.85
2	1200	0.91	1197	1.53	1831.41
3	1500	0.82	1497	1.16	1736.52
4	1210	0.93	1207	1.66	2003.62
5	120	0.84	117	1.22	142.74
6	960	0.77	957	1.02	976.14
7	300	0.75	297	0.97	288.09
8	503	0.91	500	1.53	765
9	253	0.72	250	0.91	227.5
10	13,000	0.98	12997	2.29	29763.13

11	858	0.81	855	1.13	966.15
12	1440	0.89	1437	1.42	2040.54
13	159	0.72	156	0.91	141.96
14	1090	0.63	1087	0.74	804.38
15	428	0.87	425	1.33	565.25
16	1710	0.82	1707	1.16	1980.12
17	2600	0.92	2597	1.59	4129.23
18	850	0.67	847	0.81	686.07
19	300	0.81	297	1.13	335.61
20	600	0.82	597	1.16	692.52
21	310	0.91	307	1.53	469.71
22	864	0.82	861	1.16	998.76
23	450	0.85	447	1.26	563.22
24	60	0.74	57	0.95	54.15
25	200	0.91	197	1.52	299.44
26	128	0.64	125	0.76	95
27	268	0.85	265	1.26	333.9
28	98	0.67	95	0.81	76.95
29	715	0.75	712	0.97	690.64
30	300	0.89	297	1.42	421.74
31	1000	0.72	997	0.9	897.3
32	680	0.91	677	1.52	1029.04
33	443	0.83	440	1.19	523.6
34	506	0.67	503	0.81	407.43
35	400	0.85	397	1.26	500.22
36	1360	0.87	1357	1.33	1804.81
Study	**Sample size (N)**	**Effect size (r)**	**N-3(W)**	**Fishers Zr**	**(W)(Zr)**
37	680	0.79	677	1.07	724.39
38	202	0.84	199	1.22	242.78
39	362	0.91	359	1.52	545.68
40	491	0.78	488	1.05	512.4
41	1474	0.98	1471	2.29	3368.59
42	224	0.63	221	0.74	163.54
43	297	0.79	294	1.07	314.58
44	46	0.73	43	0.93	39.99
45	10,000	0.91	9997	1.53	15295.41
46	960	0.93	957	1.66	1588.62
47	2166	0.83	2163	1.19	2573.97
48	502	0.91	499	1.53	763.47
49	707	0.84	704	1.22	858.88
50	1030	0.84	1027	1.22	1252.94

51	100	0.86	97	1.29	125.13
52	70	0.95	67	1.83	122.61
53	150	0.92	147	1.59	233.73
54	212	0.81	209	1.13	236.17
55	428	0.92	425	1.58	671.5
56	50	0.75	47	0.97	45.59
57	250	0.89	247	1.42	350.74
58	500	0.75	497	0.97	482.09
59	88	0.88	85	1.37	116.45
60	289	0.85	286	1.26	360.36
61	1500	0.84	1497	1.22	1826.34
62	600	0.87	597	1.33	794.01
63	750	0.74	747	0.95	709.65
64	1000	0.97	997	2.09	2083.73
65	600	0.83	597	1.18	704.46
66	600	0.92	597	1.59	949.23
67	480	0.87	477	1.33	634.41
68	600	0.82	597	1.16	692.52
69	300	0.81	297	1.12	332.64
70	490	0.55	487	0.61	297.07
71	224	0.87	221	1.33	293.93
72	1800	0.77	1797	1.02	1832.94
73	465	0.76	462	0.99	457.38
74	300	0.54	297	0.61	181.17
75	328	0.87	325	1.33	432.25
Study	**Sample size (N)**	**Effect size (r)**	**N-3(W)**	**Fishers Zr**	**(W)(Zr)**
76	885	0.77	882	1.02	899.64
77	372	0.58	369	0.66	243.54
78	147	0.84	144	1.22	175.68
79	370	0.65	367	0.77	282.59
80	1014	0.81	1011	1.12	1132.32
81	500	0.95	497	1.83	909.51
82	70	0.87	67	1.33	89.11
83	202	0.76	199	0.99	197.01
84	965	0.81	962	1.12	1077.44
85	550	0.83	547	1.88	1028.36
86	375	0.95	372	1.83	680.76
87	227	0.89	224	1.42	318.08
88	1500	0.87	1497	1.33	1991.01
89	520	0.89	517	1.42	734.14
90	573	0.76	570	0.99	564.3

91	650	0.84	647	1.22	789.34
92	800	0.88	797	1.37	1091.89
93	5	0.71	2	0.88	1.76
94	216	0.64	213	0.76	161.88
95	341	0.78	338	1.04	351.52
96	2063	0.68	2060	0.83	1709.8
97	350	0.84	347	1.22	423.34
98	283	0.87	280	1.37	383.6
99	21156	0.83	21153	1.88	39767.64
100	346	0.95	343	1.83	627.69
101	1309	0.82	1306	1.16	1514.96
102	2156	0.74	2153	0.95	2045.35
103	997	0.83	994	1.18	1172.92
104	698	0.64	695	0.75	521.25
105	346	0.78	343	1.05	360.15
106	484	0.71	481	0.89	428.09
107	3000	0.81	2997	1.13	3386.61
108	391	0.87	388	1.33	516.04
109	294	0.75	291	0.97	282.27
110	127	0.59	124	0.68	84.32
111	315	0.93	312	1.65	514.8
112	325	0.86	322	1.29	415.38
113	172	0.53	169	0.59	99.71
114	1092	0.88	1089	1.38	1502.82
Study	**Sample size (N)**	**Effect size (r)**	**N-3(W)**	**Fishers Zr**	**(W)(Zr)**
115	304	0.72	301	0.91	273.91
116	239	0.52	236	0.58	136.88
117	184	0.91	181	1.53	276.93
118	600	0.89	597	1.42	847.74
119	159	0.87	156	1.33	207.48
120	786	0.41	783	0.44	344.52
121	600	0.82	597	1.16	692.52
122	355	0.89	352	1.42	499.84
123	82	0.91	79	1.53	120.87
124	293	0.54	290	0.61	176.9
125	293	0.77	290	1.02	295.8
126	947	0.65	944	0.78	736.32
127	2412	0.87	2409	1.33	3203.97
128	400	0.71	397	0.89	353.33
129	203	0.78	200	1.05	210
130	3883	0.45	3880	0.49	1901.2

SUM		**125834**	**155.7**	**181330.9**
	Weighted Mean Fisher		**1.44103**	

Table 9 shows the magnitude of the effect size of the sampled studies. The Mean Fisher obtained was 1.44103, with a 'r' correspondence of 0.89, symbolising a medium effect size.

$$\text{Mean Fisher} = \frac{\Sigma(W)(Zr)}{\Sigma(W)}$$

$$= \frac{181330.9}{125834}$$

$$= 1.44103$$

Hedge 'g' was used to calculate the effect size. The fixed effect model was used for homogeneous effect sizes, while the random effect model was used for heterogeneous distributions (Veronika et al., 2019).

Hypothesis 2: The selected studies are not significantly different regarding probability level.

To test this hypothesis, the chi-square value was calculated using SPSS.

Table 10

Probability Levels and Standard Deviations of Selected Studies

S/N	Sample size	P value	Standard Normal Deviate
1	120	0.05	1.96
2	1200	0.05	1.96
3	1500	0.05	1.96
4	1210	0.05	1.96
5	120	0.05	1.96
6	960	0.05	1.96
7	300	0.05	1.96
8	503	0.05	1.96
9	253	0.05	1.96
10	13,000	0.05	1.96
11	858	0.05	1.96
12	1440	0.05	1.96
13	159	0.05	1.96
14	1090	0.05	1.96
15	428	0.05	1.96
16	1710	0.05	1.96
17	2600	0.05	1.96
18	850	0.05	1.96
19	300	0.05	1.96
20	600	0.05	1.96
21	310	0.05	1.96
22	864	0.05	1.96

23	450	0.05	1.96
24	60	0.05	1.96
25	200	0.05	1.96
26	128	0.05	1.96
27	268	0.05	1.96
28	98	0.05	1.96
29	715	0.05	1.96
30	300	0.05	1.96
31	1000	0.05	1.96
32	680	0.05	1.96
33	443	0.05	1.96
34	506	0.05	1.96
35	400	0.05	1.96
36	1360	0.05	1.96
37	680	0.05	1.96
38	202	0.05	1.96
39	362	0.05	1.96
40	491	0.05	1.96
41	1474	0.05	1.96

S/N	**Sample size**	**P value**	**Standard Normal Deviate**
42	224	0.05	1.96
43	297	0.05	1.96
44	46	0.05	1.96
45	10,000	0.05	1.96
46	960	0.05	1.96
47	2166	0.05	1.96
48	502	0.05	1.96
49	707	0.05	1.96
50	1030	0.05	1.96
51	100	0.05	1.96
52	70	0.05	1.96
53	150	0.05	1.96
54	212	0.05	1.96
55	428	0.05	1.96
56	50	0.05	1.96
57	250	0.05	1.96
58	500	0.05	1.96
59	88	0.05	1.96
60	289	0.05	1.96
61	1500	0.05	1.96
62	600	0.05	1.96
63	750	0.05	1.96

64	1000	0.05	1.96
65	600	0.05	1.96
66	600	0.05	1.96
67	480	0.05	1.96
68	600	0.05	1.96
69	300	0.05	1.96
70	490	0.05	1.96
71	224	0.05	1.96
72	1800	0.05	1.96
73	465	0.05	1.96
74	300	0.05	1.96
75	328	0.05	1.96
76	885	0.05	1.96
77	372	0.05	1.96
78	147	0.05	1.96
79	370	0.05	1.96
80	1014	0.05	1.96
81	500	0.05	1.96
82	70	0.05	1.96
83	202	0.05	1.96
84	965	0.05	1.96
S/N	**Sample size**	**P value**	**Standard Normal Deviate**
85	550	0.05	1.96
86	375	0.05	1.96
87	227	0.05	1.96
88	1500	0.05	1.96
89	520	0.05	1.96
90	573	0.05	1.96
91	650	0.05	1.96
92	33800	0.05	1.96
93	5	0.05	1.64
94	216	0.05	1.96
95	341	0.05	1.96
96	2063	0.05	1.96
97	350	0.05	1.96
98	3283	0.05	1.96
99	21156	0.05	1.96
100	3346	0.05	1.96
101	13309	0.05	1.96
102	2156	0.05	1.96
103	997	0.05	1.96
104	698	0.05	1.96
105	346	0.05	1.96
106	484	0.05	1.96

107	3000	0.05	1.96
108	391	0.05	1.96
109	294	0.05	1.96
110	127	0.05	1.96
111	315	0.05	1.96
112	325	0.05	1.96
113	172	0.05	1.96
114	1092	0.05	1.96
115	304	0.05	1.96
116	239	0.05	1.96
117	184	0.05	1.96
118	600	0.05	1.96
119	159	0.05	1.96
120	786	0.05	1.96
121	600	0.05	1.96
122	355	0.05	1.96
123	82	0.05	1.64
124	293	0.05	1.96
125	293	0.05	1.96
126	947	0.05	1.96
127	2412	0.05	1.96
128	400	0.05	1.96
129	203	0.05	1.96
130	3883	0.05	1.96

Table 10 shows that all sampled studies used the same probability levels (0.05) to test the significant level of the results. The probability levels were categorised into P> 0.05 and P<0.05 alongside individual studies. The implication is that there is no basis for comparison since the probability levels of the selected studies are the same.

Table 11

Probability Levels of Sampled Studies

Chi-Square Tests			
	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-square	1.220	1	.001
Likelihood Ratio	158.579	1	.003
No of Valid Cases	130		

The statistical significance of the probability level was obtained from a Chi-square (χ^2) value computed on SPSS. The result showed no significant difference in the probability levels of the studies (χ^2=1.220, p>0.05). Therefore, the null

hypothesis, which states that there is no significant difference in the probability levels of the studies, is accepted.

DISCUSSION

From the study, it can also be said that the magnitude of the effect size of the sampled studies is moderate according to Cohen's interpretation of moderate effect size (0.5-0.7). Since effect size is the magnitude of the relationship between variables, it can thus be said that there is a relatively average relationship between the selected studies on validating items. This result agreed with Erisen and Gunay (2015), who collected correlation coefficients from published and unpublished meta-analyses of the effectiveness of doctoral dissertations on constructivist learning. The coefficients were analysed to determine if their magnitude was related to student's academic achievement and constructive learning. Also, the effect sizes of sampled studies identified ranged from 0.41 to 0.98 based on transformations. In other words, the results of this study showed that there is a high relationship between the variables of studies (10 (0.980), 41 (0.98) and 64 (0.97)). Additionally, the study identified a low relationship between the variables of studies (120 (0.41), 130 (0.45), 116 (0.52)). Furthermore, the forest plot shows the effect size estimate of each study. It is represented by a square box proportional to the weight assigned to a particular study. The weight is assessed based on study variances. Studies 10, 41, 45, 64, 96 and 130 pulled high weight. Within the study estimate runs the horizontal line, which is the same as the width of the 95% confidence interval. The effect size of sampled studies ranged between 0.41 - 0.98. In other words, the results identified a low relationship between the variables of studies (120 (0.41) and a very high relationship between the variables studies 10 and 41 with an effect size of 0.98 each. Additionally, Fishers Zr was involved in interpreting effect size to avoid complications of the extremity of high and low correlations (Liu et al., 2021). Moreover, all the primary studies adopted 0.05 as their probability level in deciding their null hypothesis. This means that all studies adopted similar or the same probability levels. The results of this study support the findings of Greenland et al. (2016) that the probability levels of studies do not differ significantly. There was no serious basis for comparison. In order to determine whether the 130 studies on validation procedures differed significantly in their effect sizes, they are classified into low, medium and very high effect sizes. The effect sizes of sampled studies identified ranged from 0.47 to 0.98 based on transformations. In other words, the results of this study showed that there is a high relationship between variables of studies, that is, studies 10 (0.980), 41 (0.98) and 64 (0.97). Comprehensive Meta-Analysis (CMA) software reported a correlation between Effect size 'r' and Fisher 'z' to clarify the effect sizes.

CONCLUSION

This study examined the effect size and homogeneity of standardised tests in education, analysing 130 empirical studies from 1988 to 2017 using meta-analytical techniques. The findings revealed an inverse relationship between sample size and effect size, aligning with Cohen's assertion that smaller samples often produce inflated effect sizes. A high Q statistic (24940.437), I2 value (99.48%), and Tau2 greater than zero indicated significant heterogeneity among the studies. Effect sizes were categorised as large, medium or small, with many studies exhibiting strong relationships between variables. The mean Fisher Zr (1.44103) and corresponding effect size (r = 0.89) confirmed a moderate overall effect. Probability level analysis showed uniformity across studies, with no significant differences in statistical thresholds. These results highlight the importance of sample size considerations and standardised methodologies in educational assessments. The study underscores the necessity of refining effect size estimation techniques to enhance the reliability and generalizability of research outcomes. Future research should address heterogeneity challenges to improve the validity of standardised test evaluations.

REFERENCES

Abu Bakar, M., Umroh, K. A., & Hameed, F. (2023). Improving Quality Islamic Education for Today's Generation. At-Tadzkir: Islamic Education Journal, 2(2), 118–128. https://doi.org/10.59373/attadzkir.v2i2.42

Arif, M., & Aziz, M. K. N. A. (2023). Islamic Religious Education Learning Model in the 21st Century: Systematic Literature Review. Indonesian Journal of Islamic Education Studies (IJIES), 6(2), 237-262. https://doi.org/10.33367/ijies.v6i2.4417

Arif, M., Chapakiya, S., & Dewi, A. Y. (2024). Character Education in Indonesia Islamic Elementary Schools: A Systematic Literature Review (2014-2024). J-PAI: Jurnal Pendidikan Agama Islam, 11(1).

Arif, M., Dorloh, S., & Abdullah, S. (2024). A Systematic Literature Review of Islamic Boarding School (Pesantren) Education in Indonesia (2014-2024). Tribakti: Jurnal Pemikiran Keislaman, 35(2), 161-180. https://doi.org/10.33367/tribakti.v35i2.5330

Barker, T.H., Habibi, N., Aromataris, E., Stone, J.C., Leonardi-Bee, J., Sears, K., Hasanoff, S., Klugar, M., Tufanaru, C., Moola, S. & Munn, Z. (2024). The revised JBI critical appraisal tool assesses the risk of bias for quasi-experimental studies. JBI Evidence Synthesis, 22(3), 378–388.

Carroll, L. M., Morris, M. E., O'Connor, W. T., & Clifford, A. M. (2020). Is aquatic therapy optimally prescribed for Parkinson's disease? A systematic review and meta-analysis. Journal of Parkinson's disease, 10(1), 59–76.

Çoğaltay, N., & Karadağ, E. (2015). Introduction to meta-analysis. Leadership and organisational outcomes: Meta-analysis of empirical studies, 19-28.

Cook, B. G., Cook, L., & Therrien, W. J. (2018). Group–difference effect sizes: Gauging the practical importance of findings from group–experimental research. Learning Disabilities Research & Practice, 33(2), 56-63.

Dattalo, P. (2013). Analysis of multiple dependent variables. Oxford University Press.

Erisen, Y., & Gunay, R. (2015). A meta-analysis into the effectiveness of doctoral dissertations on constructivist learning. The Anthropologist, 21(1-2), 202-212.
Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. Personality and individual differences, 102, 74-78.
Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European journal of epidemiology, 31(4), 337–350.
Gurung, R. A., Chick, N. L., & Haynie, A. (Eds.). (2023). Exploring signature pedagogies: Approaches to teaching disciplinary habits of mind. Taylor & Francis.
Huda, M., Arif, M., Rahim, M. M. A., & Anshari, M. (2024). Islamic Religious Education Learning Media in the Technology Era: A Systematic Literature Review. At-Tadzkir: Islamic Education Journal, 3(2), 83–103. https://doi.org/10.59373/attadzkir.v3i2.62
Kadirhanoğulları, M. K. (2023). The Effect of Teaching with Concept Maps on Academic Success in Biology Teaching: A Meta-Analysis Study. Dokuz Eylül Üniversitesi Buca Eğitim Fakültesi Dergisi, (58), 2781-2796.
Komalasari, M., & Yakubu, A. B. (2023). Implementation of Student Character Formation Through Islamic Religious Education. At-Tadzkir: Islamic Education Journal, 2(1), 52–64. https://doi.org/10.59373/attadzkir.v2i1.16
Kulinskaya, E., Hoaglin, D. C., Bakbergenuly, I., & Newman, J. (2021). AQ statistic with constant weights for assessing heterogeneity in meta-analysis. Research Synthesis Methods, 12(6), 711–730.
Liu, H., Li, T. W., Liang, L., & Hou, W. K. (2021). Trauma exposure and mental health of prisoners and ex-prisoners: A systematic review and meta-analysis. Clinical Psychology Review, 89, 102069.
Mellinger, C., & Hanson, T. (2016). Quantitative research methods in translation and interpreting studies. Routledge.
Mengist, W., Soromessa, T., & Legese, G. (2020). Method for conducting systematic literature review and meta-analysis for environmental science research. MethodsX, 7, 100777.
Mikolajewicz, N., & Komarova, S. V. (2019). Meta-analytic methodology for basic research: a practical guide. Frontiers in Physiology, 10, 203.
Nahar, L. (2023). The effects of standardised tests on incorporating 21st-century skills in science classrooms. Integrated Science Education Journal, 4(2), 36–42.
Nuijten, M. B., Van Assen, M. A., Augusteijn, H. E., Crompvoets, E. A., & Wicherts, J. M. (2020). Effect sizes, power, and biases in intelligence research: A meta-meta-analysis. Journal of Intelligence, 8(4), 36.
Ovute, A. O. (2015). Meta-Analysis of Research Findings on Influence of School Location on Students' Achievement in Mathematics. American-Eurasian Journal of Scientific Research, 10(1), 18-21.
Paramole, O. C. (2021). A Meta-Analysis of Procedures for the Validation of Standardised Tests in Education (1988-2017) (Doctoral dissertation, Obafemi Awolowo University).
Paramole, O. C., & Afolabi, E. R. I. (2024). Meta-Analysis of Item Generation Procedures Used in Selected Standardised Tests in Education. Jurnal Pendidikan: Riset dan Konseptual, 8(4), 679-689.

Perdices, M. (2018). Null hypothesis significance testing, p-values, effects sizes and confidence intervals. Brain Impairment, 19(1), 70–80.

Pigott, T. D., & Polanin, J. R. (2020). Methodological guidance paper: High-quality meta-analysis in a systematic review. Review of Educational Research, 90(1), 24-46.

Pomroy, P. (2020). An investigation into the challenges faced by international school teachers when interpreting the results from standardised tests (Doctoral dissertation, Durham University).

Ratnah, Ali Shah, S. A., & Alam, M. (2024). Integrating Religious Moderation into Islamic Religious Education: Strategies and Impacts. At-Tadzkir: Islamic Education Journal, 3(2), 120–133. https://doi.org/10.59373/attadzkir.v3i2.67

Rohmadiyah, B., Zamroni, M. A., & Ismawati. (2024). Principal Strategies in School Management at the State Vocational High School. Kharisma: Jurnal Administrasi Dan Manajemen Pendidikan, 3(1), 1–15. https://doi.org/10.59373/kharisma.v3i1.43

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. Frontiers in Psychology, 10, 813.

Simpson, A. (2017). The misdirection of public policy: Comparing and combining standardised effect sizes. Journal of Education Policy, 32(4), 450–466.

Soni, K. D. (2023). When do we need to do a meta-analysis? Indian Journal of Anaesthesia, 67(8), 673–674.

Tanha, K., Mohammadi, N., & Janani, L. (2017). P-value: What is and what is not. Medical journal of the Islamic Republic of Iran, 31, 65.

Tellinghuisen, J. (2022). Goodness-of-Fit tests in calibration: are they any good for selecting least-squares weighting formulas? Analytical Chemistry, 94(46), 15997-16005.

Veroniki, A.A., Jackson, D., Bender, R., Kuss, O., Langan, D., Higgins, J.P., Knapp, G. & Salanti, G. (2019). Methods to calculate uncertainty in the estimated overall effect size from a random-effects meta-analysis. Research synthesis methods, 10(1), 23-43.

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: a meta-analysis. Psychological bulletin, 140(4), 1174.

Zaini, M., Barnoto, B., & Ashari, A. (2024). Improving Teacher Performance and Education Quality through Madrasah Principal Leadership. Kharisma: Jurnal Administrasi Dan Manajemen Pendidikan, 2(2), 79–90. https://doi.org/10.59373/kharisma.v2i2.23 (Original work published June 12, 2023)

Zhao, Y., Li, T., & Liu, W. (2023). The Benefits and Drawbacks of Standardized Curriculum in Education. Research and Advances in Education, 2(10), 41–47.

Zubar, F. A. (2014). A Meta-Analytic Assessment of Empirical Differences on Standard setting Procedures in Public Examinations. Unpublished PhD Thesis, Obafemi Awolowo University, Ile-Ife.