Przegląd Badań Edukacyjnych
Educational Studies Review

ISSN 1895-4308
nr 49 (1/2025), s. 163–178

THE METHODS OF DATA COLLECTION
AND ANALYSIS
IN EDUCATIONAL RESEARCH

METODY ZBIERANIA
I ANALIZY DANYCH
W BADANIACH EDUKACYJNYCH

Arkadiusz Prajzner

University of the National Education Commission, Krakow, Poland
e-mail: arkadiusz.prajzner@uken.krakow.pl
ORCID: https://orcid.org/0000-0002-8681-9636

Beyond False Positives and Negatives. Understanding Type III and IV Errors in Social Research

http://dx.doi.org/10.12775/PBE.2025.010

Abstract

Research methodologies that correlate acquired measurement data with critical values defining the null hypothesis, despite their widespread application, remain susceptible to various inference inaccuracies. The decision to reject or fail to reject the null hypothesis can result in type I or type II errors, which can undermine the validity of conclusions. A comprehensive understanding of these errors is crucial for ensuring the accuracy of inferences drawn in social research. Proper definition and control of variables are essential in minimizing the risk of incorrect conclusions. Moreover, examining the broader research context, beyond just statistical probabilities, reveals the potential for additional logical errors (Kimball, 1957). These errors, classified as type III and type IV, have been examined in prior research (Scanlon et al., 1977). This paper seeks to provide an overview of these error types and their definitions. The study identifies several causes of these errors in the conducted analysis, including the selection of statistical methods, the choice of testing area (type III errors), and the neglect of violated assumptions or misinterpretation of interaction effects (type IV errors). The underlying causes of these errors are traced to the research methodology itself, such as improper operationalization, incomplete knowledge of the process under study (type III), and failure to account for study limitations (type IV). In addition, this paper proposes strategies to minimize the occurrence of type III and type IV errors, emphasizing their potential impact and providing recommendations to mitigate these risks.

Keywords: type III error, type IV error, social research methodology.

Null hypothesis statistical testing

In social sciences, conclusions are derived by applying both quantitative and qualitative research methodologies (Józefacka et al., 2023). Despite the evident distinctions between these approaches, such as sampling techniques and data collection methods (Marvasti, 2004; Mehrad & Zangeneh, 2019), their overarching objective remains consistent – to enhance understanding of the world (Asper & Corte, 2019, p. 145). In recent years, the dominance of quantitative methods has been increasingly questioned, with mixed-method research designs gaining prominence as a complementary alternative (Xiong, 2022). Despite this shift, social science studies continue to demonstrate a strong reliance on quantitative data and its statistical analysis (Habes & Pasha, 2021). This paper focuses on addressing errors commonly associated with the quantitative approach.

The process of data collection inherently requires researchers to engage in statistical analysis. The probabilistic framework introduced by Bayes (1763) has long been used in social science data analysis (Field, 2018). Unlike classical statistical methods, Bayesian statistics diverts attention from significance testing, emphasizing probabilistic reasoning as the basis for hypothesis evaluation. While the comparative strengths and limitations of classical and Bayesian approaches lie beyond the scope of this paper, comprehensive discussions on these topics can be found in works by Krusche (2014), McElreath (2016), and Field (2018). Despite alternatives in quantitative data analysis, such as the cited example of Bayes statistics, the classical approach to significance assessment dominates scientific papers. Not without reason that one of the *spine* areas of statistics in Field's (2018) assessment is the letter N, which corresponds to the null hypothesis significance testing. Furthermore, the application of a hypothesis-challenging system, combined with an assessment of its probability of being true, has emerged as a prevalent approach in social studies to address research questions (Vascha-Haase & Thompson, 2004; Henson, 2006; Sun, Pan & Wang, 2010). The resulting p-value provides researchers with an indication of the probability of obtaining a situation that is identical to that described in the null hypothesis in a sample when the null hypothesis is true (King & Minium, 2020). A low p-value (less than 0.05) indicates that, in less than 5% of cases, a sample result may deviate to such an extent from what is described in the null hypothesis, given that the null hypothesis is true. Consequently, the significance of the result obtained in the test leads to the conclusion that the null hypothesis should be rejected. A high p-value obtained in the analysis (greater than 0.05) indicates that there is a high probability that the result in the sample does not differ significantly from the situation described in the null hypothesis (Brzeziński, 2021). At this point, it should be clarified that the non-significant results obtained in the study can only be interpreted as a lack of evidence to reject the null hypothesis. Purists would also point out that the null hypothesis can never be completely false.

The concept of statistical significance and its evaluation in research was fundamentally shaped by Fisher (1925). This approach consists of several key stages: first, selecting an appropriate statistical test based on the study's objectives; second, formulating the null hypothesis; third, calculating the p-value; and finally, interpreting the result. In the initial presentation of statistical tables, Fisher employed values of 0.05, 0.02, and 0.01 as reference points (Field, 2018, after Fisher, 1925). Initially regarded as a criterion, the result was subsequently utilized in research, with a recommendation to continue using it, even when more accurate tables were presented. Presented as a proposition, Fisher's levels of significance do not need to be rigid. A low p-value obtained in the study provides strong evidence against the null hypothesis. The greater the p-value, the weaker the evidence to reject the null hypothesis (Perezgonzalez, 2015). Despite its flexibility and suitability for exploratory research, Fisher's approach has sparked debate. Critics have pointed out several issues: a lack of correction for multivariate comparisons (MacDonald, 1997), significant interference from multiple factors (Hubbard, 2004), an absence of power assessment (Kruskal, 1980), and the lack of an alternative hypothesis (Gigerenzer, 2004) to explain the results.

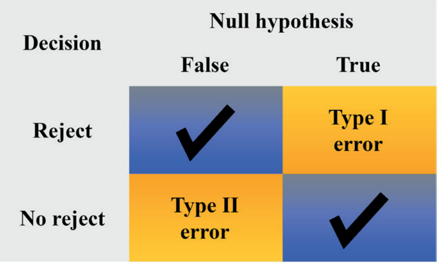
The Neyman-Pearson approach, which shares some similarities with Fisher's proposal in certain areas, addresses some of the objections raised. The more mathematical Neyman-Pearson approach also emphasized the (*a priori*) project planning stage. A notable distinction between the two approaches lies in the formulation of hypotheses: the Neyman-Pearson approach involves formulating two hypotheses: the null and the alternative. While statistical data analysis still provides evidence for rejecting the null hypothesis, as in Fisher's approach, it also offers additional insight into the results through the alternative hypothesis. The criticism of the weak statistical and mathematical foundations of inference in the Fisher model has found many solutions in the Neyman-Pearson model (Pérez-González, 2015). Indeed, it was proposed that the effect size expected in the population under the alternative hypothesis should be accounted for during the planning stage of a study. The construct proposed by Neyman and Pearson (1928a) and subsequently developed by Cohen (1988) was intended to account for the magnitude of the relationship between the constructs the study was designed to confirm (detect). Furthermore, Neyman and Pearson indicated that the statistical test's power to detect a given effect should be determined at the study's planning stage. It was emphasized that hypotheses should be clearly formulated, the required sample size to detect a specified effect should be determined, and the critical values for the statistical test should be established as part of planning the study.

The necessity to specify the power of the test, the sample size, and the effect prior to analyzing the hypotheses enabled control of type I and type II errors, a more extensive description of which is presented later in this paper. Neyman and Pearson's approach reduced several problems by becoming a more accurate and deductive approach. However, it has not escaped criticism directed at the approach being less flexible. It is not always possible, especially in exploratory studies, to plan studies that estimate, for example, population effect sizes (Perezgonzalez, 2015).

Type I and II errors

Nevertheless, the decision to reject or not to reject the null hypothesis is one of the most pivotal aspects of the statistical analysis, with ramifications for subsequent conclusions. This determination is based on the alpha and beta values, also known as the probabilities of errors of the first and second type.

They represent situations where an inaccurate decision is made regarding the acceptance (or rather failure to reject being fully correct) or rejection of the null hypothesis.



Graph 1. Type I and II errors in addiction to null hypothesis reject decision

Source: Brzeziński, 2021, p. 208.

It is evident that rejecting the null hypothesis when it is false and failing to reject it when it is true are well-founded and appropriate decisions, as demonstrated in Graph 1. It follows that an approach based on the results of a given study can be considered to be correct and thus can facilitate the generation of valid and accurate conclusions.

Rejecting a true null hypothesis

The research situation in which the true null hypothesis is rejected, if referred to as a type I error. Subsequent inference about correlations or differences in the population may be compromised and incorrect. The probability of a type I error is used by researchers to offset the risk of this error. A common threshold for alpha is set to 0.05, meaning that researchers typically accept a 5% chance of committing a type I error, although its selection was quite random. Both Fisher and Neyman-Pearson approaches provided a choice in alpha values for researchers and similarly related the statistical test results obtained to a type I error (Perezgonzalez, 2015). A p-value equal or less than alpha is considered statistically significant, and it indicates the rejection of the null hypothesis (Field, 2018). Otherwise, type I errors may occur.

Failing to reject a false null hypothesis

Not rejecting a false null hypothesis is described as a type II error. Detecting and indicating that a hypothesis is in fact false requires that the test and the selected statistical method possess the required "capability" referred to as a test power. The greater the statistical test's power, the greater the probability of the detection of existing differences or correlations (Field, 2018). As the statistical power of a test increases, the probability of committing a false positive (i.e., a type II error) declines. This relationship can be represented as follows: 1-test power = beta (Ioannidis, 2005). In empirical research, considering Neyman-Pearson approach, the value of test power and associated type II error are determined by a number of factors:

a) The magnitude of the effect that the study is attempting to demonstrate. Stronger effects are more readily discernible.
b) Significance level in the analysis. As the threshold for identifying a significant result becomes increasingly restrictive, the power of the test diminishes. A more rigorous approach in evaluating significance, coupled with tighter constraints on the alpha level, decreases the likelihood of characterizing an effect as significant (Ioannidis, 2005). An unjustified reduction in the alpha level (the probability of the first type of error) also increases the probability of committing the second type of error – rejecting the null hypothesis in the absence of an effect.
c) Number of observations. The introduction of additional data results in a reduced level of measurement error, which facilitates the identification of the effect. Filed (2018) draws a parallel between the indicated test power factor and the number of observations, comparing them with radio. The calculated statistics can be described as a signal-to-noise ratio, whereby a larger sample results in a lower level of noise (measurement error), thus facilitating the observation of a stronger signal (effect).
d) Type of statistical test. Compared to non-parametric tests, parametric tests possess greater statistical power (Perezgonzalez, 2015). Selecting a one-tailed test for a directional hypothesis enhances statistical pow-

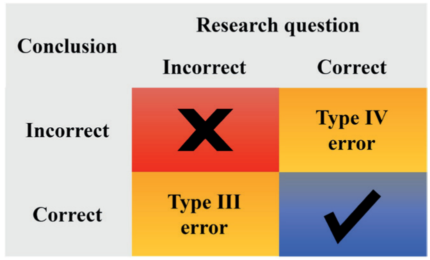
er and reduces type II errors. Consequently, planned comparisons and a priori approaches have an advantage over omnibus tests and post hoc analyses due to their greater statistical power.

e) Distribution of the variable. Unsourced variance in the research design can result in difficulties in achieving a normal distribution. An uncontrolled increase in measurement variance can diminish the power of the test, leading to a larger type II error (King & Minium, 2020). The failure to adequately assess the distribution of results can also lead to an erroneous selection of statistical tests, thereby increasing the risk of type II error (Uttley, 2019).

Researchers and students are frequently well acquainted with both type I and II errors. Awareness of the risks associated with both types of errors, and the use of contemporary data analysis methods make it more challenging for them to occur. Despite the notable attention devoted to the prevention of these errors (especially of the first type), the research produced, and the conclusions drawn may still be susceptible to other forms of shortcomings.

Type III and IV errors

Kimball identified issues that extended beyond the scope of a straightforward evaluation of the null hypothesis in his work as early as 1957, defining these as errors of the third and fourth types. The occurrence of these errors is contingent on the research question posed in the study and the resulting conclusions. As illustrated in Graph 2, a study that poses an erroneous research question and presents an incorrect answer is a clear indication of a lack of quality in the research paper. This ultimately precludes the ability to draw reasonable conclusions. Nevertheless, when the study contains a correctly designed question and presents an appropriate answer to that question, errors of type III and IV are avoided. At the same time, this allows a discussion to commence about the decision to reject or not to the null hypothesis. Consequently, providing a correct question or a suitable response does not negate the possibility of errors of types I and II (Tate, 2022).



Graph 2. Type III and IV errors in addiction to research question and given conclusion

Source: Author's research.

The definition presented by Kimball (1957) describes an error of type III. It occurs when the correct answer is presented to a poorly posed question. Conversely, a study where the incorrect answer is formulated to a well-posed question is described by the author as a type IV error. Mostly attributable to the survey methodology and data analysis, type I and type II errors can be distinct from those of type III and type IV, which originate from statistics and research design. Errors in the implementation and evaluation of research conclusions stem from logical fallacies. When evaluating the correctness of formulated arguments in a discussion, the logician also describes logical fallacies committed by the participants. A correct conclusion is guarded against formal and material errors. Material errors are defined as inadequacies of the conclusion (Kublikowski, 2011) and occur when the definition is either too narrow or too broad. Similarly, inadequate inferences may occur preceded by a false premise (Suchoń, 2005). Inadequacy may arise in operationalizing variables, defining study areas, and selecting appropriate data analysis methods. Such inadequacies may arise from inconsistencies between the defined variable and the instrument used to measure it, or between the purpose of the study and the chosen method of data analysis. As inconsistency increases, the inadequacy of the research question and the posed conclusion exacerbates implementation problem, the type III error. Formal, translational errors in deductive reasoning occur due to an absence of connection between the premise and the thesis, often indicated by a lack of logical implication (Suchoń, 2005). This lack of relationship could result from a failure to take into account the study limitations, an attempt to over-estimate its conclusions, or poor data evaluation methods. Hasty conclusions regarding cause-and-effect relationships, which overlook the limitations of the research design, as well as erroneous decisions pertaining to data analysis process, can result in challenges related to the translatability of the study and its evaluation, a type IV error.

It is important to note that the logical fallacies outlined in this study are not exhaustive or exclusive. A more comprehensive classification and illustration of logical fallacies can be found in the works of Kamiński (1962) and Suchoń (2005). The primary objective of this paper is to underscore the fact that enduring obstacles to sound conclusions, including scientific ones, in logic can impose severe limitations on research. The potential causes and consequences of these errors are discussed in detail later in the paper.

Correct answer incorrect question

A type III error can arise when the analysis and conclusions fail to meet the objective of the research (Tate, 2022). Asking the wrong question may be related to the methodology used and the statistical analysis presented. Incorrect operationalization (Tate, 2022) and a lack of complete knowledge of the process under study (Mohajan, 2017) are among the methodological causes of the third type of error. The first mentioned aspect of operationalization can be understood as a series of activities that aim to translate the theoretical definition of a domain into a set of ones and zeros in a database. These activities lead to the selection of measurement tools, focusing on the selection of appropriate definitions and concepts derived from the adopted paradigm. A poorly defined conceptual apparatus and an inappropriate choice of measurement methods can be an obstacle to the achievement of the stated objectives (Tate, 2022). Inadequate knowledge of the process under study can be a second methodological problem leading to type III errors. The formulation of a correct question requires the researcher to carefully evaluate and analyze the theoretical problem that the study is intended to measure and describe. Failing to fully understand the question may result in omitting factors important to the model/process, leading to conclusions that may be useless (Mohajan, 2017). Conversely, it is not always feasible to obtain and present all previous scientific reports on a given subject, especially in the case of exploratory studies.

Statistical reasons for making type III errors in research papers can be found in the choice of methods for analyzing data (Kimball, 1957) and in the statistical area of hypothesis testing (Shaffer, 2002). An incorrectly formulated research question may be reflected in a poor choice of methods for the analysis of the data. The formulation of a question that focuses on differences and the preparation of a correlation analysis in response, or the use of an analysis method for independent data in a repeated measures testing situation, can lead to a type III error. Although the correct answer is found and the results interpreted accurately, the conclusions pertain to an area different from the study's original aim. The choice of one-sided or two-sided statistical testing may also result in a mismatch between the research question and the aim of the study. Two-sided testing (or its accumulation in post hoc evaluation of pairwise comparisons) can be an abuse in a situation with a research aim requiring specific comparisons, leading to formulating the wrong (too broad) question. Mosteller's (1948 after: Levin & Marasculio, 1970) proposal for defining the third type of error is that it occurs when the null hypothesis is rightly rejected but for the wrong reason. This may have been caused by the incorrect selection of the data analysis method and the study area (Kaiser, 1966).

The further the study design and subsequent data analysis stray from the aim of the study, the greater the risk of a type III error, which is a good answer to a poorly described question. Failure to consider this type of risk in a study can result in an internally inconsistent study. As a result, conclusions drawn in the context of the objective may not be feasible (Scanlon et al., 1977; Shwartz & Carpenter, 1999). The occurrence of an error (type III) in operationalizing concepts and understanding the research process, or in choosing data analysis methods and the scope of testing those methods, can result in a correct answer related to the null hypothesis, but for the wrong reason (Mohajan, 2017). Marasculio and Levin (1970) compare this error to the correct conclusion that fails to achieve the objective set in the thesis.

Incorrect answer correct question

A type IV error in a survey can occur when the results are correctly interpreted but lead to erroneous conclusions (Mohajan, 2017). The incorrect answer in a paper can originate from the survey methodology or statistical analysis. The problems leading to the type IV error can be identified in the research methodology or its limitations. It is possible for a well-posed research question to receive an incorrectly formulated answer if the results are misinterpreted without considering the limitations of the research design. Such errors result from drawing hasty conclusions regarding the influence of variables and causal interaction based on correlation or difference results without using an appropriate experimental design. In addition, Yamatani and Mann (2013) observe that type IV error may also occur when a paper draws inconsequential and simplistic conclusions, thereby overlooking significant and pivotal findings.

The reasons for providing an incorrect response in a statistical context may include neglect of violations of the statistical test assumptions (Tate, 2022) and the misinterpretation of the interactions (Umesh et al., 1996). As Tate (2022) highlighted, a flawed answer or erroneous conclusion can result from a failure to verify the assumptions underlying the chosen statistical approach. The inclusion of outlier cases, the lack of homogeneity of variance, or other factors can lead to inaccurate estimate of the results. In such instances, it is essential to utilize appropriate statistical tests, often employing non-parametric ones. In examining the interaction aspect of factors, Umesh and colleagues (1996) highlighted in their paper that the mean comparison approach employed in the analysis does not correspond with an assessment of statistical interactions.

The assessment of pairwise comparisons of all averages may result in a type IV error if neither the interaction effects value nor the direction of the research question is considered. Post-hoc assessment or analysis of planned comparisons without reference to the purpose of the study and the direction of the question may also result in erroneous conclusions. Marasculio and Levin (1970) even suggested that the significance of differences between groups in measurement does not necessarily result in a significant interaction effect.

The consequence of a type IV error in the evaluation of assumptions, appraisal of consequences, or a misguided attempt at establishing a causal link in the absence of substantiated evidence can result in an inaccurate interpretation of a valid null hypothesis (Mohajan, 2017). As Marasculio and Levin (1970, p. 398) aptly observed, this can be likened to a correct diagnosis followed by the prescription of an incorrect treatment.

Conclusion

As noted by Tate (2022), errors of type III (in implementation) and type IV (in evaluation) are discernible in social research papers. The formulation of an accurate research question and a detailed description of the research problem necessitate a comprehensive examination of the existing body of literature and the current state of knowledge in the field under investigation. The theory that gives rise to the research problem must be presented to the practitioners who will be the beneficiaries of the findings, as proposed by Scanlon et al. (1977). In order to facilitate this, the research team should also draw on the experience of practitioners in the context of the psychoeducation methods used, diagnosis, therapy, or other aspects it intends to cover. This approach will enable the operationalization of concepts and the selection of measurement methods with appropriate psychometric properties and potential for practical application. A thorough grasp of the subject matter is also conducive to understanding its intricacies. The incorporation of variables, circumstances, and the interplay between them – conditions or interactions that are readily apparent to seasoned practitioners, as well as to those they serve – can be facilitated (Kimball 1957; Scanlon et al., 1977).

A research question formulated in this way must consider the testing area. In line with Cohen (1990, after: Umesh et al., 1996), a research question should include specific variables that indicate, prior to the study, whether univariate (directional analyses) or bivariate (non-directional analyses) comparisons are required. It is also essential that the research question clearly directs the specific type of analysis to be carried out. The construction of a subsequent correlation or difference analysis must follow directly from the content of the research question. Even identical data analysis methods, such as regression analysis, can yield different results based on the research questions and design. The inclusion of relevant concepts and indicators in the research, along with an understanding of the relationships between them, allows the researcher to construct the correct research question. Selecting data analysis methods appropriate to the research question will facilitate achieving the study objective. A consideration of the aforementioned factors will enhance the implementation of the conclusions and findings derived from the study.

A correct conclusion in any study can only be reached if the limitations of the resulting research plan have been taken into account. Over-generalization, caused by an unrepresentative sample, causal inference without an experimental design, or other smaller research failures, is not unusual. It is not always possible to eliminate such errors entirely. However, reflecting on these limitations and incorporating them into the conclusions can help ensure that the author does not misinterpret the research problem or reach an incorrect conclusion.

It is also important to consider problems with the analyzed data, which will help improve the quality of inferences. Addressing issues related to outliers, homogeneity, and other assumptions (Tate, 2022) and selecting an appropriate statistical method (often non-parametric tests) minimizes the risk of formulating an incorrect answer. Furthermore, data and results must not be evaluated carelessly, especially regarding interaction effects. Entirely different conclusions may result from using the omnibus F test compared to those from comparing group means (Umesh et al., 1996). Neither method is inferior nor suited to different research questions, leading to different conclusions. Awareness of methodology limitations and accurate evaluation of statistical assumptions can help develop a well-founded response to the research question, thereby ensuring the appropriate assessment of the key aspects discussed in the study.

Tate (2022) posits that efforts to reduce type III and type IV errors in research will facilitate greater reproducibility. It is possible to achieve greater uniformity in research methods in order to minimize the risk of type III error. The application of consistent statistical methods will result in a reduction in the occurrence of type IV error.

References

Aspers, P., & Corte, U. (2019). What Is Qualitative in Qualitative Research. *Qualitative Sociology*, *42*, 139–160, doi: 10.1007/s11133-019-9413-7.

Bayes, T. (1763). LII. An Essay Towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, FRS Communicated by Mr. Price, in a Letter to John Canton, AMFR S. *Philosophical Transactions of the Royal Society of London*, *53*, 370–418, doi: 10.1093/biomet/45.3-4.296.

Brzeziński, J.M. (2021). Testy istotności różnic i wskaźniki wielkości efektu ES – wybrane zagadnienia [Tests of Significance of Differences and Indicators of ES Effect Size – Selected Issues]. In: J.M. Brzeziński (Ed.), *Metodologia badań psychologicznych. Wydanie nowe* [Psychological Research Methodology. New Edition] (pp. 205–234). Warszawa: Wydawnictwo Naukowe PWN.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. New York, NY: Psychology Press, doi: 10.4324/9780203771587.

Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics*, 5th ed. London: SAGE Publications.

Fisher, R.A. (1925). *Statistical Methods for Research Workers*. Edinbugh, Scotland: Oliver & Boyd.

Gigerenzer, G. (2004). Mindless Statistics. *The Journal of Socio-Economics*, *33*(5), 587–606, doi: 10.1016/j.socec.2004.09.033.

Habes, M., Ali, S., & Pasha, S.A. (2021). Statistical Package for Social Sciences Acceptance in Quantitative Research: From the Technology Acceptance Model's Perspective. *FWU Journal of Social Sciences*, *15*(4), 34–46, doi: 10.51709/19951272/Winter-2021.

Henson, K. (2006). Effect-Size Measures and Meta-Analytics Thinking in Counseling Psychology Research. *The Counseling Psychology*, *34*(5), 601–629, doi: 10.1177/00110005283558.

Hubbard, R. (2004). Alphabet Soup. Blurring the Distinctions Between p's and α's in Psychological Research. *Theory* & *Psychology*, *14*(3), 295–327, doi: 10.1177/0959354304043638.

Ioannidis, J.P. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), e124, doi: 10.1371/journal.pmed.0020124.

Józefacka, N., Kołek, M.F., Arciszewska-Leszczuk, A., & Iwanowski, P. (2023). *Metodologia i statystyka: Przewodnik naukowego turysty* [Methodology and Statistics: A Scientific Tourist's Guide]. Warszawa: Wydawnictwo Naukowe PWN.

Kamiński, S. (1962). Systematyzacja typowych błędów logicznych [Systematization of Typical Logic Errors]. *Roczniki Filozoficzne/Annales de Philosophie/Annals of Philosophy*, *10*(1), 5–39.

Kimball, A.W. (1957). Errors of the Third Kind in Statistical Consulting. *Journal of the American Statistical Association*, *52*(278), 133–142, doi: 10.1080/01621459.1957.10501374.

King, B.M., & Minium, E.W. (2020). *Statystyka dla psychologów i pedagogów* [Statistics for Psychologists and Educators], 4th ed. Warszawa: Wydawnictwo Naukowe PWN.

Kruschke, J. (2014). *Doing Bayesian Data Analysis: A Tutorial with R. JAGS*, *and Stan, 2*. Academic Press.

Kruskal, W. (1980). The Significance of Fisher: A Review of R.A. Fisher: The Life of a Scientist. *Journal of the American Statistical Association*, *75*(372), 1019–1030, doi: 10.2307/2287199.

Kublikowski, R. (2011). Czym jest definicja? [What Is the Definition]. In: P. Kawalec, P. Lipiński, & R. Wodzisz (Eds.), *Podstawy naukoznawstwa* [Basics of Science] (pp. 185–209). Lublin: Wydawnictwo KUL.

Macdonald, R.R. (1997). On Statistical Testing in Psychology. *British Journal of Psychology*, *88*(2), 333–347, doi: 10.1111/j.2044-8295.1997.tb02638.x.

Marascuilo, L.A., & Levin, J.R. (1970). Appropriate Post Hoc Comparisons for Interaction and Nested Hypotheses in Analysis of Variance Designs: The Elimination of Type IV Errors. *American Educational Research Journal*, *7*(3), 397–421.

Marvasti, A.B. (2004). *Qualitative Research in Sociology*. London: Sage Publications Ltd., doi: 10.4135/9781849209700.

McElreath, R. (2018). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC.

Mehrad, A., & Zangeneh, M.H.T. (2019). Comparison Between Qualitative and Quantitative Research Approaches: Social Sciences. *International Journal for Research in Educational Studies, Iran*, *5*(7), 1–7.

Mohajan, H.K. (2017). Two Criteria for Good Measurements in Research: Validity and Reliability. *Annals of Spiru Haret University. Economic Series*, *17*(4), 59–82. doi: 10.26458/1746.

Neyman, J., & Pearson, E.S. (1928a). On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference Part I. *Biometrika*, *20*(1–2), 175–240, doi: 10.2307/2331945.

Neyman, J., & Pearson, E.S. (1928b). On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part II. *Biometrika*, *20A*, 263–294, doi: 10.2307/2332112.

Perezgonzalez, J.D. (2015). Fisher, Neyman-Pearson or NHST? A Tutorial for Teaching Data Testing. *Frontiers in Psychology*, *6*, 223, doi: 10.3389/fpsyg.2015.00223.

Scanlon, J.W., Horst, P., Nay, J.N., Schmidt, R.E., & Waller, J.D. (1977). Evaluability Assessment: Avoiding Type III and IV Errors. In: G.R. Gilbert, & P.J. Conklin (Eds.), *Evaluation Management: A Sourcebook of Readings* (pp. 264-282). United States Civil Service Commission – Bureau of Training.

Schwartz, S., & Carpenter, K.M. (1999). The Right Answer for the Wrong Question: Consequences of Type III Error for Public Health Research. *American Journal of Public Health*, *89*(8), 1175–1180, doi: 10.2105/AJPH.89.8.1175.

Shaffer, J.P. (2002). Multiplicity, Directional (Type III) Errors, and the Null Hypothesis. *Psychological Methods*, *7*(3), 356, doi: 10.1037/1082-989X.7.3.356.

Suchoń, W. (2005). *Prolegomena do retoryki logicznej* [Prolegomena to the Rethoric of Logic]. Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.

Sun, S., Pan, W., & Wang, L.L. (2010). A Comprehensive Review of Effect Size Reporting and Interpreting Practice in Academic Journals in Education and Psychology. *Journal of Educational Psychology*, *102*(4), 989–1004, doi: 10.1037/a0019507.

Tate, C.U. (2022). The Importance of Type III and Type IV Epistemic Errors for Improving Empirical Science. In: L.J. Jussim, J.A. Krosnick, & S.T. Stevens (Eds.), *Research Integrity: Best Practices for the Social and Behavioral Sciences* (pp. 316–337). Oxford University Press, doi: 10.1093/oso/9780190938550.003.0012.

Umesh, U.N., Peterson, R.A., McCann-Nelson, M., & Vaidyanathan, R. (1996). Type IV Error in Marketing Research: The Investigation of ANOVA Interactions. *Journal of the Academy of Marketing Science*, *24*, 17–26, doi: 10.1007/BF02893934.

Uttley, J. (2019). Power Analysis, Sample Size, and Assessment of Statistical Assumptions – Improving the Evidential Value of Lighting Research. *Leukos*, *15*, 143–162, doi: 10.1080/15502724.2018.1533851

Vacha-Haase, T., & Thompson, B. (2004). How to Estimate and Interpret Various Effect Sizes. *Journal of Counseling Psychology*, *51*(4), 473–481, doi: 10.1037/0022-0167.51.4.473.

Xiong, X. (2022). Critical Review of Quantitative and Qualitative Research. In 2022 3rd International Conference on Mental Health, Education and Human Development (MHEHD 2022) (pp. 956–959). Atlantis Press, doi: 10.2991/assehr.k.220704.172.

Yamatani, H., Mann, A., & Feit, M. (2013). Avoiding Type III, IV, and V Errors Through Collaborative Research. *Journal of Evidence-Based Social Work*, *10*(4), 358–364, doi: 10.1080/15433714.2012.664050.