Research Papers

Statistical Audit of Data Analysis of Educational Researches

S. K. TYAGI*

ABSTRACT

The present study is a part of the research project approved and supported by ICSSR, New Delhi, under its Senior Fellowship Programmme. 126 Ph.D. and 11 M.Phil. theses submitted to five universities of the five North-West Indian States were examined with a view to identify dominant statistical practices adopted by the researchers in education. The study revealed that 56.9 per cent of the research studies used t-test, while 33.6 per cent employed Analysis of Variance technique to test the hypotheses. Thus, over 90 per cent of the researches used either t-test or Analysis of Variance (ANOVA) in particular. However, it was found that in 51.3 per cent applications of the total t-tests and 21.7 per cent of the total ANOVA uses were inappropriate. Next most preferred (29.9 per cent) practice comprised of testing relationships using coefficient of correlations technique. Furthermore, the assumptions underlying these techniques were seldom tested. Consequently, only one researcher was found to have used non-parametric test of hypothesis. There was just one study wherein the researcher has reported effect size for each of the factors without interpreting it. The findings of the study could serve as a useful guide for chalking out need-based intervention programmes for researchers in education.

सार

वर्तमान अनुसंधान भारतीय सामाजिक विज्ञान अनुसंधान परिषद्, नई दिल्ली द्वारा अनुमोदित और समर्थित वरिष्ठ फेलोशिप कार्यक्रम के अंतर्गत परियोजना का एक भाग है। शिक्षा के क्षेत्र में शोधकर्ताओं द्वारा अपनाए गए प्रमुख सांख्यिकीय प्रथाओं की पहचान करने के लिए 126 पी.एच.डी. तथा 11 एम.फील. शोध प्रबंधों की जांच की गई जो कि उत्तर पश्चिम

* S. K. Tyagi, *ICSSR Senior Fellow, Former Head and Dean,* Faculty of Education, Devi Ahilya Vishwavidyalaya, Indore – 452001 (e-mail: sktyagi.sk@gmail.com).

भारतीय राज्यों के पाँच विश्वविद्यालयों में प्रस्तुत किए गए थे। अध्ययन से ज्ञात हुआ है कि 56.9 प्रतिशत शोध अध्ययनों में t.परीक्षण का उपयोग किया गया, जबकि 33.6 प्रतिशत ने ANOVA तकनीक को परिकल्पना का परिक्षण करने के लिए उपयोग किया। इस प्रकार 90 प्रतिशत से अधिक शोधों में विशेष रूप से t.परीक्षण या ANOVA का उपयोग किया गया है। हालांकि यह पाया गया कि कुल t.परीक्षण के 51.3 प्रतिशत और ANOVA के 21.7 प्रतिशत उपयोग अनुचित थे। इन परीक्षणों के अतिरिक्त सहसबंध गुंणाक की गणना अत्यधिक (29.9 प्रतिशत) प्रयुक्त तकनीक है। परंतु इन तकनीकों में अंतर्निहत मान्यताओं का शायद ही कभी परीक्षण किया गया। जिसके फलस्वरूप केवल एक शोधकर्ता ने परिकल्पना की जांच हेतु गैर-पैरामीट्रिक परीक्षण का उपयोग किया था। केवल एक शोधकर्ता ने प्रभाव विस्तार का उपयोग अपने अध्ययन में किया है, हालाँकि इसकी व्याख्या नहीं दी है। अध्ययन के निष्कर्ष, शिक्षा में शोधकताओं के लिए आवश्यकता आधारित हस्तक्षेप कार्यक्रमों को चलाने के लिए उपयोगी मार्गदर्शिका के रूप में उपयोगी साबित हो सकते हैं।

Introduction

Quality of higher education provided in the colleges and universities depends to a great extent on the quality of research being conducted by the faculty and students. Quality of research in turn is largely influenced by the knowledge and skill base possessed by the research practitioners. The University Grants Commission (UGC), an institution mandated by the Government of India for maintaining standards in higher education, contemplated various steps to enhance standards in research practices at the university level. For these purposes, UGC through its regulation on Minimum Standards and Procedure for the award of M. Phil. or Ph.D. degree (2009) has stipulated a mandatory six-monthly pre-Ph.D. Coursework in research methodology. Simultaneously, faculty guiding research scholars are required to upgrade their knowledge of research methods through various empowerment programmes, refresher courses and other online courses and Massive Open Online Courses (MOOCs) under Ministry of Human Resource Development portal, SWAYAM.

However, for ensuring the effectiveness of the designed academic interventions, the above provisions need to be tailor-made to the emergent needs of the research practitioners. Need assessment exercise can either be based on the analysis of the curricula offered at the universities for preparing would-be researchers or by an analysis of the doctoral studies produced by

the universities. Hence, an analysis of the statistical data analysis practices followed by the researchers was planned, which could serve the blue print of any targeted intervention programme for research practitioners in education.

Not many researches focusing at analysis of research methodology, particularly their statistical procedures, are available in literature. Among the three studies at which the hands could be laid, Keselman et al. (1998) examined several articles published in the prominent US educational journals and found that researchers rarely verify validity assumptions, use analysis that are typically non-robust to assumption violations, rarely report effect size statistics and hardly perform any power analysis. Elmore and Woehlke (1996) analysed all published articles appearing in the three top US journals from 1978 to 1995 and concluded that most common methods used by researchers were ANOVA and ANCOVA, multiple regression, bivariate correlation, descriptive statistics, multivariate analysis, non-parametric tests and t-tests. Govil, et al. (2015) studied 20 Ph.D. theses submitted to Aligarh Muslim University (AMU), out of which 10 were from the Department of Education and remaining 10 from the Department of Psychology. The study concluded that errors committed by research scholars while using statistical methods were serious in nature. Only 79 per cent employed appropriate methods, 47 per cent tested the underlying assumptions, 36 per cent used insufficient statistical methods and none of the research reported effect size.

Clearly, the couple of studies undertaken abroad were taken up some twenty years ago. The only Indian study was on a miniscule level, based on just 20 studies from a single institution and included only ten studies from the field of education. There is definitely a need for such review studies as the present one.

Objectives of the Study

The objectives of the study were:

- to identify the dominant statistical data analysis practices of educational researches in Indian universities at Ph.D. or M. Phil. level.
- to assess the appropriateness of the practices in view of the research designs adopted, underlying assumptions and other desirable indicators.

Method

Sample

The four states from Hindi speaking belt of Northern India, namely Uttarakhand, Uttar Pradesh, Haryana and Himachal Pradesh, were randomly selected. After that, one university from each of the selected States was selected on the basis of following criteria:

- Having Department of Education in the University
- Availability of substantial amount of Ph.D./ M. Phil. work during the last five years

Additionally, Mahatma Gandhi Antarrashtriya Hindi Vishwavidyalaya, Wardha, being Hindi-medium Central University in the Western Indian State of Maharashtra was also included in the list. Thus, the following five universities in each of the five North-West Indian states comprised the sample for the present study.

1. Mahatma Gandhi Antarrashtriya Hindi Vishwavidyalaya (MGAHV), Wardha
2. Hemvati Nandan Bahuguna University (HNBU), Sri Nagar, Garhwal
3. Maharishi Dayanand University (MDU), Rohtak
4. Chaudhary Charan Singh University (CCSU), Meerut
5. Himachal Pradesh University (HPU), Shimla

Table 1 gives the number of Ph.D. or M. Phil. theses by university and year of submission.

Table 1

University and Year-wise Breakup of the Sampled Theses (N=137)

Year	MGAHV	HNBU	HPU	CCSU	MDU	Total
2011	-	-	-	6	4	10
2012	-	20	9	11	14	54
2013	-	5	11	3	8	27
2014	-	6	5	2	10	23
2015	-	3	2	2	3	10
2016	4	1	5	-	-	10
2017	3	-	-	-	-	3
Total	7*	35	32	24	39**	137

*M. Phil. theses ** Includes 4 M. Phil. theses

As is clear from Table 1, two universities namely, MGAHV and MDU accounted for 11 M. Phil. theses while 126 Ph. D. theses were submitted to the four universities. The entire lot of 137 researches was spread over a span of seven years i.e. from 2011 to 2017. However, about 76 per cent of the theses examined, referred to the three years period ranging from 2012 to 2014.

Procedure

All the selected theses were studied with a view to identify the practices related to quantitative data analysis. Parameters for classification of the studies and the associated considerations are presented below:

1. Selection of statistical test or tests used by researchers to test the formulated hypotheses: While classifying a research study on the basis of statistical tests used by the researchers, certain points were kept in mind. For instance, research studies using t-test as a follow up procedure to ANOVA were not included in the t-test category but were rather placed under the ANOVA category. Similarly, research studies using regression necessarily use the technique of coefficient of correlation, t-test and ANOVA. Nonetheless, such theses were put under regression only, and so on.
2. Appropriateness of the selected statistical test in view of the design of the study and objectives/hypotheses: Researches using series of t-tests for comparing more than two group-means as also those breaking up a factorial design into a several pair-wise comparisons were considered inappropriate practices as they inflate Type I error. In the same manner, research studies using t-tests separately at pre- and post-experiment stage or using paired samples t-tests separately for the two groups were placed under inappropriate category.
3. Statistics used for studying the assumptions underlying the selected tests: It was also examined whether the researchers have performed distributional checks to test the assumptions underlying the selected tests and the statistics employed by them for this purpose.
4. Post Hoc tests used in view of a significant F value: Frequencies of researchers following up a significant F-value with post analysis and the strategies of post-hoc comparisons used by them were observed.

5. Interpretation of significant interactional effect strategies: Practices used by the researchers in view of a significant interactional effect were identified and their frequencies noted.
6. Non-parametric statistics used by the researchers: The incidences of use of various non-parametric tests by the researchers along with the conditions of their use were recorded under the point.

Results

The collected information was tabulated in the form of a master sheet and later analysed on the basis of the parameters fixed for the analysis. The results of the analysis are being presented below under various captions.

Statistical Tests/ Measures Adopted by the Researchers: An Overview

All the research studies examined were categorised into one or more of the categories given in Table 2, according to the statistical tests and measures used for analysing the data.

Table 2

Statistical Tests Employed by Researchers or Data Analysis (N = 137)

Statistical Test	Frequency	Per cent
Descriptive	14	10.2
t-test	78	56.9
Paired	05	3.6
Independent	73	53.3
ANOVA	46	33.6
One way	15	11.0
Many way	30	21.9
ANCOVA	01	0.7
Non-Parametric Tests	11	8.0
Chi-square Test	09	6.6
Mann Whitney Test	01	0.7
Kruskal Wallis Test	01	0.7
Correlation	41	29.9
Multiple Regression	06	4.4
Factor Analysis	01	0.7

Note: *Total is greater than 137 (and more than 100%) as some of the studies having used different statistical tests were placed under more than one category.*

It is evident from Table 2 that 10.2 per cent of the researchers used only descriptive statistics like measures of central tendency, measures of variability and graphical representations to analyse/ represent data. A total of 78 theses comprising of 56.9 per cent of the sample employed t-test for the data analysis, 5 of these (3.6%) have been paired samples t-tests. Analysis of Variance was used by 46 researchers (33.6%). The number of researches using Many-way ANOVA, One-way ANOVA and One-way ANCOVA was 21.9 per cent, 11 per cent and 0.7 per cent respectively. Next most preferred statistical technique was coefficient of correlation, used in 29.9 per cent of the theses. Only 6 researchers used multiple regression technique, while factor analysis was employed by a solitary researcher.

Among non-parametric statistics, Chi-square test was used by 11 (8%) of the researchers, while Mann Whitney and Kruskal Wallis test were applied by one researcher each.

Appropriateness of the Statistical Test

The appropriateness of the selected statistical tests used by the researchers can be gauzed by the study of Table 3.

Table 3: Test Appropriateness

Statistical Test Used	Frequency	Per cent
t-test (N=78)		
Appropriate	38	48.7
Inappropriate	40	51.3
Test deemed Appropriate		
Two-way ANOVA	30	38.5
One-way ANOVA	04	5.1
Two-way ANCOVA	04	5.1
One-way ANCOVA	02	2.6
ANOVA (N=46)		
Appropriate	36	78.3
Inappropriate	10	21.7
Test deemed Appropriate		
Two-way ANOVA	05	10.9
One-way ANOVA	02	4.3
Two-way ANCOVA	03	6.5

Total = 124, Appropriate = 74 (59.7%), Inappropriate = 50 (40.3%)

Out of 78 researches using t-test, 40 (51.3%) were found inappropriate (Table 3). In about 38.5 per cent of the researches using t-test, the technique called for was two-way ANOVA as the researchers have used twin factors. Similarly, in about 5.1 per cent of the theses, the appropriate tests should have been two-way ANCOVA and one-way ANOVA, respectively. Finally, one-way ANCOVA was an appropriate test in case of two studies. In case of use of Analysis of Variance, 10 out of 46 (21.7%) ANOVA uses were inappropriate. In place of the ANOVA model used by the researcher, two-way ANOVA was called for in 5 instances, one-way ANCOVA in one instance and two-way ANCOVA in two cases. In all, out of total 124 researchers using t-test/ANOVA, 50 (40.3%) used them inappropriately. In 28.2 per cent of the cases, two-way ANOVA was called for. One-way ANOVA, two-way ANCOVA and one-way ANCOVA should have been used in 3.2, 5.6 and 3.2 per cent of the cases, respectively.

Testing Assumptions Underlying the Statistical Tests Employed for Data Analysis: The frequency of researchers following the practice of testing the assumptions underlying statistical tests applied is given in Table 4.

Table 4: Testing of Underlying Assumptions

t-test (N=78)		
Assumption	Frequency	Per cent
Normality	06	7.7
Homogeneity	04	5.1 (Using Descriptive Statistics)
Hartley's Test	03	3.8
Leven's Test	01	1.3
ANOVA (N=46)		
Normality	05	10.9 (Using Descriptive Statistics)
Homogeneity	08	17.4
Hartley's Test	05	10.9
Barlett's Test	02	4.3
Leven's Test	01	2.2
Correlation (N=41)		
Normality	04	9.8

Homogeneity	Nil	Nil
Linearity	Nil	Nil
Multiple Regression (N=6)		
Normality	03	50
Multicollinearity	Nil	Nil
Linearity	Nil	Nil

It can be observed (Table 4) that among those who used t-test, only 7.7 per cent took care to discuss normality of distribution using descriptive statistics like mean, median, skewness, kurtosis and graphs. However, none of the researchers did so for checking the normality of distributions at each level of the independent variable. With regard to variance homogeneity, only 5.1 per cent of the researchers applied statistical tests to check whether the data meets this requirement. Only 3 of the 78 studies employed Hartley's test and one used Leven's Test. In case of total ANOVA usage, 10.9 per cent of the researchers discussed normality of the underlying distribution using descriptive statistics. None of them have applied test of normality such as Kolmogorov-Smirnov test or Shapiro-Wilk test. As for the variance homogeneity, the incidence of testing the assumptions is 17.5 per cent. The tests used by the researchers are Hartley's Test, Bartlett's Test and Leven's Test in that order of preference. Next most popular statistical test used was bi-variate coefficient of correlation 'r'. While normality test was given due attention by 9.8 per cent of the researchers, none chose to test the variance homogeneity and linearity of relationship. Finally, with the exception of normality which was considered by half of those applying multiple regression, none cared to verify the nature of the data for violation of linearity, multi-collinearity and homogeneity. Likewise, the assumptions of ANCOVA and factor analysis were not paid due attention by the only researcher using these techniques each.

Use of Different Multiple Comparison Procedures (MCP)

In case of significant F-value for a factor with three or more levels, researchers have to further subject the analysis to pair-wise comparisons for identifying pairs with significant differences. Post-Hoc tests employed to compare groups pair-wise have been classified in Table 5.

Table 5

ANOVA: Post Hoc (Multiple Comparisons) Tests Employed

Post Hoc Test	Frequency	Per cent
t-tests	23	82.1
No test used	03	10.7
Bonferroni	01	3.6
Tukey's HSD	01	3.6
Total	28	100.00

As apparent from the Table 5, 82.1 per cent (23 out of 28 eligible cases) of the researches used t-tests, whereas none used Fishers' LSD (an approach using common error term). Another 10.7% left the analysis at that point, concluding that significant differences exist. Only one researcher each used Tukey's HSD and Bonferroni test which allowed making adjustments for increased alpha error in case of a family of hypotheses.

Strategies Used for Interpreting Significant Interactions

The strategies adopted by researchers to interpret a significant interactional effect denoted by FA×B are presented in the Table 6.

Table 6

Interpretation of Significant Interaction (A×B)

Strategy adopted	Frequency	Per cent
Multiple 't's	10	47.6
Bar graphs	02	9.5
Line graph	05	23.8
No follow up	04	19.1
Total	21	100.00

It can be seen that less than one-fourth (23.8% to be precise) used line graph to discuss interaction between two variables (Table 6). Interestingly, a large number of researchers (47.6%) used multiple t-tests in a bid to 'explain' interaction. About 9.5% researchers prepared bar graphs, while 19.1 per cent did not follow up a significant F-value for interaction at all.

Non-Parametric Tests Used by the Researchers

Frequencies of different non-parametric tests used for hypotheses testing have been listed in the Table 7 below.

Table 7

Non-parametric Statistics (N=11)

Test	Frequency
Chi-square	09
One Sample Case	05
Two Samples Case	04
2 ×2 Classification	02
2 ×3 Classification	01
2 ×4 Classification	01
Mann Whitney	01
Kruskal Wallis	01

It is clear from Table 7 that only 9 research studies applied Chi-square test. Out of these, only 5 have been one sample cases while the remaining ones have been two sample cases. None of the studies has followed residual analysis in view of a significant Chi-square value with more than two categories of factors.

Discussion

An overwhelming majority of 90.5 per cent of the researchers have been found using either t-test or ANOVA techniques for hypotheses testing. This concurs with the findings by Keselman et al. (1998), Elmore & Woehlke (1996) and Govil et al. (2015). Thus, sound practices related to these tests will go a long way in avoiding large number of invalid and misleading research findings, consequently improving the quality of research. However, ignoring the fact that comparison of several group means calls for the use of One-way ANOVA is not a healthy practice. Further, if one chooses to use multiple t-tests, he/she might run the risk of escalating Type I error, termed as family-wise error (FWE). In such cases, certain null hypotheses liable to be accepted will be erroneously rejected by the researchers.

The same applies to the multiple classifications of variables and use of ANOVA for testing the difference between means. If there are two factors with two or more levels each, it is a perfect situation for the use of 'Two-way ANOVA'. There is a tendency among researchers to break it into several pair wise comparisons using multiple t-tests rather than examining two main and interactional effects through omnibus ANOVA test. Interestingly, in one case, a researcher has applied two one-way ANOVA tests separately for boys and girls

to avoid using two-way ANOVA. In every one out of five ANOVA usages, researchers have first used multiple t-tests and afterwards claim to have 'verified' the results by applying Omnibus ANOVA Test. A particular case is instructive in which the researcher finds significant difference among one of the several pairs of t-values. Later, upon applying ANOVA she gets a non-significant F-value. She admits the fact as unfortunate not to get results of the t-test verified by the F-test!

In case of ANOVA, a significant F-value indicates that there are significant differences among several means. If the factor has just two levels, it is sufficient to observe the two mean scores and no further analysis is required. Using a t-test in such case is eventually a redundant practice if not an erroneous one. Post-Hoc tests or Pair-wise comparisons are made in case of three or more groups. Out of 28 cases, barring one researcher each who has used Tukey's HSD and Bonferroni test, all others have used t-tests for pair-wise comparisons. Thus, they have failed to take advantage of tests which can control inflation of alpha error and ensure that family-wise error does not exceed 0.05 or 0.01 levels of confidence. More exposure to softwares like SPSS and 'R' can perhaps facilitate researchers to adopt these strategies.

Exploration of the nature of data for checking the underlying assumptions is yet to develop as a healthy research practice among the educational researchers. Keselman et al. (1998) also reported that researchers rarely tested the validity assumptions. The occurrence of testing underlying assumptions ranges from a minimum of 5.1 per cent to a maximum of 17.4 per cent (for N>10). Not surprisingly, therefore, there are numerous cases of illegitimate application of statistical tests for analysis of the research data resulting in many misleading and invalid research findings.
Unwillingness to check whether the data meet the requirements of statistical tests happens to the prime cause of almost total disregard of the educational researchers for non-parametric tests such as Mann-Whitney test, Wilcoxon matched pairs test and Kruskal Wallis Test by ranks.

Barring one instance, none of the researchers in education happened to use strategies to manage violation of assumption of homogeneity of variance like Welch Test, Brown-Forsythe Test or any other adjustment mechanism. Likewise, applying data transformation to deal with cases of non-normality seems to be a phenomenon not yet belonging to the realm of educational research.

Obviously, unless checking basic assumptions like normality of distribution is followed as a mandatory practice, researchers may afford to remain blissfully unaware of the data transformation practices in educational research.

The statistical significance of an observed value of a test need not imply its practical significance. If sample size is quite large, a very low value of the test statistic may turn out to be statistically significant which may not have any practical importance. For example, studies have reported coefficient of correlation as low as 0.1 to be significant, which means only 1 per cent common variance (r2 value) are being shared by the two variables. The indicator of practical significance of a test value is termed as 'effect size'. Only one of the sampled theses has reported effect size—that too without any interpretation. Govil et al. (2015) too found none of the researchers reporting the effect size, while Keselman et al. (1998) also observed it to be rarely reported in studies. It may be noted that Cohen's 'd', partial eta squared and coefficient of determination (r2) are some of the effect size measures that can be used in case of t-test, ANOVA and coefficient of correlation techniques, respectively. Lack of awareness about these measures coupled with shyness to use statistical software like SPSS might constitute the prime reasons for the trend.

With the exception of a solitary researcher, all others using parallel group pretest design have followed bizarre practices like comparing pre-test scores of experimental group with post tests scores of control group and vice-versa. Similarly, one also finds researchers applying correlated t-test separately for the experimental and the control groups and even concluding that both the treatment and conventional methods have been found effective. Also, comparison of both groups at pre and post-stage separately is quite common among researchers. The ill practices can perhaps be attributed to the lack of knowledge regarding ANCOVA and tendency to reduce any situation to familiar two-group comparisons.

The most popular non-parametric statistical test used by the educational researchers was Chi square test. Tests like Kruskal Wallis test and Mann Whitney 'U' test have been employed by only one researcher each, out of a total 11 researchers. However, none has taken recourse to residual analysis to locate significant differences within cells of the contingency table. A significant Chi square value needs to be subjected to further analysis akin to post-hoc tests in ANOVA.

The study has significant implications for planning a need-based curriculum of a Pre-Ph.D. Coursework or intervention programs to improve the data analysis skills of research practitioners in education, eventually rising in the quality of research in education.

References

ELMORE, P. B., AND D.P.L. WOEHLKE. 1996. *Research methods employed in American educational research journal, educational researcher and review of educational research from 1978 to 1995.* Paper presented at the Annual Meeting of the American Educational Research Association, New York.

GOVIL, P., M., A.N. QASEM AND S. GUPTA. 2015. *Evaluation of statistical methods used in Ph.D. theses of social science in Indian universities. International Journal of Recent Scientific Research,*vol. 6, No. 3, pp. 2926–2931.

KESELMAN, H. J., C.J. HUBERTY, L.M. LIX, S. OLEJNIK, R.A. CRIBBIE, B.R. DONAHUE, K. KOWALCHUK, L.L. LOWMAN, M.D. PETOSKEY, J.C. KESELMEN, AND J.R. LEVIN. 1998. *Statistical practices of educational researchers: An analysis of the ANOVA, MANOVA and ANCOVA Analysis.* Review of Educational Research, vol. 68, No. 3, pp. 350–386.

UNIVERSITY GRANTS COMMISSION. 2009. *UGC (Minimum Standards and Procedure for the Awards of M. Phil. / Ph. D. Degree) Regulations.* New Delhi: UGC, pp.4053-54.