International Journal of Assessment Tools in Education
2022, Vol. 9, No. 2, 290–318
https://doi.org/10.21449/ijate.946743
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

# The Methodological Quality of Experimental STEM Education Articles Published in Scholarly Journals from 2014 to 2020

**Ramazan Avcu**[1,*], **Seher Avcu**[2]

[1,2]Aksaray University, Faculty of Education, Department of Mathematics and Science Education, Aksaray, Türkiye

**ARTICLE HISTORY**

*Received: June 02, 2021*
*Revised: Feb. 08, 2022*
*Accepted: Mar. 15, 2022*

**Keywords:**
STEM education,
Methodological quality,
Experimental research,
Scholarly journal articles.

**Abstract:** Experimental studies have a considerable impact on the educational policies and practices of many countries. In Turkey, policymakers are planning to initiate a STEM education reform in K-12 schools based on experimental studies. However, the methodological flaws in these studies may lead to biased outcomes and may mislead the STEM education community. Despite the importance of methodological quality, to the best of our knowledge, there are no studies that investigate the methodological quality of experimental STEM education articles published in scholarly journals. Therefore, in this study, we conducted a methodological review to examine the methodological quality of experimental STEM education articles published in refereed Turkish journals from 2014 to 2020. During the targeted period, we located 68 articles. We analyzed these articles by developing a coding framework. We found that the selected articles suffer seriously from various methodological flaws. We discuss the findings in light of the literature on methodological quality and suggest ways to improve the rigor of the experimental designs used. Ultimately, we discuss some implications for authors, journals editors, policymakers, and curriculum developers.

## 1. INTRODUCTION

Experimental research findings have a considerable impact on the decisions taken by the policymakers about the educational practices that should be adopted in their own countries (Borman et al., 2005; Slavin, 2008). For instance, in the USA, Finn and Achilles (1999) conducted an experiment to investigate the effect of class size on elementary students' academic achievement and found out that the students in small classes (13–17 students) had superior performances compared to the students in regular classes (22–26 students). Finn and Achilles' (1999) findings initiated the educational reform entitled class-size reduction and led many states to reduce the number of students in the classrooms to improve student learning. In another experimental study, Schweinhart et al. (1993) examined the benefits of pre-school programs to children who live in poverty and who are at the risk of failing at school. They revealed that the pre-school students had significantly higher achievement scores, high school graduation rates, and earnings, while they had less crime rates and welfare use compared to the non-preschool students as of age 27. Their findings urged the legislators to deliver publicly funded programs in many states and localities in the USA.

Given the impact of experimental studies on educational policies and practices (Borman et al., 2005; Slavin, 2008), using rigorous methodological designs and techniques is of crucial

*CONTACT: Ramazan AVCU ramazanavcu@aksaray.edu.tr Aksaray University, Faculty of Education, Department of Mathematics and Science Education, Aksaray, Türkiye

e-ISSN: 2148-7456 /

importance in establishing that the observed effects in an experiment are caused by the treatment (e.g., a specific teaching technique, a newly developed curriculum, or an instructional program) but not by the extraneous variables. On the other hand, the flaws or errors in an experimental study may lead to false reports in the literature; other researchers may build theories or conduct other experiments by using these spoiled findings; and as a result, a great deal of time, money, effort, and other resources may be wasted (Gravetter et al., 2021). For instance, the United States Department of Education (2020) announced that almost 1.5 billion dollars were invested between the years 2018 and 2020 to support high-quality STEM (Science, Technology, Engineering, and Mathematics) education for students. In return for this, it expects researchers to conduct experiments with random assignments (i.e., randomized trials) and prioritizes researchers whose grant applications involve such rigorous methodological designs (Hedges & Schauer, 2018) because only in such designs it can be ascertained that the observed effects on important student outcomes such as academic achievement are caused by STEM education practices but not by other extraneous variables.

In Turkey, traditional educational practices being implemented in schools are not considered sufficient for students in solving real-world problems and gaining the knowledge and skills that are compulsory for maintaining their future careers (Akgündüz et al., 2015). For this reason, the Ministry of National Education (2016) is planning to initiate a STEM education reform in the near future to help students gain the technical knowledge and skills needed in the contemporary workplace and consequently to better prepare them for real life. Unfortunately, an action plan for implementing STEM education in Turkey has not been prepared yet (Ministry of National Education, 2018). However, the Ministry of National Education (2016) advocates the conduction of research studies on STEM education as a first step in developing this action plan. Thus, examining the methodological quality of experimental STEM education articles may help educational policymakers and curriculum developers determine a clear STEM education road map for students in all educational stages. With this idea in mind, in this study, we aimed to conduct a methodological review of experimental STEM education articles published in refereed Turkish journals to reveal whether current research practices in these journals are in agreement with the canons of educational research as described in commonly used methodology textbooks such as Creswell and Creswell (2018), Frankel et al. (2012), Gall et al. (2007), Cohen et al. (2018), and Johnson and Christensen (2020).

### 1.1. Significance and Research Questions

Although plenty of researchers conducted content analysis studies to determine the trends in STEM education research (e.g., Aydın Günbatar & Tabar, 2019; Brown, 2012; Çavaş et al., 2020; Çevik, 2018; Daşdemir et al., 2018; Elmalı & Balkan Kıyıcı, 2017; Kaya & Ayar, 2020; Li et al., 2020; Mizell & Brown, 2016), there is a dearth of studies that explore the methodological quality of educational research articles published in refereed journals (e.g., Horton et al., 1993; Shaver & Norton, 1980; Sung et al., 2019; Wallen & Fraenkel, 1988). What is more, to the best of our knowledge, there are no studies that investigate the methodological quality of experimental research articles on STEM education. Thus, this study attempts to fill this gap by analyzing the experimental STEM education articles published in refereed Turkish journals with respect to the following categories: formulating purpose statements, research questions, and hypotheses; clarifying contribution to the literature; describing the type of experimental design; describing the sample, sampling strategy, and the population; establishing instrument validity and reliability and describing their types; fulfilling the basic assumptions of the parametric tests used; attending to minimum sample size in experimental and control groups; and reporting effect sizes and statistical powers for the parametric tests used.

Since a methodological review of experimental research on STEM education in refereed Turkish journals had not been undertaken before, the findings of the study first provide the

researchers and other stakeholders with a snapshot of prevailing research reporting practices in Turkish journals. Second, they inform the STEM education community about the state-of-the-art and the soundness of experimental research practices in these journals. More importantly, since "a periodic review of common research practices in a scholarly discipline aids in improving those practices" (Horton et al., 1993, p. 858), our findings may enhance the quality of experimental research articles that will be published in these journals. Based on our findings, the editors and editorial board members of these journals may increase their article publication standards by ensuring that information about the above categories is provided by the authors who intend to publish their manuscripts in these journals. As authors pay increased attention to ensuring methodological rigor in their research manuscripts, deficiencies in their research reporting practices may diminish, and this may pave the way for more meaningful and consistent research on STEM education in Turkish journals.

In the transition from traditional education to STEM education, the Ministry of National Education (2016) deems it very significant to prepare and implement a good action plan considering the common sense of all stakeholders in the educational arena. Thus, the findings of our study may benefit but much to the Ministry of National Education. If publishing high-quality research articles becomes a standard practice for refereed journals in Turkey, the research findings about STEM education in these journals may point curriculum developers in the Ministry of National Education in the right direction. Frankly, the rigorous research findings accumulated from these journals may help curriculum developers design relevant STEM education materials and optimal STEM learning environments for students and help them integrate STEM education into Turkish school curricula in the best possible way.

Due to the above considerations, we conducted a methodological review to determine whether authors' experimental research reporting practices on STEM education in refereed Turkish journals are consistent with the commonly suggested research methods and procedures. Through this purpose, we sought to find answers to the following research questions:

1. Which research components do authors typically report in their articles?
    a. Do they report how they contribute to the scholarly literature?
    b. Do they report purpose statements?
    c. Do they report research questions?
    d. Do they report hypotheses?
    e. Do they report the type of experimental research design used?
    f. Do they describe the sample, the sampling strategy, and the population?
    g. Do they report instrument validity and reliability and describe their types?
    h. Do they report the basic assumptions that must be fulfilled for the parametric tests used?
    i. Do they attend to the minimum sample size required for experimental and control groups?
    j. Do they report effect sizes and statistical powers?

## 2. METHOD

### 2.1. Research Design

We conducted a *methodological review* to determine the methodological quality of experimental STEM education articles published in refereed Turkish journals. Methodological reviews describe the research designs, methods, and procedures used in scientific research and they foreground the strengths and weaknesses of methodological tools used in such research (Dochy, 2006). They are used in many fields to "improve research practice, inform debate, and identify islands of practice" (Randolph et al., 2013, p. 2). In these reviews, the focus is on identifying *how* research studies are conducted (i.e., the research methodologies used) rather than on identifying *which* research outcomes (i.e., the findings) are presented (Shukla, 2017).

Taking all these together, our methodological review helped us uncover authors' prevailing research reporting practices in the articles published in Turkish journals, determine the publishing standards of these journals, and suggest ways to improve the methodological quality of the articles published in these journals.

### 2.2. Data Sources

To locate the STEM articles published in refereed Turkish journals, we first formulated the following search terms (i.e., keywords or descriptors): STEM, STEM education, integrated STEM education, FeTeMM, FTMM (Turkish equivalents of STEM), FeTeMM eğitimi, FTMM eğitimi (Turkish equivalents of STEM education), entegre FeTeMM eğitimi, and entegre FTMM eğitimi (Turkish equivalents of integrated STEM education). Next, we typed these keywords in the following databases: TR Index (https://trdizin.gov.tr/), DergiPark (https://dergipark.org.tr/tr/), and Google Scholar (https://scholar.google.com.tr/). TR Index and DergiPark are national databases in Turkey, while Google Scholar is a search engine that is widely used all over the world. Our reason for using Google Scholar is that it "provides a simple way to do a broad search for scholarly literature, including peer-reviewed papers, theses, books, abstracts and articles" (Fraenkel et al., 2012, p. 55). More importantly, Google Scholar searches the entire internet. By this means, we were able to locate the articles that were not produced by the TR Index and DergiPark. In Google Scholar, we limited our search to "Turkish pages" to locate the articles published in Turkish journals and not to locate too many references. However, we used Google Scholar only as a supplement to TR Index and DergiPark and not as a substitute for them.

We delimited our search to the articles published till December 30, 2020. Our search elicited many studies with different research designs such as survey studies, correlational studies, theoretical papers, literature reviews, meta-analysis studies, scale development studies, and content analysis studies. However, to act in accordance with the purpose of the current study, we considered only the articles that used purely experimental research designs and the articles that combined experimental research designs with qualitative research designs (i.e., mixed methods studies). Thus, in the current study, 44 purely experimental research articles, 24 mixed methods articles, and in total 68 STEM education articles from 52 different journals underwent content analysis.

### 2.3. Coding Framework

To formulate the coding categories that are pertinent to our study, we first developed a tentative coding framework based on previous research on methodological quality (e.g., Horton et al., 1993; Shaver & Norton, 1980; Sung et al., 2019; Wallen & Fraenkel, 1988). Namely, we first used the predetermined categories developed by past researchers. As we coded the journal articles selected for our study, we had to make some changes to some of the codes or categories included in the tentative coding framework. Namely, we added some new codes or categories, deleted some of the codes or categories that were specified a priori, and refined some of these a priori codes or categories until the remaining codes and categories totally reflected the structure of our data. A final coding framework was developed when we were able to code all the journal articles exhaustively and explicitly with the codes and categories at hand. The categories and codes included in our final coding framework are explained below.

#### 2.3.1. *Contribution to the literature*

Contribution to the literature refers to relating the intended study to previous studies in a planned way (Nelson & Shaver, 1985). In other words, it refers to situating the intended study in the context of the existing body of literature simply to "avoid reinventing the wheel" (Orne, 1981, p. 1). There are four ways to report how a piece of research contributes to the literature: *i*) filling a gap or void in the literature, *ii*) replicating past research, *iii*) extending past research,

and *iv*) developing new ideas in the scholarly literature (Brown & Dant, 2008). Filling a void refers to examining concepts or ideas not addressed in the existing literature (Creswell, 2015). Replication refers to repeating a past study using a different group of participants and under different conditions such as different places, abilities, and socioeconomic status (Fraenkel et al., 2012). Extending past research refers to broadening a published study to a new topic or field or simply carrying out the study more deeply and exhaustively, for example, by incorporating new variables into the study (Ary et al., 2014; Creswell, 2015). Developing new ideas in the scholarly literature means dealing with new problems (i.e., problems that have not been explored before) that concern researchers and practitioners (Brown & Dant, 2008).

### 2.3.2. *Purpose statement*

Creswell and Creswell (2018) defined a purpose statement as a passage that conveys the overall intent of a research study in one or more sentences. They emphasized that a good purpose statement must be clear, specific, and informative and proposed the following design features in writing a good purpose statement: (*i*) words such as *purpose*, *intent*, *aim*, and *objective* should be used to draw attention to the central topic of the study, (*ii*) the study should be narrowed to a single phenomenon, concept, or idea, (*iii*) action verbs such as *examine*, *explore*, *discover*, *develop*, *generate*, and *understand* should be used to convey how the topic of the study will be learnt, (*iv*) directional language should be avoided and instead, neutral words or phrases should be used, (*v*) working definitions should be formulated especially for the terms that are not normally known by a large number of people in the research community, (*vi*) words that specify the research design of the study should be used, (*vii*) the participants of the study should be mentioned, (*viii*) the research site should be identified, and (*ix*) the research participants and sites should be delimited. In the current study, the journal articles that did not consider any one or more of these design features in their purpose statements were categorized as articles having *unclear* purpose statements.

### 2.3.3. *Research question*

A research question is a statement that is used to narrow the purpose statement to specific questions that a researcher attempts to answer by carrying out a study (Plano Clark & Creswell, 2015). Research questions are "concrete questions, carefully composed in order to address the research objectives, to constitute a fair operationalization and embodiment of a valid set of indicators for addressing the research objectives, providing answers which address the research purposes with warranted data" (Cohen et al., 2018, p. 165). The journal articles analyzed in the present study were divided into two as articles containing research questions and articles not containing research questions.

### 2.3.4. *Hypothesis*

A hypothesis is a prediction of the anticipated findings from scientific research (Fraenkel et al., 2012; Gall et al., 2014). Stating a hypothesis in a research study helps to ponder more thoroughly and precisely on the findings anticipated from a study, build a body of knowledge, and notice whether relationships between different variables are, or are not, examined (Fraenkel et al., 2012). There are two opposing hypotheses as null hypothesis ($H_0$) and alternative hypothesis ($H_1$) and hypothesis testing works under the premise that the null hypothesis is true (Gravetter et al., 2021). More precisely, researchers start with the null hypothesis, cast their research in the form of a null hypothesis, and turn to the alternative hypothesis when their data do not support the null hypothesis (Cohen et al., 2018). Simply put, "the null hypothesis is the focal point in hypothesis testing because it is the null hypothesis, not the alternative hypothesis, that is tested directly" (Johnson & Christensen, 2020, p. 514). In the current study, the journal articles were categorized into two as articles reporting a hypothesis and articles not reporting a

hypothesis. Articles reporting a hypothesis were further categorized into two as articles reporting a null hypothesis and articles reporting an alternative hypothesis.

### 2.3.5. *Type of experimental research design*

In experimental research studies, the effect of a treatment or an intervention (an independent variable) on an outcome (a dependent variable) is tested by attempting to control for all other factors (extraneous variables) that may influence that outcome (Creswell & Creswell, 2018). Experimental research studies provide the best way to establish cause-effect relationships between different variables (Fraenkel et al., 2012). Namely, they produce the strongest evidence of causality (Johnson & Christensen, 2020). In the current study, we considered the typology proposed by Fraenkel et al. (2012) and Johnson and Christensen (2020) and divided the experimental research designs used in the articles selected for analysis into three types: true experimental designs, quasi-experimental designs, and weak experimental designs. In true experimental designs, there are both experimental and control groups and the study participants are randomly assigned to these groups. In quasi-experimental designs, there are again experimental and control groups, but the study participants are not randomly assigned to these groups. In weak experimental designs, there is either no control group (i.e., there is only an experimental group) or the study participants are not randomly assigned to the groups (i.e., the groups are pre-existing or intact/static groups). Thus, true experiments are the most rigorous experimental design types, while weak experiments are the least rigorous ones.

### 2.3.6. *Sampling strategy*

The strategy used while selecting a sample from a population is called the sampling strategy (Johnson & Christensen, 2020). Educational research textbooks describe two main types of sampling as random sampling (probability sampling) and nonrandom sampling (nonprobability sampling). In these textbooks, the most commonly reported random sampling strategies are simple random sampling, systematic sampling, stratified sampling, and cluster random sampling, while the most commonly reported nonrandom sampling strategies are convenience sampling, purposive sampling, and quota sampling (Ary et al., 2014; Cohen et al., 2018; Johnson & Christensen, 2020; Mills & Gay, 2016). In the present study, when coding the selected articles, we considered the abovementioned strategies. For the articles that did not report any specific strategy, we coded only the main sampling strategy used. For the articles that did not provide any information about the main or specific sampling strategy used, we used the code “not reported”.

### 2.3.7. *Sample description*

Sample description refers to the information given about societal, demographic, economic, and other characteristics of the subjects who take part in a research study (Erdoğan, Marcinkowski, & Ok, 2009). In the current methodological review study, the subjects recruited were pre-service teachers and K-12 students. The socio-demographic characteristics of the pre-service teachers were age, gender, university fund type (privately versus publicly funded university), department studied, year level, cumulative grade point average, and type of high school graduated (traditional high school, foreign language intensive high school, or vocational and technical high school). Location of the university (the region or city where the university is located) was also another characteristic reported in the articles that recruited pre-service teachers as participants.

The socio-demographic characteristics of the K-12 students included their age, gender, ethnicity, educational stage (preschool, elementary school, middle school, and high school), school fund type (privately versus publicly funded school), giftedness, achievement level, attitude level, English language proficiency, English as a second language, special education status, at-risk status, and high school type (traditional high school or inclusive STEM high

school). K-12 students' familial characteristics included socio-economic status (SES), economic status, and residence (state, region, city, district, or village). K-12 students' other characteristics as reported in the selected articles were class size and educational opportunity.

To operationalize the extent of sample description in each journal article, we used the following categorization including three levels: poor description, mediocre description, and rich description. Poor description refers to articles that describe at most three different characteristics of the sample. Mediocre description refers to articles that describe four, five, or six different characteristics of the sample. Rich description refers to articles describing more than six different characteristics of the sample.

### 2.3.8. *Population*

The notion of science rests entirely on the idea of generalization (Fraenkel et al., 2012). In quantitative studies, researchers obtain information from a small group of individuals and usually wish to generalize their findings to a larger group of individuals (Fraenkel et al., 2012). This larger group, which includes all possible members of a group of people, events, or objects, is called a population (Ary et al., 2014). Briefly, population refers to the "set of all the individuals of interest in a particular study" (Gravetter et al., 2021, p. 4). Defining a population helps researchers determine the extent of generalizability of their findings (Mills & Gay, 2016). In respect to this, Fraenkel et al. (2012) emphasized that researchers should avoid narrowly defined populations as much as possible because in such studies the usefulness of the obtained findings is severely restricted. They also remarked that it is not worth spending a considerable amount of time, energy, and money on studies that produce low applicable findings.

In the present study, the selected STEM articles were categorized into two as those that reported a population and those that did not report a population. Moreover, for those that reported a population, we also evaluated population sizes.

### 2.3.9. *Instrument validity and reliability*

Instrument validity refers to the "appropriateness, correctness, meaningfulness, and usefulness of the specific inferences researchers make based on the data they collect" (Fraenkel et al., 2012, p. 148) and instrument reliability refers to the "consistency of the scores obtained" (Fraenkel et al., 2012, p. 154). In the current study, we first categorized whether the authors of the selected journal articles developed their own instruments or used pre-existing instruments developed by others. Next, we coded the availability of validity and reliability information about the instruments used by the authors no matter who developed these instruments. In other words, we also paid particular attention to coding availability of validity and reliability information for the articles in which already developed instruments were administered. This is because even formerly developed instruments with perfect validity and reliability do not guarantee that they will function in the same way in the latter studies. Differences in participants and contexts may make earlier validity and reliability coefficients non-transferable to novel participants and contexts. Moreover, validity is always contingent upon the goals and interpretations of the researchers (Fraenkel et al., 2012).

We further categorized the journal articles with respect to the types of validity and reliability used in them. We delimited our analysis of instrument validity to the following three major types: content validity, criterion-related validity, and construct validity (Ary et al., 2014; Cohen et al., 2018; Fraenkel et al., 2012; Mills & Gay, 2019). Similarly, we considered the following commonly reported reliability types in the educational literature when categorizing the selected articles: internal consistency (*i*. Cronbach's alpha, *ii*. Kuder-Richardson, and *iii*. split-half), test re-test, equivalent-forms, and interrater agreement (Ary et al., 2014; Creswell, 2015; Fraenkel et al., 2012; Johnson & Christensen, 2020; Mills & Gay, 2019).

### 2.3.10. *Basic assumptions of parametric tests*

Experimental research studies involve comparing scores obtained from two or more groups or under different conditions (Gravetter et al., 2021). Parametric tests are a subcategory of inferential statistics tests and are usually used to compare differences between the groups or conditions (Pagano, 2013). However, they require the fulfillment of several assumptions about the population and nature of data (Pallant, 2016).

In the present study, to examine how well the selected journal articles fulfilled the basic assumptions required for conducting parametric tests, we used the judgment tree proposed by Sung et al. (2019). According to this tree, *t*-tests and between-groups ANOVAs must meet the basic assumptions of normality and homogeneity of variance. Thus, articles that examined these assumptions and that did not report any violations were categorized as fulfilling the assumptions. However, if the articles did not examine normality and homogeneity of variance for *t*-tests and between-groups ANOVAs, we considered the following two criteria: (1) Are the number of participants in each group or cell equal to or greater than 30? (2) Are there an equal number of participants in each cell or group? Articles that met these criteria were also categorized as fulling the assumptions because *t*-tests and ANOVAs are robust with respect to violations of the normality and homogeneity of variance assumptions (Howell, 2017; Pagano, 2013).

According to Sung et al.'s (2019) judgment tree, to conduct repeated-measures ANOVA and mixed-design ANOVA, homogeneity of regression slopes in addition to normality and homogeneity of variance must be met (Hair et al., 2019; Kirk, 2013). Thus, the journal articles that met the three assumptions for repeated-measures ANOVA and mixed-design ANOVA were categorized as fulfilling the basic assumptions; otherwise, they were categorized as not fulfilling the basic assumptions. Similarly, to conduct ANCOVA, researchers must satisfy the sphericity assumption in addition to normality and homogeneity of variance (Hair et al., 2019; Kirk, 2013). Thus, articles meeting these three basic assumptions for ANCOVA were also coded as fulling the basic assumptions.

### 2.3.11. *Sample size*

Sample size refers to the number of participants in a research study (Frey, 2018). It is important to note that by sample size we refer to the final sample size, not to the designated sample size (Shapiro, 2008), because the number of participants in a designated sample may be much fewer if a considerable number of individuals drop out of experimental research studies. Authors of commonly used educational research textbooks (e.g., Ary et al., 2014; Fraenkel et al., 2012; Mills & Gay, 2016) recommend a minimum of 30 participants in each cell or group (i.e., experimental and control groups). Including a minimum of 30 participants in each group is significant because a sample size less than 30 for each group may lead to low statistical power and this may, in turn, endanger the validity of experimental research (Cheung & Slavin, 2012). A summary of commonly used experimental research designs and the sample sizes needed to conduct the corresponding statistical tests are presented in Table 1.

**Table 1.** *The required sample size for statistical tests conducted under different experimental research designs when statistical power is 0.80, effect size is moderate, and* α = 0.05 (Sung et al., 2019, p. 18).

| | Within-subject design | | Mixed design | Between-subject design | | | | |
|---|---|---|---|---|---|---|---|---|
| Experimental | Paired *t*-test | Multi-factor ANOVA (2*2 levels, interaction) | Multi-factor ANOVA (2*2 levels, interaction) | Independent *t*-test | ANOVA (2 levels) | ANOVA (3 levels) | ANCOVA (2 levels) | Multi-factor ANOVA (2*2 levels, interaction) |
| Pre-experiment | 34 | | | | | | | |
| Quasi-experiment | | | | | | | | |
| Pre test | | | | 128 | 128 | 159 | | |
| Gain Score | | | | 128 | 128 | 159 | | |
| ANCOVA | | | | | | | 128 | |
| Counterbalance | 34 | 30 | | | | | | |
| Multi-factor | | 30 | 34 | | | | | 179 |
| True-experiment | | | 34 | 128 | 128 | 159 | 128 | 179 |

Upon examining the selected articles, we used the following categories for the number of participants in each group or cell: 10–19, 20–29, 30–49, 50–99, and 100 and above.

### 2.3.12. *Effect size*

Effect size is a measure that quantifies the magnitude of difference between two groups (Coe, 2021). In experimental research, it refers to the treatment effect (Gravetter et al., 2021). It supplements statistical significance because statistical significance alone does not provide enough evidence for the importance of the findings (Warner, 2013). Besides, studies with large sample sizes can easily reach statistical significance even if the difference between the groups has little or no practical significance at all (Pallant, 2016). On the other hand, effect size is independent of sample size (Gravetter et al., 2021). Thus, it is not influenced by very small or large sample sizes. For this reason, reporting effect sizes in addition to statistical significance tests plays a crucial role in adopting a more rigorous approach to determining the effectiveness of experimental interventions and consequently encouraging a more scientific approach to the accumulation of research findings (Coe, 2021). As effect sizes are valuable means to report and interpret educational effectiveness, in this study, we categorized the selected articles into two as those that reported effect sizes and those that did not report effect sizes.

### 2.3.13. *Statistical power*

Statistical power refers to the "probability that the test will identify a treatment effect if one really exists" (Gravetter et al., 2021, p. 275). In experimental studies, it refers to the experiment's sensitivity to detect a treatment effect that really exists (Pagano, 2013). Statistical power is dependent upon three factors: sample size, effect size, and alpha level set by the researcher. It should be cautioned that in studies that are carried out with a quite small sample size (e.g., 20 participants), non-significant results may be obtained because of low statistical power. Thus, statistical power demonstrates how much confidence researchers should have in the results when they fail to reject the null hypothesis (Pallant, 2016). Besides, Cohen (1998) recommended that the power of a statistical test should be at least 0.80 (i.e., 80% probability of detecting an effect if there is actually one). Similarly, Sung et al. (2019) pointed out that if the power of a statistical test is less than 0.50, then obtaining a significant or non-significant result will be similar to guessing. Accordingly, given the importance of reporting statistical power, in this study, we categorized the selected journal articles into two as those that reported statistical power and those that did not report statistical power.

### 2.4. Data Analysis

Content analysis is used to analyze "written or visual materials for the purpose of identifying specified characteristics of the material" (Ary et al., 2014, p. 488). This research method helps to study human behavior indirectly usually through analysis of documents such as textbooks, essays, and magazine articles (Fraenkel et al., 2012). It is commonly used by educational researchers for several reasons such as revealing textbook biases, prejudices, and propaganda; analyzing error types in learners' writings; identifying prevailing practices; determining the difficulty level of a textbook content; and finding out the importance given to and the interest shown in some topics compared to the other ones (Ary et al., 2014). In the current study, we used this method to reveal the methodological quality of experimental STEM education articles published in refereed Turkish journals. When conducting our content analysis, we followed the steps recommended by Ary et al. (2014). These steps are explained below.

Specifying the phenomenon to be investigated: The phenomenon that we investigated in our content analysis was the *methodological quality* of publications on STEM education in the scholarly literature. More clearly, the phenomenon explored was Turkish educational researchers' prevailing research reporting practices. By examining this phenomenon, we aimed to determine the consistency between authors' research reporting practices and the commonly suggested research methods and procedures by well-known methodology textbooks.

Selecting the media from which the observations are to be made: The media selected for investigating the phenomenon of methodological quality were *peer-reviewed articles* published in Turkish journals. However, not all articles on STEM education were subjected to content analysis. That is, we only analyzed purely experimental STEM education articles and mixed methods articles on STEM education that used any type of experimental research design in their quantitative dimensions. There are several reasons for delimiting our analysis to peer-reviewed journal articles on STEM education in which experimental designs are used either completely or partially. First, journal articles provide the most recent research for the audience (Stebbins, 2006). Second, they are primary sources because the authors report their findings directly to the readers through them (Fraenkel et al., 2012). Third, they are expected to maintain higher standards to ensure quality (Creswell, 2015). Last and foremost, experimental research is "the only type of research that directly attempts to influence a particular variable", and consequently, is "the best way to establish cause-and-effect relationships among variables" (Fraenkel et al., 2012, p. 265). It is for this reason that policymakers and other stakeholders consider experimental research findings when making decisions about nationwide educational practices (Gall et al., 2007).

Formulating coding categories: Based on previously developed coding frameworks (e.g., Horton et al., 1993; Sung et al., 2019; Wallen & Fraenkel, 1988) and recommendations of well-known methodology textbook authors (e.g., Creswell, 2015; Frankel et al., 2012; Cohen et al., 2018; Johnson & Christensen, 2020) for conducting more rigorous research, we designed a comprehensive coding framework that comprises the following categories: formulating purpose statements, research questions, and hypotheses; clarifying contribution to the literature; describing the type of experimental design; describing the sample, sampling strategy, and the population; establishing instrument validity and reliability and describing their types; fulfilling the basic assumptions of the parametric tests used; attending to minimum sample size in experimental and control groups; and reporting effect sizes and statistical powers for the parametric tests used. In the previous section, these coding categories were defined and explained in some detail.

Deciding on the sampling plan to be used: STEM education research does not have a long history in Turkey. Several researchers (e.g., Aydın Günbatar & Tabar, 2019; Daşdemir et al., 2018; Elmalı & Balkan Kıyıcı, 2017) indicated that, in Turkey, STEM education research was

first initiated in 2014. Similarly, Özcan and Koca (2019) expressed that STEM education research in Turkey has gained momentum only over the past 5 years. Furthermore, our extensive review of literature also shows that there are a limited number of research studies on STEM education in Turkey. More importantly, we could locate a significantly fewer number of refereed journal articles on STEM education that employed an experimental research design. This extensive literature review helped us decide on the sampling plan to be used. Namely, in our content analysis, we attempted to locate the entire population of experimental STEM education articles published in refereed Turkish journals and thereby aimed to obtain an almost perfectly representative sample.

Training the coders: The first and second author of the current study coded the methodological quality of the journal articles. Before the actual coding, the two authors conducted sample coding for several experimental articles (different from the 68 articles selected for actual analysis) published in refereed Turkish journals. They independently coded the sample articles. Next, they held several sessions to discuss their independent coding and resolve the conflicting codes. These sessions also helped to clarify the meanings of the categories, make them more complete, and consequently revise and refine the coding framework. After the training session, the two authors separately coded all articles with respect to methodological quality by using the final form of the coding framework. In the first round of coding, the intercoder agreement (Miles et al., 2014) between the two coders was around 80%. Miles et al. (2014) recommended that "intercoder agreement should be within the 85% to 90% range, depending on the size and range of the coding scheme" (p. 85). Thus, the coders first identified the conflicting codes and re-examined the corresponding articles. In the second round of coding, the intercoder agreement reached 93%. The two authors discussed the rest of the conflicting codes periodically until they negotiated and arrived at a full consensus.

Analyzing the data: Once we have finalized coding our data, we counted the frequency of each code under each category. We also calculated percentages for these codes. Next, as recommended by Fraenkel et al. (2012), we assigned a label for each article (i.e., A1–A68) to facilitate data analysis. In these labels, the letter A stands for "Article" and the numbers ranging between 1 and 68 denote articles' IDs. The frequencies and percentages about each code under each category helped us summarize and interpret our research data. Namely, through frequencies and percentages, we were able to reveal refereed Turkish journals' trends in STEM education regarding methodological quality. More specifically, we could detect the decreasing and increasing trends in authors' use of research methods and procedures to improve the rigor of their articles.

### 2.5. Trustworthiness of the Study

To establish the trustworthiness of our study, we applied the following criteria proposed by Lincoln and Guba (1985): credibility, transferability, dependability, and confirmability. To ensure credibility, we explained our rationale for using a methodological review, thickly described the categories of our coding framework, examined previous research findings on the methodological quality of journal articles, and compared our findings with these previous research findings in the discussion section. Moreover, we used both data and investigator triangulation, spent prolonged time reading the full texts of the articles (over 3 months) to become familiar with the data and to obtain rich data for analysis, and finally, we used peer debriefing. That is, we had an associate professor review our data and examine our codes and categories. He has considerable experience in experimental research and methodological reviews. Through his review and examination, we received constructive feedback and thereby improved the quality of our findings.

To ensure transferability, we clearly described how we selected the articles for content analysis, explained our potential inclusion criteria, and described the main characteristics of the articles

so that other researchers who would like to examine the methodological quality of experimental studies may evaluate whether the findings drawn from the articles analyzed in the current study are applicable to other scientific documents such as books and proceedings. To establish dependability, first, we tried to maintain consistency across the entire study period including the starting point of research, data collection, and analysis. Meanwhile, we described each of the research steps transparently. Second, we explained our data analysis process as clearly as possible and used tables to report our findings to help other researchers evaluate the whole data coding process and replicate our study if desired. To achieve confirmability, we tried to do our best to control our biases and paid careful attention to shaping our findings solely by the data collected from the journal articles. More specifically, we used Ahern's (1999) ten tips to achieve reflexive bracketing.

## 3. FINDINGS

In this section, we report the findings related to the methodological quality of experimental STEM education articles published in refereed Turkish journals. In what follows, we present the number and percentage of articles with respect to each category of our coding framework.

### 3.1. Contributions to the Literature

Articles' contribution types to the STEM education literature are presented by year of publication in Table 2.

**Table 2.** *Articles' contribution types to the STEM education literature.*

| Contribution to the literature | 2014 | | 2015 | | 2016 | | 2017 | | 2018 | | 2019 | | 2020 | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % |
| Filling a gap or void in the literature | 1 | 50 | 1 | 25 | - | - | 1 | 17 | 5 | 38 | 3 | 27 | 11 | 39 | 22 | 32 |
| Replicating past research | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Extending past research | - | - | - | - | - | - | - | - | - | - | - | - | 1 | 4 | 1 | 1 |
| Developing new ideas in the scholarly literature | - | - | - | - | - | - | 2 | 33 | 6 | 46 | 2 | 18 | 8 | 29 | 18 | 26 |
| Not reported | 1 | 50 | 3 | 75 | 4 | 100 | 3 | 50 | 2 | 15 | 6 | 55 | 8 | 29 | 27 | 40 |

*Note.* Numbers inside the parentheses are percentages and the 2018, 2020, and Total columns do not add up to 100% due to round-off errors.

As shown in Table 2, by and large, less than half of the articles (40%) did not report how they contributed to the STEM education literature. Besides, only one article (A14) was designed to extend the findings of past research. More strikingly, none of the articles attempted to replicate past research. When the articles are examined on a yearly basis, it can be seen that 50% or more of the articles did not report how they contributed to the literature in 2014, 2015, 2016, 2017, and 2019. Encouragingly, this percentage decreased drastically in 2018 (15%) and 2020 (29%). It also appears that filling a gap or void in the literature was a more standard reporting practice for the articles published from 2014 to 2020 because, in each year excluding 2016, at least one article used this contribution type. On the other hand, developing new ideas in the literature seems to be a more recent practice since it was only used in the articles published from 2017 to 2020.

### 3.2. Purpose Statements

The breakdown of articles with respect to the formulation of purpose statements and publication years are presented in Table 3. As indicated in Table 3, it is encouraging to find that, all told, most of the articles (81%) contained clearly formulated purpose statements. Besides, the pattern across the 7-year period indicates that each year at least 50% of the articles provided purpose statements and this percentage reached its peak in 2018 (85%), 2019 (91), and 2020 (82%). On

the other hand, eight out of 68 articles (A2, A7, A14, A15, A36, A46, A47, and A52; 12%) included unclear purpose statements.

**Table 3.** *Articles' formulation of purpose statements.*

| Formulation of purpose statements | 2014 | | 2015 | | 2016 | | 2017 | | 2018 | | 2019 | | 2020 | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % |
| Purpose statement is clear | 1 | 50 | 3 | 75 | 3 | 75 | 4 | 67 | 11 | 85 | 10 | 91 | 23 | 82 | 55 | 81 |
| Purpose statement is not clear | - | - | - | - | - | - | 1 | 17 | 2 | 15 | 1 | 9 | 4 | 14 | 8 | 12 |
| Purpose statement is not formulated | 1 | 50 | 1 | 25 | 1 | 75 | 1 | 17 | - | - | - | - | 1 | 4 | 5 | 7 |

*Note.* Numbers inside the parentheses are percentages and the 2017 column does not add up to 100% due to round-off errors.

### 3.3. Research Questions

The distribution of articles with regards to the formulation of research questions and publication years are given in Table 4.

**Table 4.** *Articles' formulation of research questions.*

| Formulation of research questions | 2014 | | 2015 | | 2016 | | 2017 | | 2018 | | 2019 | | 2020 | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % |
| Research questions are formulated | - | - | 3 | 75 | 2 | 50 | 6 | 100 | 13 | 100 | 7 | 64 | 21 | 75 | 52 | 76 |
| Research questions are not formulated | 2 | 100 | 1 | 25 | 2 | 50 | - | - | - | - | 4 | 36 | 7 | 25 | 16 | 24 |

*Note.* Numbers inside the parentheses are percentages.

As depicted in Table 4, overall, about three-quarters of the articles (76%) formulated their research questions. When the articles are examined on a yearly basis, it can be seen that none of the articles specified research questions in 2014, while at least half of them specified research questions from 2015 to 2020. Notably, all of the articles reported research questions in 2017 and 2018, while there was some decrease in articles' research question reporting percentages in 2019 (64%) and 2020 (75%).

### 3.4. Hypotheses

The distribution of articles with respect to the formulation of hypotheses and publication years are given in Table 5.

**Table 5.** *Articles' formulation of hypotheses.*

| Formulation of hypotheses | 2014 | | 2015 | | 2016 | | 2017 | | 2018 | | 2019 | | 2020 | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % |
| Hypotheses are formulated | - | - | - | - | 1 | 25 | 2 | 33 | - | - | - | - | - | - | 3 | 4 |
| Hypotheses are not formulated | 2 | 100 | 4 | 100 | 3 | 75 | 4 | 66 | 13 | 100 | 11 | 100 | 28 | 100 | 65 | 96 |

*Note.* Numbers inside the parentheses are percentages and the 2017 column does not add up to 100% due to round-off errors.

It appears from Table 5 that formulation of hypotheses is not an accepted standard for experimental research articles published in refereed Turkish journals. All in all, only three out of 68 articles (A55, A58, and A62) formulated hypotheses. One of these articles (A62) was published in 2016 and the remaining two articles (A55 and A58) were published in 2017. Furthermore, the hypotheses formulated in these articles were all in null hypothesis ($H_0$) form.

### 3.5. Types of Experimental Research Designs

The distribution of articles by experimental design types and publication years is presented in Table 6.

**Table 6.** *Distribution of experimental design types.*

| Experimental design types | 2014 | | 2015 | | 2016 | | 2017 | | 2018 | | 2019 | | 2020 | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % |
| Weak experimental design | 2 | 100 | 2 | 50 | 2 | 50 | 3 | 50 | 4 | 31 | 6 | 55 | 9 | 32 | 28 | 41 |
| Quasi-experimental design | - | - | 2 | 50 | 2 | 50 | 3 | 50 | 9 | 69 | 5 | 45 | 17 | 61 | 38 | 56 |
| True experimental design | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Not specified | - | - | - | - | - | - | - | - | - | - | - | - | 2 | 7 | 2 | 3 |

*Note.* Numbers inside the parentheses are percentages.

As given in Table 6, altogether, more than half of the STEM education articles (56%) used quasi-experimental designs and roughly 40% of the articles used weak experimental designs. However, none of the articles adopted true experimental designs. In two articles (A68 and A28) experimental research designs were used but their types were not specified. An examination of the trend over the targeted period reveals that the emphasis on weak experimental designs declined gradually from 2014 to 2020 and that quasi-experimental designs became more prevalent in the articles published in recent years, especially in 2020.

### 3.6. Sampling Strategies

The classification of articles by sampling strategy and year of publication is presented in Table 7.

**Table 7.** *Sampling strategies used in the articles.*

| Sampling strategies | | 2014 | | 2015 | | 2016 | | 2017 | | 2018 | | 2019 | | 2020 | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % |
| I | Stratified random sampling | 1 | 50 | 1 | 25 | - | - | - | - | - | - | - | - | - | - | 2 | 3 |
| | Not specified | - | - | - | - | - | - | 1 | 17 | - | - | - | - | 4 | 14 | 5 | 7 |
| II | Convenience sampling | - | - | 1 | 25 | 1 | 25 | 1 | 17 | 2 | 15 | 6 | 55 | 11 | 39 | 22 | 32 |
| | Purposive sampling | - | - | - | - | - | - | 1 | 17 | 1 | 8 | 1 | 9 | 2 | 7 | 5 | 7 |
| | Not specified | - | - | - | - | - | - | 1 | 17 | - | - | - | - | 1 | 4 | 2 | 3 |
| Not reported | | 1 | 50 | 2 | 50 | 3 | 75 | 2 | 33 | 10 | 77 | 4 | 36 | 10 | 36 | 32 | 47 |

*Note.* I represents random sampling strategies and II represents nonrandom sampling strategies. Numbers inside the parentheses are percentages and the 2017 and Total columns do not add up to 100% due to round-off errors.

As seen in Table 7, on the whole, nearly half of the articles (47%) did not report the sampling strategy used. Meanwhile, about one-third of them (32%) used samples of convenience, generally intact classrooms that are easily available to the STEM education researchers, and only 10% of them used random sampling strategies. From 2014 to 2020, each year at least one-third of the articles did not report their sampling strategies. Convenience sampling was used in all years excluding 2014 and the tendency to use this sampling strategy increased drastically in 2019 (55%) and 2020 (39%). A year-by-year examination also shows that, excluding 2020, each year either one or none of the articles used random sampling strategies. Purposive sampling was used in 2017 and thereafter. In 2017, 2018, and 2019 one article and in 2020 two articles reported the use of this sampling strategy.

### 3.7. Description of Samples

As mentioned previously, the selected STEM education articles recruited either K-12 students or pre-service teachers as their subjects. The level of description of K-12 students in the selected articles is presented in Table 8.

**Table 8.** *The level of description of K-12 students in the selected articles.*

| Level of description | 2014 | | 2015 | | 2016 | | 2017 | | 2018 | | 2019 | | 2020 | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % |
| Poor description | - | - | 2 | 67 | 1 | 25 | 1 | 33 | 2 | 25 | 1 | 17 | 8 | 36 | 15 | 32 |
| Mediocre description | 1 | 100 | - | - | 3 | 75 | 2 | 67 | 6 | 75 | 5 | 83 | 14 | 64 | 31 | 66 |
| Rich description | - | - | 1 | 33 | - | - | - | - | - | - | - | - | - | - | 1 | 2 |

*Note.* Poor description refers to reporting at most three different sample characteristics, mediocre description refers to reporting four, five, or six different sample characteristics, and rich description refers to reporting more than six different sample characteristics. Numbers inside the parentheses are percentages.

As shown in Table 8, overall, 31 out 47 articles (66%) provided mediocre description, about 30% of the articles provided poor description, and only one article provided rich description for their samples. It also appears from this table that there was not a detectable pattern in terms of sample description across the 7-year period. The level of description of pre-service teachers in the selected articles is presented in Table 9.

**Table 9.** *The level of description of pre-service teachers in the selected articles.*

| Level of description | 2014 | | 2015 | | 2016 | | 2017 | | 2018 | | 2019 | | 2020 | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % |
| Poor description | - | - | - | - | - | - | 1 | 33 | 1 | 20 | - | - | 1 | 17 | 3 | 14 |
| Mediocre description | 1 | 100 | 1 | 100 | - | - | 2 | 67 | 4 | 50 | 4 | 80 | 3 | 50 | 15 | 71 |
| Rich description | - | - | - | - | - | - | - | - | - | - | 1 | 20 | 2 | 33 | 3 | 14 |

*Note.* Numbers inside the parentheses are percentages and the Total column does not add up to 100% due to round-off errors.

As given in Table 9, altogether, 15 out of 21 articles (71%) provided a mediocre description and three articles provided a rich description for their samples, while the remaining three articles poorly described their samples. A year-by-year examination shows that there is an increasing trend towards mediocre description in 2017 and 2018 and towards rich description in 2019 and 2020.

### 3.8. Populations

The breakdown of articles with respect to defining a population and year of publication is presented in Table 10.

**Table 10.** *Articles' description of their populations.*

| Description of populations | 2014 | | 2015 | | 2016 | | 2017 | | 2018 | | 2019 | | 2020 | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % |
| Population is defined | 1 | 50 | 1 | 25 | - | - | 2 | 33 | - | - | 1 | 9 | 3 | 11 | 8 | 12 |
| Population is not defined | 1 | 50 | 3 | 75 | 4 | 100 | 4 | 67 | 13 | 100 | 10 | 91 | 25 | 89 | 60 | 88 |

*Note.* Numbers inside the parentheses are percentages.

As demonstrated in Table 10, nearly 90% of the articles did not define their populations. Moreover, it appears that reporting populations became an almost overlooked research reporting practice, especially in the last three years. Of the eight articles that reported population, six described very narrow populations (i.e., A4, A6, A10, A39, A53, and A67). That is, these articles reported their populations as only the schools or faculties from which the samples were selected.

### 3.9. Validity and Reliability of Data Collection Instruments

The selected STEM education articles either used the instruments that were developed in previous studies (i.e., pre-existing instruments) or developed their own instruments to collect their own data. Eighty-nine pre-existing instruments (74%), 32 self-developed instruments (26%), and altogether 121 different instruments were used in these articles. Unfortunately, for the pre-existing instruments, only three articles (i.e., A30, A49, and A53) reported the validity

of data collected from their own samples, while the rest of them solely reported the validity of data obtained from the original studies. The validity types reported for the instruments used in the STEM education articles are shown in Table 11.

**Table 11.** *The validity types reported for the instruments used in the STEM education articles.*

| Validity of the instruments | 2014 | | 2015 | | 2016 | | 2017 | | 2018 | | 2019 | | 2020 | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % |
| Likert scale | | | | | | | | | | | | | | | | |
| Construct validity | 1 | 50 | - | - | 1 | 100 | 7 | 70 | 7 | 41 | 11 | 85 | 15 | 60 | 42 | 62 |
| Content and construct validity | - | - | - | - | - | - | - | | - | - | - | - | 1 | 4 | 1 | 1 |
| Not reported | 1 | 50 | - | - | - | - | 3 | 30 | 10 | 59 | 2 | 15 | 9 | 36 | 25 | 37 |
| Achievement test | | | | | | | | | | | | | | | | |
| Content and construct validity | - | - | 2 | 40 | - | - | 3 | 100 | 6 | 60 | 3 | 75 | 7 | 39 | 21 | 50 |
| Content validity | - | - | - | - | - | - | - | - | - | - | - | - | 2 | 11 | 2 | 5 |
| Construct validity | 1 | 100 | - | - | - | - | - | - | - | - | - | - | 2 | 11 | 3 | 7 |
| Not reported | - | - | 3 | 60 | 1 | 100 | - | - | 4 | 40 | 1 | 25 | 7 | 39 | 16 | 38 |
| Questionnaire | | | | | | | | | | | | | | | | |
| Content validity | - | - | - | - | - | - | 1 | 100 | - | - | - | - | - | - | 1 | 17 |
| Not reported | - | - | - | - | 4 | 100 | - | - | - | - | - | - | 1 | 100 | 5 | 83 |
| Performance test | | | | | | | | | | | | | | | | |
| Content validity | - | - | - | - | - | - | - | - | - | - | - | - | 2 | 100 | 2 | 100 |
| Semantic differential scale | | | | | | | | | | | | | | | | |
| Construct validity | - | - | - | - | 1 | 100 | - | - | - | - | - | - | - | - | 1 | 100 |
| Ability test | | | | | | | | | | | | | | | | |
| Not reported | - | - | - | - | - | - | - | - | - | - | - | - | 1 | 100 | 1 | 100 |

*Note.* Thirty-two articles used only one instrument, 24 articles used two different instruments, 10 articles used three different instruments, and 2 articles used 5 different instruments. Numbers inside the parentheses are percentages.

As presented in Table 11, the STEM education articles mainly used Likert scales and achievement tests as data collection instruments. Overall, for a large proportion of the Likert scales (62%), only construct validity was reported. Both content and construct validity were provided for only one Likert scale. On the other hand, for a considerable proportion of Likert scales (37%), validity information was not provided. On a yearly basis, construct validity again seems to be a more standard validity reporting practice for the articles that used Likert scales as data collection instruments. Altogether, for half of the achievement tests used (50%), both content and construct validity were reported. On a yearly basis, reporting content and construct validity was also a predominant practice for the achievement tests used in the articles published from 2014 to 2020.

Pre-existing instruments used in the articles were Likert scales, achievement tests, questionnaires, and semantic differential scales. Sixty-seven pre-existing Likert scales were used in the articles. For 29 of them (40%), reliability information from both the original studies and their own data were reported. For 27 of them (40%), only reliability information from the original studies was provided. For 7 of them (10%), only reliability information from their own data was reported. For 4 of them (6%) reliability information was not provided. Sixteen pre-existing achievement tests were used in the articles. For 11 of them (65%), only reliability information from the original studies was provided. For 4 of them (24%), reliability information from both the original studies and their own data was reported. For one of them (6%), only reliability information for its own data was provided. For the remaining one (6%), reliability information was not provided. Five pre-existing questionnaires were used in the articles. For 4 of them (80%), only reliability information for their own data was provided and for the remaining one (20%), reliability information was not reported. Finally, one pre-existing semantic differential scale was used in A61 and reliability information from both the original study and its own data was reported.

Self-developed instruments used in the articles were achievement tests, performance tests, ability tests, and Likert scales. Twenty-six self-developed achievement tests were used in the selected STEM education articles and for 22 of them (85%), reliability information was provided. Two self-developed performance tests were used in A15 and reliability information was provided for both of them. One self-developed Likert scale was used in A22 and one self-developed questionnaire was used in A57 and reliability information was reported for both instruments. Finally, one self-developed ability test was used in A11. However, reliability information was not reported for this instrument. The reliability types reported for the instruments used in the selected STEM education articles are given in Table 12.

**Table 12.** *The reliability types reported for the instruments used in the STEM education articles.*

| Reliability of the instruments | 2014 | | 2015 | | 2016 | | 2017 | | 2018 | | 2019 | | 2020 | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % |
| Likert scale | | | | | | | | | | | | | | | | |
| Cronbach's alpha | 1 | 50 | - | - | 1 | 100 | 8 | 80 | 15 | 88 | 13 | 100 | 23 | 92 | 61 | 90 |
| Cronbach's alpha and test-retest | - | - | - | - | - | - | - | - | 1 | 6 | - | - | 1 | 4 | 2 | 3 |
| Internal consistency not specified | - | - | - | - | - | - | - | - | 1 | 6 | - | - | - | - | 1 | 1 |
| Not reported | 1 | 50 | - | - | - | - | 2 | 20 | - | - | - | - | 1 | 4 | 4 | 6 |
| Achievement test | | | | | | | | | | | | | | | | |
| KR-20 | 1 | 100 | 1 | 20 | - | - | 1 | 33 | 9 | 90 | 2 | 50 | 12 | 67 | 26 | 62 |
| Cronbach's alpha | - | - | - | - | - | - | - | - | 1 | 10 | 2 | 50 | 3 | 17 | 6 | 14 |
| Inter-rater agreement | - | - | - | - | - | - | - | - | - | - | - | - | 2 | 11 | 2 | 5 |
| KR-21 | - | - | - | - | - | - | 1 | 33 | - | - | - | - | - | - | 1 | 2 |
| Internal consistency not specified | - | - | 1 | 20 | 1 | 100 | - | - | - | - | - | - | - | - | 2 | 5 |
| Not reported | - | - | 3 | 60 | - | - | 1 | 33 | - | - | - | - | 1 | 6 | 5 | 12 |
| Questionnaire | | | | | | | | | | | | | | | | |
| Cronbach's alpha | - | - | - | - | 3 | 75 | - | - | - | - | - | - | - | - | 3 | 50 |
| Inter-rater agreement | - | - | - | - | - | | 1 | 100 | - | - | - | - | 1 | 100 | 2 | 33 |
| Not reported | - | - | - | - | 1 | 25 | - | - | - | - | - | - | - | - | 1 | 17 |
| Performance test | | | | | | | | | | | | | | | | |
| Inter-rater agreement | - | - | - | - | - | - | - | - | - | - | - | - | 2 | 100 | 2 | 100 |
| Semantic differential scale | | | | | | | | | | | | | | | | |
| Cronbach's alpha | - | - | - | - | 1 | 100 | - | - | - | - | - | - | - | - | 1 | 100 |
| Ability test | | | | | | | | | | | | | | | | |
| Cronbach's alpha | - | - | - | - | - | - | - | - | - | - | - | - | 1 | 100 | 1 | 100 |

*Note.* Thirty-two articles used only one instrument, 24 articles used two different instruments, 10 articles used three different instruments, and 2 articles used 5 different instruments. Numbers inside the parentheses are percentages and the 2017 and 2020 columns for achievement tests do not add up to 100 due to round-off errors.

Table 12 indicates that the selected STEM education articles mainly used Likert scales and achievement tests as data collection instruments and that they mostly used internal consistency estimates when reporting reliability. More specifically, for most of the Likert scales (93%), Cronbach's alpha estimates were calculated. Similarly, for more than half of the achievement tests (64%), Kuder-Richardson formulas (KR-20 and KR-21) were used. When examined per year, it can be seen that the tendency to use Cronbach's alphas and KR-20s for reporting reliabilities of Likert scales and achievement tests is especially more evident in the last three years (i.e., between 2018 and 2020). On the other hand, other methods such as test-retest and inter-rater agreement were seldom used to report reliabilities of data collection instruments.

### 3.10. Basic Assumptions of Parametric Tests

Table 13 demonstrates the extent to which the basic assumptions of the parametric tests are fulfilled in the selected articles. As shown in Table 13, the STEM education articles mainly used paired-samples *t*-test (55%) and independent-samples *t*-test (28%) as parametric tests. However, only less than half of the articles using paired-samples *t*-test (46%) could fulfill the corresponding basic assumptions. Similarly, only 37% of the articles using independent-

samples *t*-test satisfied the basic assumptions related to this test. On an annual basis, there seems to appear a stable trend for the STEM education articles towards not fulling the basic assumptions of parametric tests. For the paired-samples *t*-test, every year, at least nearly half of the articles did not check the basic assumptions. For independent-samples *t*-test, this is far more manifest because, each year excluding 2019, articles not fulling the basic assumptions outnumbered the ones that fulfilled the basic assumptions.

**Table 13.** *The fulfillment of basic assumptions of parametric tests used in the STEM education articles.*

| Fulfillment of basic assumptions | 2014 | | 2015 | | 2016 | | 2017 | | 2018 | | 2019 | | 2020 | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % |
| Paired-samples t-test | | | | | | | | | | | | | | | | |
| Fulfilled | - | - | 2 | 67 | - | - | 2 | 50 | 1 | 14 | 4 | 57 | 8 | 57 | 17 | 46 |
| Not fulfilled | 2 | 100 | 1 | 33 | - | - | 2 | 50 | 6 | 86 | 3 | 43 | 6 | 43 | 20 | 54 |
| Independent-samples t-test | | | | | | | | | | | | | | | | |
| Fulfilled | - | - | - | - | - | - | - | - | 1 | 17 | 3 | 100 | 3 | 38 | 7 | 37 |
| Not fulfilled | - | - | 1 | 100 | - | - | 1 | 100 | 5 | 83 | - | - | 5 | 63 | 12 | 63 |
| One-way between-groups ANOVA | | | | | | | | | | | | | | | | |
| Fulfilled | - | - | - | - | - | - | 1 | 100 | 1 | 100 | - | - | 1 | 100 | 3 | 100 |
| Not fulfilled | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| One-way repeated-measures ANOVA | | | | | | | | | | | | | | | | |
| Fulfilled | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Not fulfilled | - | - | - | - | - | - | - | - | - | - | - | - | 1 | 100 | 1 | 100 |
| One-way ANCOVA | | | | | | | | | | | | | | | | |
| Fulfilled | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Not fulfilled | - | - | - | - | - | - | - | - | 1 | 100 | - | - | - | - | 1 | 100 |
| Two-way ANCOVA | | | | | | | | | | | | | | | | |
| Fulfilled | - | - | - | - | - | - | - | - | - | - | - | - | 1 | 100 | 1 | 100 |
| Not fulfilled | | | | | | | | | | | | | - | - | - | - |
| One-way mixed-design ANOVA | | | | | | | | | | | | | | | | |
| Fulfilled | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Not fulfilled | - | - | - | - | - | - | - | - | - | - | - | - | 2 | 100 | 2 | 100 |
| Two-way mixed-design ANOVA | | | | | | | | | | | | | | | | |
| Fulfilled | - | - | - | - | - | - | 1 | 50 | - | - | - | - | - | - | 1 | 33 |
| Not fulfilled | - | - | - | - | - | - | 1 | 50 | 1 | 100 | - | - | - | - | 2 | 67 |

*Note.* Thirty-four articles used only one parametric test, 15 articles used two different parametric tests, one article used three different parametric tests, 16 articles used nonparametric tests, one article used hierarchical linear modeling, and the remaining one used the Wald test for multi-group analysis. Numbers inside the parentheses are percentages and the 2020 column for independent-samples *t*-test does not add up to 100 due to round-off errors.

### 3.11. Sample Sizes

The number of participants used in each group or cell of the selected articles is categorized in Table 14. As can be calculated from Table 14, overall, 60% of the articles had less than 30 participants in their experimental and/or control groups. In 38% of the articles, the number of participants in each group ranged between 20 and 29 (38%), while only three articles used 100 or more participants in each group. Table 14 also shows that there does not appear an increasing and deliberate attempt to use at least 30 participants in experimental and control groups from 2014 to 2020.

**Table 14.** *The number of participants used in each group or cell.*

| Number of participants | 2014 | | 2015 | | 2016 | | 2017* | | 2018* | | 2019* | | 2020 | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % |
| 10–19 participants | - | - | - | - | - | - | - | - | 3 | 19 | 4 | 33 | 9 | 32 | 16 | 22 |
| 20–29 participants | 2 | 100 | - | - | 2 | 50 | 3 | 43 | 7 | 44 | 4 | 33 | 10 | 36 | 28 | 38 |
| 30–39 participants | - | - | 1 | 25 | - | - | 1 | 14 | 6 | 38 | - | - | 5 | 18 | 13 | 18 |
| 40–49 participants | - | - | 1 | 25 | - | - | 1 | 14 | - | - | 1 | 8 | 1 | 4 | 4 | 5 |
| 50–99 participants | - | - | 1 | 25 | - | - | 2 | 29 | - | - | 3 | 25 | 3 | 11 | 9 | 12 |
| 100 or more participants | - | - | 1 | 25 | 2 | 50 | - | - | - | - | - | - | - | - | 3 | 4 |

*Note.* In columns marked with *, A38 (2019), A40 (2018), A45 (2018), A46 (2018), and A53 (2017) were counted twice due to unequal number of participants in the experimental (EG) and control groups (CG). In A38, EG = 20 and CG = 13. In A40, EG1 = 28, EG2 = 33, and CG = 26. In A45, EG = 28 and CG = 30. In A46, EG = 34 and CG = 22. In A53, EG1 = 30, EG2 = 26, and CG = 22. Numbers inside the parentheses are percentages and the 2017, 2018, 2019, and Total columns do not add up to 100% due to round-off errors.

### 3.12. Effect Sizes

The STEM education articles' reporting of effect sizes for the parametric tests they used are presented in Table 15. As depicted in Table 15, overall, more than half of the articles did not report effect sizes for the paired-samples *t*-tests (65%), independent-samples *t*-tests (63%), and two-way mixed-design ANOVAs (67%) they used. For one-way between-groups ANOVA, one-way repeated-measures ANOVA, and one-way ANCOVA, none of the articles reported effect sizes. All of the articles that used two-way ANCOVA (i.e., A27) and one-way mixed-design ANOVA (i.e., A15 and A24) reported effect sizes. When examined on a yearly basis, it can be seen that, each year excluding 2018, more than half of the articles did not report effect sizes for the paired-samples *t*-tests they conducted. Moreover, on a yearly basis, more than 60% of the articles that used independent-samples *t*-test did not report effect sizes.

**Table 15.** *Articles' reporting of effect sizes for the parametric tests used.*

| Effect sizes reported | 2014 | | 2015 | | 2016 | | 2017 | | 2018 | | 2019 | | 2020 | | Total | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % | *n* | % |
| Paired-samples *t*-test | | | | | | | | | | | | | | | | |
| Reported | - | - | 1 | 33 | - | | 1 | 25 | 4 | 57 | 3 | 43 | 4 | 29 | 13 | 35 |
| Not reported | 2 | 100 | 2 | 67 | - | | 3 | 75 | 3 | 43 | 4 | 57 | 10 | 71 | 24 | 65 |
| Independent-samples *t*-test | | | | | | | | | | | | | | | | |
| Reported | - | - | - | - | - | | 1 | 100 | 2 | 33 | 1 | 33 | 3 | 38 | 7 | 37 |
| Not reported | - | - | 1 | 100 | - | | - | | 4 | 67 | 2 | 67 | 5 | 63 | 12 | 63 |
| One-way between-groups ANOVA | | | | | | | | | | | | | | | | |
| Reported | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Not reported | - | - | - | - | - | - | 1 | 100 | 1 | 100 | - | - | 1 | 100 | 3 | 100 |
| One-way repeated-measures ANOVA | | | | | | | | | | | | | | | | |
| Reported | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Not reported | - | - | - | - | - | - | - | - | - | - | - | - | 1 | 100 | 1 | 100 |
| One-way ANCOVA | | | | | | | | | | | | | | | | |
| Reported | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Not reported | - | - | - | - | - | - | - | - | 1 | 100 | - | - | - | - | 1 | 100 |
| Two-way ANCOVA | | | | | | | | | | | | | | | | |
| Reported | - | - | - | - | - | - | - | - | - | - | - | - | 1 | 100 | 1 | 100 |
| Not reported | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| One-way mixed-design ANOVA | | | | | | | | | | | | | | | | |
| Reported | - | - | - | - | - | - | - | - | - | - | - | - | 2 | 100 | 2 | 100 |
| Not reported | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | |
| Two-way mixed-design ANOVA | | | | | | | | | | | | | | | | |
| Reported | - | - | - | - | - | - | 1 | 50 | - | - | - | - | - | - | 1 | 33 |
| Not reported | - | - | - | - | - | - | 1 | 50 | 1 | 50 | - | - | - | - | 2 | 67 |

*Note.* Numbers inside the parentheses are percentages.

### 3.13. Statistical Powers

Regrettably, of the 68 articles reviewed, none reported the statistical power of the parametric tests used. This indicates that statistical power is an overlooked statistical measure for the experimental STEM education articles published in national indexed journals.

## 4. DISCUSSION and IMPLICATIONS

In this study, we examined the methodological quality of experimental STEM education articles published in the refereed Turkish journals from 2014 to 2020. In this way, we attempted to reveal the degree to which current research reporting practices in these journals are in agreement with the standards of educational research described in the commonly used methodology textbooks. In what follows, we discuss the findings in light of the literature on methodological quality.

### 4.1. Contributions to the Literature

The present study found that 40% of the articles did not clarify how they contributed to the STEM education literature. This implies that national indexed journals are largely lacking a systematic effort to build a cumulative knowledge base in the area of STEM education. What is more, researchers who are publishing in these journals may not be aware that similar or related research might have been or is being conducted elsewhere by other colleagues and it is most likely that they will continue to conduct isolated studies in the future. The researchers' failure to relate their studies to past research may have serious consequences on the quality of STEM education research. As emphasized by Nelson and Shaver (1985), not clarifying contributions to the literature may lead to the "repetition of unproductive prior research and a disconnectedness of studies on similar topics" (p. 410). Thus, conducting isolated studies on STEM education may highly be counterproductive to knowledge building in this area.

We also found that none of the articles reported a direct or systematic replication of previous research. Although replication can be used as a strategy to compensate for weaknesses in generalizability (Horton et al., 1993), it was not accepted as a research reporting practice by the STEM education researchers who publish in the refereed Turkish journals. Several other researchers also found that replication is a neglected practice in educational research. For instance, Horton et al. (1993) examined the methodological quality of articles published in the Journal of Research in Science Teaching from 1985 to 1989 and found that only four of them (3%) replicated past research. Similarly, Shaver and Norton (1980) examined two social studies journals and found that only four (13%) and three (14%) of the articles in these journals replicated previous research.

### 4.2. Purpose Statements, Research Questions, and Hypotheses

We found that around 80% of the STEM education articles contained clearly formulated purpose statements. It is encouraging to find that a large proportion of the articles included purpose statements. These statements clarify the primary objective or focus of our research and thus are the most important ones in research studies. Moreover, they signal the procedures we should use during data collection and they point to the types of findings we expect to obtain in our research (Creswell, 2015).

It is also good news that roughly 75% of the articles formulated their research questions. Research questions hint at the methodology used in a research study and to the data analysis methods that are relevant to that study (Aktemur, 2015). For example, the research question "what is the effect of X on Y?" infers an experimental research methodology and subsequently the statistical tests used during the analysis of data such as *t*-tests, ANOVAs, and ANCOVAs. Therefore, it can be said that three-quarters of the STEM education articles enabled the audience

to easily determine whether the research methodologies and corresponding data analysis procedures used in them were correct or not.

On the other hand, the findings revealed that only three (4%) articles formulated their hypotheses. It seems that formulating a hypothesis is not within the STEM education authors' research reporting routines. Hypotheses refer to researchers' expectations about how specific phenomena work and affect, while experiments are procedures conducted to confirm, rebut, or ascertain the validity of these hypotheses (Horváth, 2016). Thus, hypotheses are an important key tenet of experimental designs. For this reason, well-formulated hypotheses are compulsory for carrying out more rigorous experimental studies.

### 4.3. Types of Experimental Research Designs

It is sobering to find that none of the STEM education articles used true experimental designs. One possible reason for the absence of true experimental designs in the STEM education articles might be that researchers find it difficult to obtain random samples of students for their studies. However, as argued by Campbell and Boruch (1975), school settings do provide natural laboratories in which random assignment could usually be used.

We also found that the most frequently used research design was a quasi-experimental design (56%). This finding is in line with the findings of previous research that examined the design quality of experimental studies (e.g., Sung et al., 2019) or research trends in education (e.g., Baydaş et al., 2015; Duman et al., 2015). For instance, Sung et al. (2019) investigated the quality of experimental designs in mobile learning research from 2006 to 2016 and revealed that 63% of the articles used a quasi-experimental design. Baydaş et al. (2015) examined educational technology research trends from 2002 to 2014 and similarly found that 48% of the experimental studies used a quasi-experimental design. Moreover, Duman et al. (2015) analyzed research trends in the mobile-assisted language learning articles published from 2000 to 2012 and found that 12 out of 26 experimental studies (46%) employed a quasi-experimental design.

Alarmingly, we found that almost half of the articles (41%) used weak experimental designs. However, an important drawback of weak experimental designs is that they are subject to numerous threats to validity. Mills and Gay (2019) cautioned that weak experimental designs should be avoided as much as possible. They further indicated that the findings obtained from studies using weak experimental designs are very questionable and thus "they are not useful for most purposes except, perhaps, to provide a preliminary investigation of a problem" (p. 310). Weak experimental designs provide little or no control of extraneous variables (Ary et al., 2014) and consequently, it is almost impossible to refute rival hypotheses or explanations. Thus, extreme caution must be exercised in interpreting and generalizing the findings obtained from the weak experimental STEM education articles that were analyzed in the current study.

### 4.4. Description of Sampling Strategies, Sample Characteristics, and Populations

The present study found that 42% of the articles used nonrandom sampling strategies. More specifically, nearly one-third of the articles (32%) used convenience sampling. These high percentages signal that many of the articles published in national indexed journals from 2014 to 2020 used biased samples because such samples almost always differ systematically from the population with respect to particular characteristics (Johnson & Christensen, 2020). Besides, it is a major error to use significance tests in studies where samples are obtained nonrandomly (Wallen & Fraenkel, 1988). Random sampling is the fundamental basis of inferential statistics and "one must raise serious questions about the use of inferential statistics when the lack of randomness makes probability statements indeterminate" (Shaver & Norton, 1980, p. 8). In the present study, only seven articles (10%) used random samples. Thus, it can be said that only these articles can appropriately use the significance tests.

We also found that almost 90% of the articles did not define their populations and only a very small portion of the articles provided rich descriptions for their samples (i.e., one article for K-12 students and three articles for pre-service teachers). Providing rich descriptions for the samples being studied is crucial because it offers researchers some basis for deciding whether their findings are generalizable to the relevant contexts (Shaver & Norton, 1980). More explicitly, describing the details of experimental studies or the contexts of nonexperimental studies as vividly as possible may help other researchers evaluate the applicability of particular findings to their own situations (Fraenkel, 1987). Thus, our findings demonstrate that relevant generalizations beyond the samples described may indeed not be possible for most of the STEM education articles published in the national indexed articles. These findings are not surprising given that such reporting practices were found to be extremely widespread in many of the previous studies conducted on methodological quality (e.g., Aktemur, 2015; Horton et al., 1993; Shaver & Norton, 1980; Wallen & Fraenkel, 1988).

### 4.5. Validity and Reliability of Data Collection Instruments

Sound measurement is the keystone of rigorous research and it is very significant for high-quality experimentation (D'agostino, 2005). In addition, sound measurements rely on the validity and reliability of instruments. Thus, improving the validity and reliability of instruments enhances the rigor of research and the quality of experiments in particular (D'agostino, 2005). In the current study, the selected STEM education articles mainly used Likert scales and achievement tests as data collection instruments. Despite the emphasis on using valid instruments, validity information was not documented for 37% and 38% of these Likert scales and achievement tests, respectively. Therefore, a considerable proportion of the Likert scales and achievement tests used in the STEM education articles might have jeopardized the validity of conclusions drawn from these articles. Frankly speaking, the STEM education articles might have used faulty Likert scales and achievement tests and these instruments might have yielded biased outcomes (D'agostino, 2005). Similar findings were obtained in previous studies about the documentation of instrument validity (e.g., Sung et al., 2019; Wallen & Fraenkel, 1988). For instance, Wallen and Fraenkel (1988) examined articles published in Theory and Research in Social Education over an eight-year period and found out that only 30% of the articles empirically checked the validity of instruments used. Similarly, Sung et al. (2019) revealed that only 24% of the mobile-learning articles provided information about instrument validity.

Encouragingly, we found that reliability information was provided for 94% of the Likert scales and 88% of the achievement tests used in the STEM education articles. More specifically, Cronbach's alpha reliabilities were reported for all types of instruments without any exceptions. However, test-retest reliability and inter-rater reliability were seldom reported and split-half reliability was never reported in these articles. These reliability types deal with different kinds of test consistencies. For instance, test-retest reliability measures the stability of scores over time and in high-quality journals, it is almost always reported in company with internal consistency reliability (Johnson & Christensen, 2020). Moreover, Mills and Gay (2016) stress that test-retest reliability is particularly crucial for instruments that are used for making predictions because predictions are based to a large extent on the assumption that the scores are stable over time.

### 4.6. Basic Assumptions of Parametric Tests

In this study, we used Sung et al.'s (2019) judgment tree to determine how well the selected STEM education articles fulfilled the basic assumptions of the parametric tests used. We found that 17 out of 37 articles (46%) met the basic assumptions of the paired-samples *t*-test and 7 out of 19 articles (37%) met the basic assumptions of the independent samples *t*-test. Moreover, for

one-way repeated-measures ANOVA, one-way ANCOVA, and one-way mixed-design ANOVA, the basic assumptions were not satisfied by any one of the articles. This finding shows that the selected STEM education articles largely overlooked the significance of basic assumptions when conducting certain parametric tests. However, violation of the basic assumptions gives rise to invalid probability inferences from these parametric tests (Aron et al., 2019). For instance, ANCOVA mandates that the relationship between the dependent variable and the covariate must be the same for each group (i.e., homogeneity of regression slopes). Unequal regression slopes demonstrate that there is an interaction between the treatment and the covariate and that the findings will be misleading in case ANCOVA is conducted (Tabachnick & Fidell 2019).

Similarly, one-way repeated-measures ANOVA necessitates that "variance of the population difference scores for any two conditions are the same as the variance of the population difference scores for any other two conditions" (i.e., sphericity; Pallant, 2016, p. 287). If this assumption is violated, one-way repeated-measures ANOVA will become too liberal and provoke inflation of Type I error rates (Tabachnick & Fidell, 2019). In such cases, significance tests such as Greenhouse-Geisser or Huynh-Feldt may be used as alternatives to avoid biased conclusions (Tabachnick & Fidell, 2019). Nevertheless, the articles that violated the basic assumptions of one-way repeated-measures ANOVA (i.e., A22), one-way mixed-design ANOVA (i.e., A15 and A24), and two-way mixed-design ANOVA (i.e., A43 and A56) did not report the use of such corrective measures.

Our findings show, by and large, that fulfilling the basic assumptions of the statistical tests is not a cut-and-dried practice for most of the STEM education articles that we examined. The findings also demonstrate that the STEM education articles failed to make appropriate adjustments when they did not meet the basic assumptions of the parametric tests and therefore, they seriously threatened the validity of findings on STEM education.

### 4.7. Sample Sizes, Effect Sizes, and Statistical Powers

Although many educational research textbooks (e.g., Ary et al., 2014; Fraenkel et al., 2012; Mills & Gay, 2016) recommend a minimum of 30 participants in each of the experimental and control groups, the current study revealed that the most proportion of the articles (38%) had 20–29 participants in their groups. This shows that using 20–29 participants for each cell might have been accepted as a sample size standard for the STEM education articles published in national indexed journals. What is worse, 22% of the articles had 10–19 participants in their groups. This is especially alarming because studies that recruit an inadequate number of participants for each cell or group are most likely to produce low statistical power (Cheung & Slavin, 2012). For instance, a medium effect size ($0.2 < d < 0.8$), as most commonly reported in educational research studies, with 25–30 participants produces a statistical power of 0.47 (Cohen, 1988). Since this value is below 50%, it appears that for 60% of the selected STEM education articles, the probability of correctly detecting a real treatment effect is worse than guessing. This alerts that the selected STEM education articles may highly be prone to Type II errors.

Moreover, due to the existence of *between-persons errors*, recruiting a fewer number of participants for between-subjects designs may have far more serious consequences on the accuracy of findings compared to within-subject designs and mixed-designs (Privitera, 2019). For between-subjects designs and the related parametric tests (e.g., independent-samples *t*-tests, between-groups ANOVAs, and ANCOVAs) at least 128 participants must be recruited to obtain a statistical power of 0.80 when there is a medium effect size (see Table 1). In the present study, we found that 19 articles (28%) used independent-samples t-tests, three articles used between-groups ANOVAs (4%), and 2 articles (3%) used ANCOVAs for their between-subjects designs. Since such parametric tests were often used in STEM education articles, future studies in this

area must recruit a greater number of participants to achieve the widely accepted standard of statistical power of 0.8 for these tests.

Our findings also indicated that more than 60% of the articles did not report effect sizes for each of the parametric tests used (see Table 15). This demonstrates that the authors of these articles relied solely on statistical significance tests when reporting their findings. However, statistical significance tests alone would be misleading and may lead to many different conclusions. Namely, studies with large sample sizes can readily reach statistical significance despite demonstrating very small practical significance. Conversely, studies with very few sample sizes may not achieve statistical significance despite having very large practical significance. Unlike statistical significance tests, effect sizes are not contingent upon sample sizes (Gravetter et al., 2021). Besides, since effect sizes add "a more exact numerical statement of facts" (Hanel & Mehler, 2019, p. 469), they are more informative than statistical significance (Cohen, 1994). Thus, for a more accurate interpretation of findings, STEM education researchers who publish in national indexed journals should pay more attention to supplementing statistical significance tests with effect sizes.

Ultimately, we found out that none of the STEM education articles reported statistical powers for the parametric tests they conducted. This suggests that the authors of these articles may not have sufficient knowledge and awareness of statistical power. There are several software packages such as G*Power, PASS, Power and Precision, and nQuery to calculate powers of statistical significance tests based on sample sizes, effect sizes, and the alpha levels. There are also different ways to assess statistical power. For instance, G*Power 3 allows for the calculation of the following five different types of power analysis: a priori power analyses, post hoc power analyses, compromise power analyses, sensitivity analyses, and criterion analyses (Faul et al., 2007). Currently, the use of statistical power analysis is absent in the national STEM education literature. However, given the accumulated body of information about statistical power and the diversity of computer programs available, there is not any reason to overlook statistical power when planning research studies and analyzing their findings.

### 4.8. Implications

Our findings demonstrate that the experimental STEM education articles published in the refereed Turkish journals from 2014 to 2020 suffer from serious methodological flaws. Thus, the methodological quality of these articles should remain a concern for the STEM education community including authors, journal editors, editorial board members, reviewers, practitioners, readers, and particularly for policymakers and curriculum developers who are responsible for developing and reforming national curricula in the Ministry of National Education.

To improve the methodological quality of STEM education articles, first, authors should develop some competence and awareness in experimental research designs. Universities or other institutions may design workshops and deliver some training to the authors to have them gain substantial expertise in experimental research methodology. In these workshops, several novel and fruitful approaches may be used to help the authors gain a more thorough understanding of experimental research designs. For instance, LaCosse et al.'s (2017) active-learning approach may be used. LaCosse et al. (2017) examined the impact of project-oriented active-learning techniques on psychology undergraduates' understanding of research methods and found that these techniques increased the participants' understanding significantly.

Scholarly journals are a primary means for disseminating research findings. Thus, to improve the rigor of STEM education research, editors, editorial board members, and reviewers may adopt clear and efficient quality guidelines or criteria such as the standards and procedures specified by the What Works Clearinghouse (2020a, 2020b), the Consolidated Standards of

Reporting Trials (Schulz et al., 2010), and the Study Design and Implementation Assessment Device (Valentine & Cooper, 2008). They may use these guidelines, standards, or criteria as requirements for the manuscripts submitted to the journals. Meanwhile, authors may use the coding framework developed in the current study and similar evaluation tools developed in previous studies such as the Checklist for the Rigor of Education-Experiment Designs (Sung et al., 2019) to check the experimental design quality of their manuscripts and remedy the deficiencies existing in their manuscripts before submission. Thus, the coding framework proposed in the present study may particularly serve experimental research authors as an effective self-checking and self-improvement tool.

Ultimately, using valid designs is significant for maintaining the sustainability and practicability of STEM education research. Using rigorous experimental designs will undoubtedly contribute to the development of theories and practices in the area of STEM education. On the other hand, the deficiencies in the experimental designs will be an obstacle to the sustainability of STEM education research. In addition, these deficiencies may mislead the STEM education community about the effectiveness of STEM education practices. In the current study, we found that the selected articles had considerable deficiencies in their research designs. Thus, policymakers and curriculum developers in the Ministry of National Education should be very cautious when using the findings of the STEM education articles published in the national journals and ruminate much on these findings before implementing STEM education curricula in K–12 schools. This is because initiating such reforms in educational environments requires too much time, energy, and resources and if STEM curricula do not lead to superior outcomes, contrary to the existent literature, then all the investment will be wasted.

## Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. Ethical approval for the current study was provided by Nevşehir Hacı Bektaş Veli University Ethics Committee (Date of Application: 31.03.2021, Number of Application: 2100014216).

## Authorship Contribution Statement

The authors worked collaboratively in all phases of the manuscript preparation.

## Orcid

Ramazan AVCU https://orcid.org/0000-0002-0149-5178
Seher AVCU https://orcid.org/0000-0003-4938-7325

## REFERENCES

Ahern, K.J. (1999). Ten tips for reflexive bracketing. *Qualitative Health Research, 9*(3), 407–411. https://doi.org/10.1177/104973239900900309

Akgündüz, D., Aydeniz, M., Çakmakçı, G., Çavaş, B., Çorlu, M.S., Öner, T., & Özdemir, S. (2015). *STEM eğitimi Türkiye raporu: Günün modası mı yoksa gereksinim mi?* [A report on STEM education in Turkey: A provisional agenda or a necessity?]. Scala Press.

Aktemur, Ş. (2015). *Review of aviation research: A content analysis of articles published in the Collegiate Aviation Review, 2007–2012* [Unpublished master's thesis]. Florida Institute of Technology.

Aron, A., Coups, E.J., & Aron, E.N. (2019). *Statistics for the behavioral and social sciences: A brief course* (6th ed.). Pearson.

Ary, D., Jacobs, L.C., Sorensen, C.K., & Walker, D. (2014). *Introduction to research in education* (9th ed.). Wadsworth Cengage Learning.

Aydın Günbatar, S., & Tabar, V. (2019). Türkiye'de gerçekleştirilen STEM araştırmalarının içerik analizi [Content analysis of Science, Technology, Engineering and Mathematics (STEM) research conducted in Turkey]. *Yüzüncü Yıl University Journal of Education, 16*(1), 1054–1083. http://dx.doi.org/10.23891/efdyyu.2019.153

Baydaş, Ö., Küçük, S., Yılmaz, R. M., Aydemir, M., & Göktaş, Y. (2015). Educational technology research trends from 2002 to 2014. *Scientometrics, 105*, 709–725. https://doi.org/10.1007/s11192-015-1693-4

Borman, G.D., Slavin, R.E., Cheung, A., Chamberlain, A.M., Madden, N.A., & Chambers, B. (2005). Success for all: First-year results from the national randomized field trial. *Educational Evaluation and Policy Analysis, 27*(1), 1-22. https://doi.org/10.3102/01623737027001001

Brown, J. (2012). The current status of STEM education research. *Journal of STEM Education: Innovations and Research, 13*(5), 7–11.

Brown, J.R., & Dant, R.P. (2008). On what makes a significant contribution to the retailing literature. *Journal of Retailing, 84*(2), 131-135. https://doi.org/10.1016/j.jretai.2008.05.002

Campbell, D.T., & Boruch, R.F. (1975). Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In C. A. Bennett & A. A. Lumsdaine (Eds.), *Evaluation and experiment: Some critical issues in assessing social programs* (pp. 195–296). Academic Press.

Çavaş, P., Ayar, A., Bula Turuplu, S., & Gürcan, G. (2020). Türkiye'de STEM eğitimi üzerine yapılan araştırmaların durumu üzerine bir çalışma [A study on the status of STEM education research in Turkey]. *Yüzüncü Yıl University Journal of Education, 17*(1), 823–854. https://doi.org/10.33711/yyuefd.751853

Çevik, M. (2018). Impacts of the project based (PBL) science, technology, engineering and mathematics (STEM) education on academic achievement and career interests of vocational high school students. *Pegem Journal of Education and Instruction, 8*(2), 281–306. http://dx.doi.org/10.14527/pegegog.2018.012

Cheung, C.K., & Slavin, R.E. (2012). How features of educational technology applications affect student reading outcomes: A meta-analysis. *Educational Research Review, 7*(3), 198–215. https://doi.org/10.1016/j.edurev.2012.05.002

Coe, R.J. (2021). Effect size. In J. Arthur, M. Waring, R. Coe, & L. V. Hedges (Eds.), *Research methods and methodologies in education* (3rd ed., pp. 368–377). Sage.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education* (8th ed.). Routledge.

Creswell, J.W. (2015). *Educational research: Planning, conducting and evaluating quantitative and qualitative research* (5th ed.). Pearson Education.

Creswell, J.W., & Creswell, J.D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Sage.

D'Agostino, J. (2005). Measuring learning outcomes: Reliability and validity issues. In G. D. Phye, D. H. Robinson, & J. Levin (Eds.). *Empirical methods for evaluating educational interventions* (pp. 113–145). Elsevier Academic Press.

Daşdemir, İ., Cengiz, E., & Aksoy, G. (2018). Türkiye'de FeTeMM (STEM) eğitimi eğilim araştırması [An investigation of research trends in the field of STEM education in

Turkey]. *Yüzüncü Yıl University Journal of Education, 15*(1), 1161–1183. http://dx.doi.org/10.23891/efdyyu.2018.100

Dochy, F. (2006). A guide for writing scholarly articles or reviews for the Educational Research Review. Retrieved March 15, 2021, from https://www.elsevier.com/__data/promis_misc/edurevReviewPaperWriting.pdf

Duman, G., Orhon, G., & Gedik, N. (2015). Research trends in mobile assisted language learning from 2000 to 2012. *ReCALL, 27*(2), 197-216. https://doi.org/10.1017/S0958344014000287

Elmalı, Ş., & Balkan Kıyıcı, F. (2017). Türkiye'de yayınlanmış FeTeMM eğitimi ile ilgili çalışmaların incelenmesi [Review of STEM studies published in Turkey]. *Sakarya University Journal of Education, 7*(3), 684–696. https://doi.org/10.19126/suje.322791

Erdoğan, M., Marcinkowski, T., & Ok, A. (2009). Content analysis of selected features of K–8 environmental education research studies in Turkey, 1997–2007. *Environmental Education Research, 15*(5), 525–548. https://doi.org/10.1080/13504620903085776

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146

Finn, J.D., & Achilles, C.M. (1999). Tennessee's class size study: Findings, implications, misconceptions. *Educational Evaluation and Policy Analysis, 21*, 97–109. https://doi.org/10.3102/01623737021002097

Fraenkel, J.R. (1987). Toward improving research in social studies education. *Theory & Research in Social Education, 15*(3), 203-222. https://doi.org/10.1080/00933104.1987.10505546

Fraenkel, J.R., Wallen, N.E., & Hyun, H.H. (2012). *How to design and evaluate research in education* (8th ed.). McGraw-Hill.

Frey, B.B. (2018). *The SAGE encyclopedia of educational research, measurement, and evaluation.* Sage.

Gall, J.P., Gall, M.D., & Borg, W.R. (2014). *Applying educational research: How to read, do, and use research to solve problems in practice* (6th ed.). Pearson Education.

Gall, M., Gall, J., & Borg, R. (2007). *Educational research: An introduction* (8th ed.). Pearson Education.

Gravetter, F., Wallnau, L., Forzano, L., & Witnauer, J. (2021). *Essentials of statistics for the behavioral sciences* (10 ed.). Cengage Learning.

Hair, J.F., Black, W.C., Babin, B. J., & Anderson, R.E. (2019). *Multivariate data analysis* (8th ed.). Cengage Learning.

Hanel, P.H., & Mehler, D.M. (2019). Beyond reporting statistical significance: Identifying informative effect sizes to improve scientific communication. *Public Understanding of Science, 28*(4), 468–485. https://doi.org/10.1177/0963662519834193

Hedges, L.V., & Schauer, J. (2018). Randomised trials in education in the USA. *Educational Research, 60*(3), 265–275. https://doi.org/10.1080/00131881.2018.1493350

Horton, P.B., McConney, A.A., Woods, A.L., Barry, K., Krout, H.L., & Doyle, B.K. (1993). A content analysis of research published in the Journal of Research in Science Teaching from 1985 through 1989. *Journal of Research in Science Teaching, 30*(8), 857–869. https://doi.org/10.1002/tea.3660300805

Horváth, I. (2016). Theory building in experimental design research. In P. Cash, T. Stankovic, & M. Storga (Eds.), *Experimental design research: Approaches, perspectives, applications* (pp. 209–231). Springer International Publishing.

Howell, D.C. (2017). *Fundamental statistics for the behavioral sciences* (9th ed.). Cengage Learning.

Johnson, R.B., & Christensen, L.B. (2020). *Educational research: Quantitative, qualitative, and mixed approaches* (7th ed.). Sage.

Kaya, A., & Ayar, M.C. (2020). Türkiye örnekleminde STEM eğitimi alanında yapılan çalışmaların içerik analizi [Content analysis of STEM education studies in Turkey]. *İstanbul Aydın University Journal of Education, 6*(2), 275–306.

Kirk, R.E. (2013). *Experimental design: Procedures for the behavioral sciences* (4th ed.). Sage Publications.

LaCosse, J., Ainsworth, S.E., Shepherd, M.A., Ent, M., Klein, K.M., Holland-Carter, L.A., Moss, J.H., Licht, M., & Licht, B. (2017). An active-learning approach to fostering understanding of research methods in large classes. *Teaching of Psychology*, *44*(2), 117–123. https://doi.org/10.1177/0098628317692614

Li, Y., Wang, K., Xiao, Y., & Froyd, J. E. (2020). Research and trends in STEM education: A systematic review of journal publications. *International Journal of STEM Education, 7*, 1. https://doi.org/10.1186/s40594-020-00207-6

Lincoln, Y.S., & Guba, E. (1985). *Naturalistic inquiry*. Sage.

Miles, M.B., Huberman, M.A., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). Sage.

Mills, G.E., & Gay, L.R. (2016). *Educational research: Competencies for analysis and applications* (11th ed.). Pearson Education.

Mills, G.E., & Gay, L.R. (2019). *Educational research: Competencies for analysis and applications* (12th ed.). Pearson.

Ministry of National Education. (2016). *STEM eğitimi raporu* [STEM education report]. Innovation and Educational Technologies General Directorate.

Ministry of National Education. (2018). *Küresel bağlamda STEM yaklaşımları* [STEM approaches in a global context]. Innovation and Educational Technologies General Directorate.

Mizell, S., & Brown, S. (2016). The current status of STEM education research 2013-2015. *Journal of STEM Education*, 17(4), 52–56.

Nelson, J.L., & Shaver, J.P. (1985). On research in social education. In W. B. Stanley (Ed.), *Review of research in social studies education: 1976-1983* (pp. 401–433). National Council for the Social Studies.

Orne, M.T. (1981). The why and how of a contribution to the literature: A brief communication. *International Journal of Clinical and Experimental Hypnosis, 29*(1), 1–4. https://doi.org/10.1080/00207148108409137

Özcan, H., & Koca, E. (2019). The impact of teaching the subject "pressure" with STEM approach on the academic achievements of the secondary school 7th grade students and their attitudes towards STEM. *Education and Science, 44*(198), 201–227. http://dx.doi.org/10.15390/EB.2019.7902

Pagano, R.R. (2013). *Understanding statistics in the behavioral sciences* (10th ed.). Wadsworth Cengage Learning.

Pallant, J. (2016). *SPSS survival manual: A step by step guide to data analysis using the SPSS program* (6th ed.). Open University Press.

Plano Clark, V.L., & Creswell, J.W. (2015). *Understanding research: A consumer's guide* (2nd ed.). Pearson Education.

Privitera, G.J. (2019). *Essential statistics for the behavioral sciences* (2nd ed.). Sage.

Randolph, J.J., Griffin, A.E., Zeiger, S.R., Falbe, K.N., Freeman, N.A., Taylor, B.E., Westbrook, A.F., Lico, C.C., Cristy, N. S., Sprull, N. M., Holt, C., Smith, K., & McAnespie, H. (2013). A methodological review of the articles published in Georgia Educational Researcher from 2003-2010. *Georgia Educational Researcher Online Edition, 10*(1), Article 1. https://doi.org/10.20429/ger.2013.100101

Schulz, K.F., Altman, D.G., & Moher, D. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *Trials, 11*(1), Article 32. https://doi.org/10.1136/bmj.c332

Schweinhart, L.J., Barnes, H.V., & Weikart, D.P. (1993). *Significant benefits: The High/Scope Perry Preschool Study through age 27*. High/Scope Press.

Shapiro, G.M. (2008). Sample size. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (Vol. 2, pp. 781–783). Sage.

Shaver, J.P., & Norton, R.S. (1980). Populations, samples, randomness, and replication in two social studies journals. *Theory & Research in Social Education, 8*(2), 1–10. https://doi.org/10.1080/00933104.1980.10506078

Shukla, A. (2017). Literature review: An oblivious yet grounding task of research. *Management Insight, 13*(1), 7–15. https://doi.org/10.21844/mijia.v13i01.8363

Slavin, R.E. (2008). Perspectives on evidence-based research in education—What works? Issues in synthesizing educational program evaluations. *Educational Researcher, 37*(1), 5–14. https://doi.org/10.3102/0013189X08314117

Stebbins, L.F. (2006). *Student guide to research in the digital age: How to locate and evaluate information sources*. Libraries Unlimited.

Sung, Y.T., Lee, H.Y., Yang, J.M., & Chang, K.E. (2019). The quality of experimental designs in mobile learning research: A systemic review and self-improvement tool. *Educational Research Review, 28*, 100279. https://doi.org/10.1016/j.edurev.2019.05.001

Tabachnick, B.G., & Fidell, L.S. (2019). *Using multivariate statistics* (7th ed.). Pearson.

United States Department of Education. (2020). *ED delivers historic investment in STEM*. https://content.govdelivery.com/accounts/USED/bulletins/2ad85c3

Valentine, J.C., & Cooper, H. (2008). A systematic and transparent approach for assessing the methodological quality of intervention effectiveness research: The study design and implementation assessment device (Study DIAD). *Psychological Methods, 13*(2), 130–149. https://doi.org/10.1037/1082-989X.13.2.130

Wallen, N.E., & Fraenkel, J.R. (1988). An analysis of social studies research over an eight year period. *Theory & Research in Social Education, 16*(1), 1-22. https://doi.org/10.1080/00933104.1988.10505553

Warner, R.M. (2013). *Applied statistics: From bivariate through multivariate techniques* (2nd ed.). Sage.

What Works Clearinghouse. (2020a). *Standards handbook, version 4.1*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Standards-Handbook-v4-1-508.pdf

What Works Clearinghouse. (2020b). *Procedures handbook, version 4.1*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/WWC-Procedures-Handbook-v4-1-508.pdf