In the Name of God

The Merciful

English Language Department

M.A Thesis in English Language Teaching

Exploring Internal, External and Construct Validity of Experimental Findings in ELT Domain

By:

Habibeh Hakimi

Supervisor:

Dr.Seyyed Ali Ostovar-Namaghi

February 2021

شماره: ۳/ل/۹۹/۲۸٤
تاریخ: ۹۹/۱۱/۲۸

باسمه‌تعالی

مدیریت تحصیلات تکمیلی

فرم شماره (۳) صورتجلسه نهایی دفاع از پایان نامه دوره کارشناسی ارشد

با نام و یاد خداوند متعال، ارزیابی جلسه دفاع از پایان نامه کارشناسی ارشد خانم / آقای حبیبه حکیمی با شماره دانشجویی ۹۷۰۶۱۳۴ رشته زبان انگلیسی گرایش آموزش تحت عنوان Exploring internal, external and construct validity of experimental findings in ELT domain که در تاریخ ۱۳۹۹/۱۱/۲۸ با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح ذیل اعلام می‌گردد:

الف) درجه عالی: نمره ۲۰-۱۹ ☑ ب) درجه خیلی خوب: نمره ۱۸/۹۹ – ۱۸ ☐
ج) درجه خوب: نمره۱۷/۹۹– ۱۶ ☐ د) درجه متوسط: نمره ۱۵/۹۹-۱۴ ☐
ه) کمتر از ۱۴ غیر قابل قبول و نیاز به دفاع مجدد دارد ☐
نوع تحقیق: نظری ☐ عملی ☐

عضو هیأت داوران	نام ونام خانوادگی	مرتبهٔ علمی	امضاء
۱ـ استادراهنمای اول	دکتر سید علی اسنوار نامقی	دانشیار	
۲- استادراهنمای دوم			
۳- استاد مشاور			
۴- نماینده تحصیلات تکمیلی	دکتر فاطمه مظفری	استادیار	
۵- استاد ممتحن اول	دکتر ابوطالب ایرانمهر	استادیار	
۶ـ استاد ممتحن دوم	دکتر سید حمزه موسوی	استادیار	

نام و نام خانوادگی رئیس دانشکده:
تاریخ و امضاء و مهر دانشکده:

Dedication

This thesis is dedicated to my husband and my mother for all their supports, patience and kindness they had during two years of my MA.

Acknowledgment

First and uppermost, I would like to thank Allah, who has always supported me during my life. I would like to express my special appreciation to my supervisor Dr. Seyyed Ali Ostovar-Namaghi who gave me the best opportunity to do this project. Not only was he the best teacher ever for me but also a perfect inspiration of success. I would also like to thank my husband and mother who helped me a lot in finalizing this project within the limited time frame. Finally, I would also like to acknowledge Dr. Iranmehr and Dr. Mousavi form the English department at Shahrood University of Technology, and I am thankfully indebted to them for their very insightful comments and attentive feedback on this thesis.

تعهدنامه

اینجانب حبیبه حکیمی دانشجوی دوره کارشناسی ارشد رشته آموزش زبان انگلیسی دانشگاه صنعتی شاهرود نویسنده پایان نامه بررسی اعتباراداخلی، اعتبار خارجی و اعتبار سازه در تحقیقات کمی در حوزه آموزش زبان انگلیسی تحت راهنمایی دکتر سید علی استوار نامقی متعهد می‌شوم:

- تحقیقات در این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های محققین دیگر به مرجع مورد استفاده استناد شده است.
- مطالب مندرج در این پایان نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود است و مقالات مستخرج با نام «دانشگاه صنعتی شاهرود» و یا «Shahrood University of Technology» به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی رساله تأثیرگذار بوده‌اند در مقالات مستخرج شده از رساله رعایت می‌گردد.
- در کلیه مراحل انجام این رساله، در مواردی که از موجود زنده (یا بافت‌های آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است.
- در کلیه مراحل انجام این رساله، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری، ضوابط و اصول اخلاق انسانی رعایت شده است.

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته شده است) متعلق به دانشگاه صنعتی شاهرود است. این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در این رساله بدون ذکر مرجع مجاز نمی‌باشد.

Abstract

The validity of research findings has always been a debatable issue in educational research, in general, and ELT research, in particular. Such an important issue has been more controversial especially when it came to experimental or quasi-experimental design in ELT research. Bearing this in mind, the researcher set to investigate the construct, external and internal validity criteria as defined in the primary sources of research in 40 papers published in Iranian (n = 20) and non-Iranian (n = 20) journals in ELT domain. The sample was randomly chosen from the body of papers including experimental or quasi-experimental design published from 2010 to 2020. Adopting an exploratory approach to the sample, they were meticulously analyzed in terms of their observing the threats to construct, external and internal validity. Having extracted the frequencies of the criteria, chi-square goodness of fit test and chi-square independence test were used to analyze the data. The results showed that the papers were generally deviant from the ideal state of observing construct, external and internal validity threats. In addition, a further comparison of the papers published in all journals indicated that among the measured internal validity threats, instrumentation was the less frequent threat followed by selection bias, history, maturation, diffusion and testing effect and selection-maturation interaction, regression, mortality, experimenter-subject effect were the less important ones, and the papers generally failed to meet the criteria, namely, setting-treatment interaction, pretest-treatment interaction and then selection-treatment interaction and subject-experimenter effect in the sample papers as the threats to external validity. Implications of the study for academicians were also discussed in accordance with the findings of the study.

Keywords

Construct validity, ELT, external validity, internal validity

TABLE OF CONTENTS

List of Tables

List of Figure

CHAPTER ONE:

INTRODUCTION

1.1.Overview

Papers in the field of English language teaching (ELT) are expected to entail a reliable means of scholarly exchange of academic information and innovation. Accordingly, a published research report which has been systematically reviewed and evaluated by a journal board of referees is generally incorporated into the plateau of scientific knowledge based on which further research and argumentation is being conducted in addition to replication and meta-analysis. Considering the fact that scholarly knowledge not only cumulative but also cooperative, a researcher certainly wants to predicate his/her study on a reliable unreliable study, so that its reliability is of importance. Besides, this issue is crucial when a journal is evaluated. Moreover, the decisions made based on the findings reported in the papers of ELT domain influence the academic lives of many learners. Therefore, it is strongly justified ELT papers to be evaluated. This is reflected in Donovan's (2007) statement peer-reviewed journal papers receive the most attention from promotion panels and research committees.

If we have a short look at the validity history we find that during years classifications of Threats to validity have been changed. After some years in 1979 Cook and Campbell recorded four types of validity threats in quantitative experimental analysis: internal validity, construct validity of putative causes and effects, and external validity and statistical conclusion validity. Then, Concerning qualitative research, (Maxwell. 1992) provided a general categorization of threats that can be mapped to Cook and Campbell's categorization as follows: interpretive validity (statistical conclusion validity), theoretical validity (construct validity), and generalizability (internal, external validity). In this regard, the present study aims to investigate the internal, external validity, and validity of the measure of 40 experimental ELT research and then, their frequency by Chi-square statistical method.

1.2.Statement of the Problems

Considering the significance of the quality of the papers in the domain of ELT, several instances of research can be found concentrating on reviewing and evaluating the papers published in the academic journals of the ELT domain emphasizing the soundness of different sections of the papers. For example, Koopman (1997) proposed a checklist for evaluating abstract and highlighted

five essential sections, namely (1) motivation, (2) problem statement, (3) approach, (4) results, and (5) conclusions. Durant (1994) also evaluated the method sections in the papers and maintained that "a substantial proportion of articles contain sufficient statistical and/or methodological mistakes to cast doubts on the stated conclusions" (p. 4). However, his checklist may not apply in the ELT domain since it focuses on the method section of the papers only. Similarly, Barbour (2001) concentrated on rigor in research papers, through investigating five technical fixes i.e. grounded theory, multiple coding, purposive sampling, and respondent validation as well as triangulation. Besides, he warned against the use of prescriptive checklists in evaluating qualitative research. However, Barbour (2001) did not discuss rigor in qualitative research, and the points he proposed are too broad and are restricted to the method section only.

In the same vein, Bornhöft *et al.* (2006) studied the external validity of papers, which is ironically neglected in checklists used for research paper evaluation. Accordingly, they addressed internal validity, external validity, and modal validity of research papers. However, its focus on clinical purposes can hardly be used for ELT research papers. Letts *et al.* (2007) introduced some guidelines for a critical review of a qualitative study consisting of some factors such as citation, purpose, literature review, design, qualitative methods, sampling, data collection and analysis, rigor, and conclusion. In line with these components, the researchers have offered some procedures to help readers critically appraise qualitative research studies. As claimed by the authors, these guidelines are appropriate for evaluating multidisciplinary papers; thus, they may be too general for ELT paper evaluation. As Crack, Gieves, and Lown (2011) maintained such guidelines are suitable for authors to complete before submitting their papers. However, such a checklist may not warranty the high quality of contents of research papers (Lovejoy, Revenson, & France, 2011). Besides empirical attempts, some other works such as Derntl (2014) and Henningsen (2015), have only offered some guidelines about paper preparation, resubmissions, and replying to the editors or reviewers.

Bearing these facts in mind, the researcher came up with the existing gap in the body of research on ELT papers published in Iranian and non-Iranian journals concerning the lack of evidence on the validity threats which had been controlled. Our study aimed at exploring the (UN)

controlled validity threats in the papers published in the ELT domain in Iranian and non-Iranian journals.

1.3. Purpose of the Study

The present study aims at investigating the internal and external validity and validity of the measure of experimental findings in ELT papers; essentially, this study aims at exploring the existing threats in Iranian and non-Iranian papers published in the ELT domain and comparing the extent to which these threats were controlled in these journals.

Research Questions:

Based on the problems stated above and considering the purposes of the study, the following research questions were raised:

1. Is there any significant difference in frequency between studies that assured internal validity through the random assignment and what was ideally expected to be observed in these papers?
2. Is there any significant difference in frequency between studies that assured external validity through the random selection and what was ideally expected to be observed in these papers?
3. Is there any significant difference in frequency between studies that assured construct validity through adequate specification of construct and what was ideally expected to be observed in these papers?
4. Is there any significant difference in frequency between studies that assured internal, external validity and valid measure between Iranian and non-Iranian papers?
5. Is there any significant difference in frequency between studies that addressed the threat to internal validity and what was ideally expected to be observed in these papers?
6. Is there any significant difference in frequency between studies that addressed the threat to external validity and what was ideally expected to be observed in these papers?

Research Hypotheses:

The following null hypotheses were formulated based on the research questions of the study:

1. There is not any significant difference in frequency between studies that assured internal validity through the random assignment and what was ideally expected to be observed.
2. There is not any significant difference in frequency between studies that assured external validity through random selection and what was ideally expected to be observed.
3. There is not any significant difference in frequency between studies that assured construct validity through adequate specification of construct and what was ideally expected to be observed.
4. There is not any significant difference in frequency between Iranian and non-Iranian papers in terms of internal, external and construct validity.
5. There is not any significant difference in frequency between studies that addressed the threat to internal validity and what was ideally expected to be observed.
6. There is not any significant difference in frequency between studies that addressed the threat to external validity and what was ideally expected to be observed.

1.4. Significance of the Study

This study will be useful for researchers, university teachers and policymakers. Helping investigators to be more reflective at every stage of the research process. They have to notice the truth that controlling the confined variable that is naturally kind of dynamic systems and is out of control is impossible .so, it would be a significant shortcoming of experimental researches. And, this paper makes it clear that every research contains multiple threats to internal, external validity, and validity of the measure, and that teachers should exercise extreme caution when making conclusions based on one or a few studies. Additionally, the study highlights the importance of assessing sources of invalidity in every research study and at different stages of the research process. For example, just because threats to internal and external validity have been minimized at one phase of the research study does not mean that sources of invalidity do not prevail at the other stages. And at the end, it is believed that studies like this would help, make a more accurate evaluation of ELT paper quality through varied threats. and having a better

understanding of the current status of different types of validity in the published journal papers in the ELT domain would not only contribute to the transparency and quality of research papers that would be published in the future but help journal editors and referees make more informed decisions in the domain of ELT. Regarding to the policymakers, it helps them make informed decisions based on the accurate exploring of forty authentic papers through validity criteria, rather than a general frame of one qualitative or quantitative paper.

1.5. Definition of key Terms

Validity: Validity can be formally defined as "whether the means of measurement are accurate and whether they are measuring what they are intended to measure" (Winter 2000, p. 3).

Internal validity: An experiment has internal validity (and would meet the methodological threshold standard for publishing ability) only to the extent that it has been carefully and systematically designed with sufficient care to give a high degree of confidence that the results reported accurately measure the nature and the effect size of the intervention tested. Wilson (2016).in our study we explore just the most important aspect of internal validity named random assignment.

External validity: External validity, or generalizability, is the degree to which a finding (i.e. an estimated causal effect of treatment on the outcome) can be applied to a target population (Steckler & McLeroy, 2008). In our study, we just explore the most important aspect of external validity named random selection.

Construct validity: The validity of inferences about psychological constructs involved in the subjects, setting, treatments, and observations used in the experiment. Ary *et al*. (2018). In our study, we just explore the most important aspect of the construct validity named valid measure test.

Confined variable: Such a variable in experimental research can influence the dependent variable and make our research result unrealistic.

Threats: Threats are extraneous variables, we call them threats because unless they are controlled they may produce an effect that could be mistaken for the effect of the experimental treatment.

1.6. Limitations of the Study

Like any studies, this study has its limitation. Regarding the limitation of some online databases, the selected studies were collected from free open-access databases and some rich studies are not included in the study since they were not accessible. The other limitation is that the generalization of the result of the current study will be restricted because the study was conducted in the summer of 2020.and we investigate some ELT research just during recently ten years, so the study findings were limited to that time.

1.7. Delimitation of the Study

To limit the problems and clear the exact border of the research, we define some characteristics of experimental research to be chosen. That is, studies which did not meet these criteria were excluded. They had the following characterizations:

- They have to be written in English
- They have to be published between the years 2010 and 2020.
- They have to include sufficient information for estimating their internal, external, and construct validity.
- They have to be just in the ELT domain.

CHAPTER TWO:

LITERATURE REVIEW

2.1. Overview

This chapter presents the relevant literature of the study into two main sections of theoretical perspectives and empirical findings.in the first section theories underlying validity of measurement and then of research design consist of internal, external and construct validity will be discussed. The next section deal with empirical findings from the areas relevant to the present study.

2.2. Theoretical Perspectives

Validity is not a single, universal concept, but rather a dependent construct, obviously grounded in the processes and intentions of particular research methodologies and projects. The exact nature of 'validity' is a highly debated topic in both educational and social research since there exist no single or common definition of the term. Validity definitions are concerned; two common strands begin to emerge: Firstly, whether the means of measurement are accurate. Secondly, whether they are actually measuring what they are intended to measure? Winter (2000).It is fluid, changing with the times and values. “Thus, validity and values are one imperative, not two, and test [measure] validation implicates both the science and the ethics of assessment, which is why validity has force as a social value” (Messick, 1995, p. 749).Therefore, in order to understand something of the range of meanings attached to 'validity', it is essential to review validity of measurement and then validity of research design during times.

Decades ago, a typical definition of validity could be that a test is valid “for anything with which it correlates” (e.g., Guilford, 1946, p.429) or “if it measures what it purports to measure” (Shepard, 1993, p.410). As stated by Cronbach (1971), “One validates, not a test, but an interpretation of data arising from a specified procedure” (p. 447).

Its theoretical concept has developed considerably over time (e.g., American Psychological Association, 1954, 1966; American Psychological Association, American Educational Research Association and National Council on Measurement in Education, 1974, 1985, 1999).When validity standards were first organized in 1954 (American Psycho- logical Association, 1954), regarding to the aims of test, four types of validity were recognized. Content validity was obligatory for tests describing an individual's performance on a defined universe of tasks. Predictive validity was

called for when a test was used to predict future performance .Concurrent validity, the other type of validity concerning an external criterion, was immediate when a new test was proposed as a substitute for a less convenient measure that was already accepted. Construct validity was needed when making inferences about unobserved traits such as intelligence. In a subsequent revision of the standards (American Psychological Association, 1966), the two types of validity apply to an outside criterion were abridged to one category, criterion- related validity.

A significant change in validity theory was Cronbach's (1971) statement that: "one does not validate a test, but an interpretation of data arising from a specific procedure" (447). This meant that the importance of the validity investigation shifted from the specific and definite instrument to the interpretations of the measurement result. This view is still worthy, and most researchers agree that it is not the test itself but how its outcome is interpreted and used that should be the focus in a validation process.

2.2.1. Theory of Messick (Unitary Validity Framework)

Over the years, the different types of validity have still been useful, without any clear distinctions between them. In modern validity theory, they are often referred to as a unitary validity framework. In this framework, construct validity has a central position by approval almost all forms of validity evidence. The unitary concept of validity was gradually developed, and eventually described by Messick (1989), who established the agreement that construct validity is the unifying concept of validity. Messick (1989, p.13) also describes validity as an, "integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores". social consequences of measurement outcomes was substantial part in the unitary concept of validity theory .In Messick's well-known model (see Figure 1 below) he identified a number of validity aspects, and decreases them to a system. The model emphasized on test score, grade, or other form of assessment outcome. It clarified how the outcome is interpreted and used, and what type of evidence and consequences will be the result, both in terms of value implications and other types of consequences. Then Messick makes a clear distinction between the aspects of the model, he emphasizes that they are tangled and regarding to validating an instrument all these aspects should be into consideration. Although Social consequences had previously been somehow important, but

not necessarily a matter of validity, but in his model social consequences due to construct-irrelevant variance and construct under-representation were introduced as an important aspect of the unitary concept.

Between those who advocate an extensive validity perspective, this model of validity became the most established and well know. Although he was not the only one to make a wider framework (for instance, Cronbach 1988; Crooks, Kane, and Cohen 1996; Kane 1992), but his unitary concept of validity was accepted by lots of scholars.

	TEST INTERPRETATION	TEST USE
EVIDENTIAL BASIS	Construct validity	Construct validity +Relevance/utility
CONSEQUENTIAL BASIS	Value implications	Social consequences

Figure 1: Messick's Facets of Validity Framework

Since border concept of validity was included by consequence, it provided some problematic issues for the practitioner. They were uncertain about two aspects .first about the feasibility of such an approach and secondly about whether social consequences should be part of the validity framework at all. some (e.g., Messick 1989; Shepard 1993; Stobart 2001 and Moss 1998) agreed and argued that investigating the social consequences of measurement is reasonable and even necessary because consequences are directly related to the purpose of a measurement procedure.

On the other side, Popham (1997), for instance, argued that including social consequences lead to confusion, not simplicity. Mehrens (1997) articulated a similar view and suggested that in this way the validity term would be expanded instead of shortened. They claimed that they are aware of importance of consequences related to measurement, but not in this way. These views are supported in recent articles, too. Borsboom, Mellenberg, and van Herden (2004, P.1061) conclude

that validity has become so complex that “the theoretically oriented are likely to get lost in the, intricate subtleties of validity theory, whereas the practically oriented are unlikely to derive a workable conceptual scheme with practical implications from it”. Nevertheless, the broadened concept of validity has also been commended for increasing the quality of validity investigations. Messick’s model has been argued to be useful for organizing and deriving questions related to validity (Gersten and Baker 2002; Törnkvist and Henriksson 2006; Wikström 2006; Wolming1998, 1999).

2.2.2. Theory of Kane (Argument–based Approach)

Kane (1992) has prolonged Cronbach's (1988) recommendation. He conceptualizes validation as the evaluation of interpretive argument .he continued in order to validate a test-score interpretation we should do as following: 1- first of all, based on test scores decides on the statements and decisions 2-then identifies the assumption and inferences leading from the test scores to these statements and conclusions 3-recognizes potential opposing interpretations, and at the end 4-search for proof supporting the inferences and assumptions in the proposed explanatory argument and disproving potential counterarguments. (Kane, 1992).

In 2006, the fourth edition of Educational Measurement was published, which also meant a new division on validity, again by Michael Kane. He was agree with Messick’s chapter, but highlights practical feasibility through an ‘argumentative approach’ to validity. In this approach, two types of arguments are in focus, first the interpretive argument and then the validity argument. The interpretive argument should be a specification of the planned interpretations and uses of the results of the test, from theory to conclusions and decisions. Kane emphasizes that the clarity of the interpretive argument of the measurement process is related to the degree of validity of the instrument and whether the validity argument, which is the evaluation of this interpretive argument, appears that the interpretive argument is coherent and reasonable and so the assumptions are plausible. The argumentative approach and modern view on validity are in the same direction, they emphasizes that validity has to do with how a test is used (and the interpretations made), however it is presented from a somewhat different viewpoint.

Compared to Messick's definition of validity Kane didn't consider consequences and especially social consequences important aspect. It was the main difference of this two. According to Kane (2006), the main benefit of the argumentative approach is that it provides guidance not theoretically, practical guidance in how to assign research efforts and in assessing process in the validation effort.

2.2.3. Types of Validity

2.2.3.1. Construct validity

Construct validity involves its historical conduct because the theory appealed today by the same name is noticeably different from the 1954 version. The early version of construct validity was both too modest and too ambitious compared with our present-day understandings. The 1954 standards and Cronbach and Meehl (1955) introduced construct validation as the weak parallel to the existing view of validity, as an indirect method of validation, so they considered it as a substitute when a real criterion or content domain was not available. "Construct validity is ordinarily studied when the tester has no definitive criterion measure of the quality with which he is concerned and must use indirect measures to validate the theory" (American Psychological Association, 1954, p. 14).

The most important results from these developments are (1) the presentation in the current Standards for Educational and Psychological Testing (American Education Research Association/American Psychological Association/National Council on Measurement in Education, 1999), in which validity is conceptualized differently than in the previous versions, it is unified and (2) an extensive integration of several aspects of validity into a comprehensive framework (Messick, 1989, 1995). According to both developments, no longer may validity be considered to consist of separate types, as emphasized in the last version of the Standards for Educational and Psychological Tests. The separate types of validity (construct validity, criterion-related validity, and content validity) were differentially suitable, depending on test use. Instead, the concept of construct validity is now expressed within a unifying framework of construct validity. Cronbach and Meehl (1955) that were members of the Technical Recommendations Committee, published their paper, 'Construct validity in psychological tests, which identified validation measures to

acquire evidence relevant to construct validity. This evidence comprised various aspects of criterion validity and content validity, so they didn't mention construct validity as a new type of validity, it came to be seen as the unifying concept of validity and scholars shifted from the validity of the test to the validity of test score interpretations. The ultimate concept of validity refers to whether a test or a measurement instrument, measures what it purports to measure. As below you can find some scholars point of view about it:

- The degree to which the test truly measures what it rationales to measure (Anastasia, 1954)
- "Construct validity used to refer to the vertical correspondence between a construct which is at an unobservable, conceptual level and a purported measure of it which is at an operational level." (Peter, 1981, p. 134).

Two traditional types of validity content validity and criterion-related validity are conceptualized as different sources of evidence for construct validity by Messick (1989). Put more simply, construct validity is a measure of whether researchers are studying the constructs they believe to be studying (Churchill, 1979; Cronbach & Meehl, 1955; Heeler & Ray, 1972; Peter, 1981). Construct validity is defined as the ability to make evidence-supported inferences from sampled indicators to the constructs they are intended to represent (Shadish, Cook, & Campbell, 2002). Shadish *et al*. (2002) believed to construct validity of experiments. Construct validity of experiments is defined as the validity of the inferences made about a construct based on the measures, treatment, subjects, and settings used in an experimental study.

Construct validity is also concerned with the subjects and setting of the experiment. Regarding this, Ary *et al*. (2018) made some notes. Imagine a research study using depressed children. How does the investigator describe the term depressed? About "depressed" the definition might differ greatly depending on whether the diagnosis was made by a competent psychologist or by a counselor using only scores from a personality inventory. They then continue about the concern of the construct representation of the settings for experimental research. According to their saying, comparative look makes it clear that the construct representations of treatments and subjects receive more attention than the setting, except for research dealing with the effect of environment and culture.

In construct validity theory, the construct (e.g. intelligence, empathy, etc.) is a hypothesized or theoretical concept that is defined by its position in a network of other constructs. The relationships among the constructs in the network are defined by scientific laws that bond the constructs and form the network. Cronbach and Meehl referred to this as a 'nomological network', which is a network of laws that relates constructs: scientific theory. Cronbach and Meehl (1955). Construct validity, then, is established by any evidence that supports the nomological network of constructs and laws that holds the construct.

2.2.3.2. Threats to Construct Validity

The threats to construct validity concern how well the study's operations match the constructs used to describe those operations. As below you can find threats to construct validity announced by Array *et al*. (2018):

1. Measure of the construct: As a result of poor operational definition, the construct was not accurately measured.
2. Manipulation of the construct: The construct should be manipulated accurately in the study; defective manipulation causes incorrect and improper inferences.
3. Reactivity to the experimental situation: Subjects' perceptions of the experimental situation can impact and become a portion of the treatment construct being tested.
4. Experimenter effect: Sometimes the explorer transport anticipations about desired responses, as a result, those expectations converted part of the treatment construct being studied.

Shadish *et al.* (2002) suggested the following steps to increase construct validity of experiments:

1. Clear explanation of the persons, setting, treatment, and outcome constructs of interests very important and should be mentioned at first.
2. Carefully select occasions that match those constructs.
3. To avoid any slippage between instances and constructs, assess the match between these two.
4. Revise construct descriptions accordingly.

2.2.3.3. Concepts of Internal and External Validity

Since Donald Campbell had a vital role in the expansion of validity definition, and their classification of validity types has evolved, so we mentioned his ideas regarding internal and external validity and their threats during years as below:

In 1957 Campbell introduced the concepts ‘internal’ and ‘external’ validity as a primary to detailing the various threats to validity that face experimental and quasi-experimental research. He has got two definitions of internal and external validity. In his original article, Campbell describes internal validity as a basic minimum and as being concerned with whether the experimental stimulus made “some significant difference in this specific instance” (Campbell, 1957, p. 297). Subsequently, he and others have developed this scheme (Campbell & Fiske, 1959; Campbell and Stanley 1963; Cook and Campbell 1979); and it has come to be widely accepted among social researchers. Following you can find their definition of validity:

Discriminate validity ensures that the test is not related to other instruments excessively (Campbell & Fiske, 1959) and the validity of generalizability indicates how appropriate the test is to test-takers in a variety of settings. In the later and better-known treatment (Campbell & Stanley, 1963, p. 5) internal validity is defined similarly but a little more fully as “a basic minimum without which any experiment is uninterruptable”. They identified the following eight threats to internal validity: history, maturation, testing, and instrumentation, and statistical regression, differential selection of participants, mortality, and interaction effects (e.g., selection-maturation interaction) (Gay & Airasian, 2000).

Their definition of external validity was changed over the years. In his 1957 section of writing Campbell defined external validity as “representativeness or generalizability, to what populations, settings, and variables can this effect be generalized?” (p. 297). This construction is recurring in Campbell and Stanley (1963), but once again Cook and Campbell (1979) provide a somewhat different account. According to their saying internal validity is “the validity with which statements can be made about whether there is a causal relationship from one variable to another

in the form in which the variables were manipulated or measured". (p. 38). And "external validity is the approximate validity with which we can infer that the presumed causal relationship can be generalized to and across alternate measurements of the cause and effect and across different types of persons, settings and times". (p. 37). They stated that external validity involves generalizing (1) to particular target persons, settings, and times and (2) across types of persons, settings, and times. They elaborated this previous classification to four types of validity: internal validity, external validity, construct validity, and statistical conclusion validity.

Regarding to the notion of validity there is a crucial question, does 'Validity' Concern the whole process of research, or certain key stages? Winter (2000) believed that there is no solitary form or concept that can universally be demanded to define validity terms. Neither, however, can validity be said to be an individually identifiable element of any research project, which is skilled of being located at various and specific stages within the research. He also ascertained that "The concept of 'validity' defines extrapolation from, or categorization within, any research project. For some researchers (mainly qualitative), 'validity' is not a singular acid test that can be applied to the research process as a whole. The 'validity' measure can be applied differently depending upon the researcher's beliefs as to what stage of the research process requires validation. Such an approach may perceive validity as referring only to measurement, observers, scores, instruments, relationships between scores, or observable variations, rather than to the whole research process. Within this approach, 'validity' is claimed either by viewing it as a resident in a particular stage of the research process or as combinations of certain stages." (Winter, 2000, p.4). Contrary, Maxwell (1992, p. 285), was one of the scholars who considered validity for the various stage of research. He identifies five typologies of 'validity' as "1- Descriptive Validity 2-interpretive validity3-theoretical validity 4-generalizability 5-evaluative validity". Winter (2000) disagreed with Maxwell's classification and mentioned it as unnecessary and paradoxical. He believed that dividing the concept of validity and then allocate them to stages of the research process is a needless conceptualization and such an approach couldn't guarantee validity, and continued Maxwell's typologies, although being systematic, are contradictory and pointless.

2.2.3.4. Threats to Internal Validity

Campbell and Stanley (1963) identified eight extraneous variables that frequently Signify threats to the internal validity of the research. These variables are named threats because except they are controlled, they may yield an effect that could be mistaken for the effect of the experimental treatment. If uncontrolled, these extraneous variables raise doubts about the accuracy of the experiment because they permit an alternative explanation of the experimental findings. Thus, to have a better grasp of the range of definitions attached to the term validity threats, it is necessary to review a number of definitions offered by leading authors. (McMillan, 2007; Yu & Ohlund, 2010; Ary *et al.,* 2018).

1. **History:** Yu and Ohlund (2010) stated the particular events which ascend between the first and second measurement. Ary *et al.* (2018) believed specific events or situations, that aren't related to experimental treatment, may occur between the beginning of the treatment and the posttest measurement and may yield changes in the dependent variable. History happens at the same time that the experimental treatment is being applied, and doesn't refer to past occasions. These may be major political, economic, or cultural events or some rather minor troublesome factors that occur during the conduct of the experiment. So, the longer the period between the pre-and post-measurements on the subjects will lead to greater the history threat.

2. **Maturation:** Yu and Ohlund (2010) believed, its function of the passage of time. e. g. if the plan lasts a long period, most participants may advance their performance regardless of treatment. According to the saying of Ary *et al.* (2018) the term maturation refers to changes (biological or psychological. That may occur within the subjects simply as a function of the passage of time. Since these changes produce effects that could mistakenly be attributed to the experimental treatment so they consider as threats to internal validity. Topics may implement differently on the dependent variable measure simply because they are older, wiser, hungrier, more fatigued, or less motivated than they were at the time of the first measurements.

3. **Testing:** Yu and Ohlund (2010) believed sometimes, the pretest becomes a form of "treatment." This means the effects of taking a test on the outcomes of taking a second test. Ary *et al.* (2018) believed regardless of any treatment, taking a test once may affect the subjects' performance when the test is taken again, this is named the testing effect. So, In designs using a pretest has got some disadvantages as following, subjects may do better on the posttest because they have learned subject matter from a pretest, have become familiar with the format of the test and the testing environment, have developed a strategy for doing well on the test, or are less anxious about the test the second time.

4. **Instrumentation:** McMillan (2007) believed some weaknesses can't control by random assignments, like how data are collected. The question is that, whether data gathering in a different way impacts the results in either the experimental and control groups. This could occur with the observer, rater, or recorder error or bias, ceiling and floor effects, and in changing measures with single group longitudinal studies. Yu and Ohlund (2010) believed all changes which may yield changes in outcomes like, in the instrument, observers, or scorer. According to the Ary *et al. (2018)* the best recommendation is to avoid any changes in the measuring instruments during a study. Changes may involve the type of measuring instrument, the difficulty level, the scorers, the way the tests are administered, using different observers for pre-and post-measures, and so on.

5. **Statistical regression**: Yu and Ohlund (2010) believed this phenomenon was first exposed by British statistician Francis Galton in the 19th century. And is also known as regression towards the mean. In research design, the threat of regression towards the mean is caused by the selection of subjects based on extreme scores or characteristics. If there are fifty poor students in the treatment, they will likely show some enhancement after the treatment. However, if the students are extremely poor and thus are unfeeling to any treatment, then it is called the floor effect. Ary *et al.*(2018) believed the term statistical regression refers to the tendency for extremely high or extremely low subjects who score on a pretest to score closer to the mean (regression toward the mean) on a posttest. Statistical regression

is a threat to internal validity when a Subgroup is selected from a larger group based on the subgroup's extreme scores (high or low) on a measure.

6. **Selection bias**: Yu and Ohlund (2010) believed the biases which may consequence in the selection of comparison groups. Randomization (Random assignment) of group membership is a counter-attack against this threat. According to the saying of Ary *et al*. (2018) selection is a threat when differences between the experimental and control groups exist even before the experiment begins. If the groups are not equivalent before the study, we cannot know whether any difference observed later is due to the treatment or the pretreatment difference.

7. **Selection-maturation interaction**: Yu and Ohlund (2010) stated the selection of comparison groups and maturation interacting which may lead to confounding outcomes, and erroneous interpretation that the treatment caused the effect. Ary *et al.* (2018) believed selection and maturation may interact in such a way that the combination results in an effect on the dependent variable that is mistakenly attributed to the effect of the experimental treatment. Such interaction may happen in a quasi-experimental design in which we don't use random selection for both experimental and control groups, but instead are preexisting intact groups, such as classrooms. Although a pretest may indicate that the groups are equivalent at the beginning of the experiment, the experimental group may have a higher rate of maturation than the control group, and the increased rate of maturation accounts for the observed effect. If more rapidly maturing students are "selected" into the experimental group, the selection–maturation interaction may be mistaken for the effect of the experimental variable.

8. **Experimental mortality (attrition)**: McMilan (2007) stated when after random assignment subjects in the intervention group dropped out of a study at rates that are different from subjects in a control or comparison group, it is likely that such treatment-correlated attrition will be confounded in unknown ways. According to the saying of Yu and Ohlund (2010) it is the same as the loss of subjects. Ary *et al*. (2018) stated that this threat occurs when there is a discrepancy loss of participants from the comparison groups.

Even in the absence of treatment, this differential loss may result in differences in the consequence measure.

9. **Experimenter effect**: McMillan (2007) stated that most of the time researchers try to prove a point rather than taking the viewpoint of an unbiased, objective investigator. This leads to experimenter effects that are both deliberate (bias) and unintentional. On some occasions, there is a need to show that specific interventions have a positive impact on student learning. So researchers have a vested interest in the study there is motivation to find results that will enhance their position. And regarding his attitudes, values, and biases, and needs that may foul the study. Since experimenters typically have research hypotheses about the outcomes, it is important to include procedures that minimize the plausibility that these effects could constitute rival hypotheses. Yu and Ohlund (2010) believed it happens when the experimenter unintentionally puts greater expectations or extra care to the participants. Ary *et al.* (2018) stated that the experimenter effect refers to unintentional effects that the researcher has on the study. Personal characteristics of the researcher, such as gender, race, age, and position, can affect the performance of subjects. Sometimes the actual operation of the experiment unintentionally gives the experimental group an accidental advantage over the control group. For example, in an experiment comparing the effectiveness of two teaching methods, the more professional teacher may be assigned to the experimental group. Internal validity is threatened if the experimenter has a personal bias toward one method over another. These preferences and expectancies on the part of the experimenter may be unconsciously conveyed to subjects in such a way that their behavior is affected.

10. **Diffusion of Intervention (Treatment):** McMilan (2007) stated that an important principle of good randomized studies is that the intervention and control groups are completely independent, without any effect on each other. This condition is often problematic in a field study. When the effect of the intervention spreads to the control group, or when the control group knows about the intervention, behavior and responses can be initiated that otherwise would not have occurred. Sometimes subjects affected by an intervention interact with control subjects because they are nearby, such as friends in

treatment and control groups, by being in the same school or neighborhood, or by being in the same class. In this circumstance, the changes caused by the intervention are diffused to the control subjects through their interaction with each other. Ary *et al.* (2018) believed that when participants in one group (naturally the experimental group) start to communicate information about the treatment to subjects in the control group and as a result influence the latter's behavior on the dependent variable. Also, it can happen for teachers of both groups.it means one of them in one group share information about methods and materials with teachers of the opposite group.

11. **Experimenter and *Subject effects*:** Mc Milan (2007) stated that some threats occur quite independently from whether or not there is random assignment of units, and are especially troublesome when it is clear to subjects that what outcome is expected. These behaviors can be attributed to the subjects because of the sampling and procedures used in an experiment. These are consist of compensatory rivalry and equalization, resentful demoralization, Hawthorne effect, demand characteristics, social desirability, and subjects wanting to please the experimenter. Most of these factors insipid the effect of the treatment and make it more difficult to show differences. Others, like resentful demoralization, expand true differences. Yu and Ohlund (2010) believed sometimes participants change their behaviors when they are aware of their role as research subjects, like John Henry and The Hawthorne effect. Ary *et al.* (2018) stated subjects' attitudes that can be a threat to internal validity are developed in response to the research situation and called subject effects.

2.2.3.5. Minimize the Effect of Threats on Internal Validity

The most important aspect of experiment design is to avoid or at least minimize the effect of Threats to internal validity. Since all the time there are some preexisting subject differences in the comparison group, so, the researcher's first obligation must be toward controlling for any relevant preexisting differences between subjects in the comparison groups. In all experimental researches, One of the most important and harmful threats to internal validity is selection. It is a result of a failure in the random assignment of participants. And the problem is that in the research studies

field most of the time it is not easy to randomly assign the participant to a comparison group. There is 2 attractive viewpoints by (Mc Milan, 2007; Ary *et al*., 2018) about the random assignment and how it leads to controlling threats to internal validity.

Mc Milan (2007) designed a study about the importance of randomization in experimental research and how does it increase the internal validity of the research. He found that randomization of a study doesn't guarantee threats to internal validity. It just makes sure the researcher about controlling some threats depending on the nature of the experiment, still, many possible threats remain. So, experiments, whether randomized or not, have a crowd of possible threats to internal validity. However, according to the saying of Mc Milan (2007), this emphasis on doing randomized experiments may be misleading unless attention is paid to three important conditions. The first is being sure that the design completes the reason for using random assignment – to achieve statistical equivalence of the experimental and control group before, during, and after the intervention is executed. The second is the need to evaluate internal validity according to threat factors that are common in field studies. Third, determining causality, which means why experiments are conducted, is heavily dependent on contextual factors peculiar to each study. It is a compliment to Don Campbell and Julian Stanley that their seminal publication Experimental and Quasi-Experimental Designs (1963) has had such staying power. In particular, their eight internal threats to validity, along with their labels, continue to be the ones still emphasized in educational research textbooks (some now list a few more, such as experimenter effect or diffusion of treatment). In approving his opinion, I found a state by Chatterji, he believed that at best, RFTs control many possible threats, but not all. At worst, relying too much on RFTs without appropriate consideration of all possible threats to internal validity will result in misleading conclusions about program effectiveness (Chatterji, 2007). So, it is crystal clear that the responsibility of researchers is to recognize possible threats and then include design features that will gather information to reduce the probability that the threat is reasonable.

The second investigation in this field was done by Ary *et al.* (2018). They have some statements about dealing with threats to internal validity. According to their saying, intersubject differences have a crucial role in controlling internal validity, so researchers have to be controlled by 6 below procedures: (1)Random assignment (2) randomized matching, (3) homogeneous

selection, (4) building variables into the design, (5) statistical control (analysis of covariance), and (6) use of subjects as their controls.

Random assignment: The researcher's first attempt for the study is assigning subjects to groups in a system that functions independently of personal judgment and the characteristics of the subjects. Ary *et al*. (2018) Believed that Random assignment is a powerful instrument to control selection bias, because defines whether subjects are placed in the experimental or the control group, by only chance and in this way all members has an equivalent chance of being assigned of the groups. One important assertion by them was the difference between random assignment and random selection .he believed that "random assignment is not the same thing as random selection. Random selection is the use of a chance procedure to select a sample from a population. Random assignment is the use of a chance procedure to assign subjects to treatments."(p.285).

Randomized matching: This sort of matching select when random assignment is not possible. It means researchers sometimes select pairs of individuals with equal or almost equal characteristics and randomly assign one member of the matched pair to treatment A and the other to treatment B. important step is that at first the subjects should be matched on relevant variables and then randomly assigned to treatments. So, the researcher first decides what variables are suitable for the matching process. (IQ, mental age, socioeconomic status, age, gender, reading, pretest score, or other variables).

Homogeneous selection: Here researcher, Select samples that are as homogeneous as possible on extraneous variable .in this way make groups practically comparable on that extraneous variable. If the experimenter doubts age as a variable that could disturb the dependent variable, he would select only children of a particular age. Although this method is effective and operative as a good controller but has a disadvantage.it will decrease the generalizability of populations.

Building variables into the design: In this method, we have two independent variables. For example, if you want to control gender in an experiment and you can't use the homogeneous

selection, you could add gender as another independent variable. You would comprise both males and females in the study. As result to see the effect of both gender and independent variable we can use analysis of variance.

Statistical control: The other way to control for the effect of an extraneous variable known to be correlated with the dependent variable is the Analysis of covariance (ANCOVA) as a Statistical technique. The variable used in ANCOVA to regulate score is named the covariate.

Using subjects as their controls: through this technique, subjects would be assigned to all experimental conditions. First under one experimental treatment and then under another. This process of control is effective when practicable, but in some conditions, it cannot be used. You cannot, for example, teach children how to division elements one way and then try to remove their memory and teach it another way. As a result regarding several ways suggested by Ary *et al.* (2018) to be sure about omitting intersubject differences between groups, the researcher must choose the best way to have confidence about comparison groups.

2.2.3.6 Threats to External Validity

Researchers want the results of a study to provide information about a larger Dominion of subjects, conditions, and operations that were investigated. To make external validity of experimental research and generalizations from the observed to the unobserved, researchers need to assess how well the sample of events studied represents the larger population to which results are to be generalized. In the end, if inferences about a causal relationship hold over changes in subjects, settings, and treatments, the experiment has external validity.

The research design has vital role in controlling threats to internal validity, but controlling the threats to external validity is not clear and simple. You need to examine carefully and logically the similarities and differences between the experimental setting and the target setting concerning subjects and treatments, before you can assume external validity, As below you can find two

summarized opinions about threats to external validity by (Yu & Ohlund, 2010; Ary *et al*., 2018). Threats to external validities enumerated by Ary *et al.* (2018) are as follows:

1. Selection–treatment interaction: Investigator should use a big, random sample of participants because a result found with certain subjects might not apply if other kinds of subjects were used.
2. Setting–treatment interaction: An outcome found in one kind of setting may not happen if other kinds of settings were used.
3. Pretest–treatment interaction: Pretest may inform subjects to treatment .so as a result, they can yield an effect not generalizable to an unprotected population.
4. Subject effects: Subjects' attitudes and arrogance developed during the study could affect the generalizability of the results. Examples are the Hawthorne and the John Henry effects.
5. Experimenter effects: Unique Characteristics of a specific experimenter may limit generalizability to conditions with a different experimenter.

And, factors that jeopardize external validity according to Yu and Ohlund (2010) are as follows:

1. Reactive or interaction effect of testing: a pretest might increase or decrease a subject's sensitivity or responsiveness to the experimental variable. Indeed, the effect of pretest on subsequent tests has been empirically substantiated.
2. Interaction effects of selection biases and the experimental variable
3. Reactive effects of experimental arrangements: it is difficult to generalize to non-experimental settings if the effect was attributable to the experimental arrangement of the research.
4. Multiple treatment interference: as multiple treatments are given to the same subjects, it is difficult to control for the effects of prior treatments.

2.2.3.7 Relation between Internal Validity and External Validity

A great body of studies has been conducted on Relation between internal validity and external validity, most of them believed that there aren't separable. (e.g., Cronbach, 1982; Hammersley,

1991; Briggs, 2008; Wilson, 2016) opposed to Campbell and Stanley (1963).In this regard, Cronbach (1982) is opposed to the notion of Campbell and Stanley. He argued that if a treatment is expected to be relevant to a broader context, the causal inference must go beyond the specific conditions. If the study lacks generalizability, then the so-called internally valid causal effect is useless to decision-makers. Hammersley (1991) came to the conclusion that not only does the concept of external validity obscure the different possible interpretations of a failure to replicate, but it cannot be distinguished from internal validity. Findings are either true or false (or approximately true to some degree), they cannot be true in one sense but false in another. The idea that a hypothesis could be internally valid but externally invalid is therefore incoherent. This distinction is based on the false assumption that we can separate the discovery of causal relationships from the question of whether they apply to other cases. In a similar attitude, Briggs (2008) asserted that although statistical conclusion validity and internal validity together affirm a causal effect, construct validity and external validity are still necessary for generalizing a causal conclusion to other settings. After that Wilson (2016) argued that internal validity is a necessary condition for external validity. The distinction between internal and external validity could thus be thought of as analogous to the distinction between validity and soundness in arguments. An experiment that leads to results that happen to be widely applicable, but does so despite a lack of rigor in its research design would be equivalent to an invalid argument with a true conclusion. About the importance and role of External validity in thought, experimental research Wilson (2016) believed that although the philosopher is eager to focus on internal validity external validity is both more important and more difficult to ensure. Taking external validity seriously should lead philosophers to be far more involuntary about the limits of thinly thought experiments. Opposed to a great number of studies that believed in mutual relationships between two mentioned variables, and emphasizing on the importance of both of them equally, Campbell and Stanley (1963) asserted that internal validity is more important than external validity.

2.3. Empirical Findings

A small body of research has been conducted to explore the internal, external, and construct validity of research .based on their domain, these studies could be classified into two main groups. studies in ELT domain (e.g., Christ, 2007; Rahimi *et al.*, 2008; Mahboob *et al*., 2016; Razmjoo, 2016; Huggins-manley *et al*.,2019; Park *et al.*, 2019; Ebrahimi *et al*., 2020; Olubela & Adebanjo, 2020; Vosgerau , *et al*.,2020) and studies in non-ELT domain (e.g., Smith, 2005; Siegmund *et al*., 2015; Wilson, 2016; Orquin & Holmqvist, 2018; Patino & Ferreira, 2018; Zorlu & Sezek, 2019; Kool & Giller, 2020).

2.3.1. Empirical Findings of ELT Domain

Many studies have been done in the education field to investigate the internal , external and construct validity of researches (e.g., Christ, 2007; Rahimi *et al.*, 2008; Mahboob, *et al*., 2016; Razmjoo, 2016; Huggins-manley *et al*., 2019; Park *et al.*, 2019; Ebrahimi et al.2020; Olubela & Adebanjo, 2020; Vosgerau, *et al*., 2020).

Some researchers (Christ, 2007; Ebrahimi, 2020; Olubela & Adebanjo, 2020) tried to increase internal validity through the design of the study. For example, Christ (2007) provided a critical review of the scientific merit of both concurrent and nonconcurrent multiple baseline (MB) designs, relative to their capacity to assess threats of internal validity and establish experimental control. He believed that both nonconcurrent MB and concurrent designs are experimental designs. So, they can control and ruled out various threats of internal validity .each is capable in one aspect, nonconcurrent designs are more powerful in assessing the intervening effects of history, but might be more prone to threats of mortality. Finally, Christ (2007) asserted that MB designs need to have some characteristic to promote internal validity and experimental control include:"(a) specified experimental manipulations (i.e. IV conditions), (b) a priori hypotheses, (c) formative assessment schedules, (d) marked changes in the DV that coincide with IV manipulations, and (e) replication across an adequate number of data series" (p.457).And in a study conducted by Ebrahimi *et al*. (2020) among Iranian advanced learners, the researcher wanted to develop cultural identity between participants. So, he implemented a multiphase design. Firstly he used a qualitative design and to validate a model he investigate different factors of cultural identity interviews with 20 EFL

learners. Then, in the quantitative phase, the 30-item questionnaire (that was consist of four factors of cultural identity, a questionnaire was constructed which reflected these factors) went through an exploratory factor analysis for the sake of validity. The other study was done by Olubela and Adebanjo (2020) to find Learning Styles and Gender Effects effective on Secondary School Students' Learning Outcomes. In this study researchers, Adopted a pre-test, post-test, control group switching replication quasi-experimental design. A switching replication quasi-experimental design was used to increase the internal validity of the study and to establish the efficacy of the treatment.

Several studies (Rahimi *et al.*, 2008; Razmjoo, 2016) tried to increase the validity through the clear specification of instrument. For example, in a research done by Rahimi *et al.* (2008), one of the instruments to data collection was three questioners, so, to enhance the validity of research, the researchers translate the questioner into Persian to ensure that participants thoroughly understood the content of the question. In the same vein, the other study was presented by Razmjoo (2016), he mentioned: A clear picture of the instruments like questionnaires or tests must be provided along with their reliability and validity, and also dependability and credibility for observations and interviews.

Mahboob *et al. (*2016) tried to determine causal-like relationships between more than one independent variable (e.g., types of instruction, feedback) and more than one dependent variable (e.g., language acquisition, learning behaviors).to arrive at a valid conclusion about the causal-like relationship, the researcher systematically controls other independent variables that can interfere with the effect of the target-independent variable by holding them constant across the conditions. So to ensure that the result of the study is valid, he made a stable confined variables.

Moreover, some studies attempt to investigate construct validity (Huggins-Manley *et al*., 2019; Vosgerau, *et al*., 2020). For example, Huggins-Manley *et al*. (2019) tried to investigate construct validity when using operational virtual learning environment data. They consider it crucial to define desired constructs and their threats, so they map their VLE-based data and desired educational constructs onto the formal threats to construct validity, and developed actionable solutions and they consider it useful for other research projects. Ultimately, they conclude that

evidence of statistical conclusion validity, internal validity, and external validity, all are related together and to evidence of construct validity. Lately, to provide a clear example of how to establish construct validity in practice for self-control research conducted by Vosgerau *et al.* (2020). They first proposed a behavioral measure of self-control. Second, they used a measure in the main study to provide support for the construct validity. They indicated that the majority of articles they studied on self-control flop to support the construct validity of their measures. Thus, they warn us that it is possible that "the observed effects do not represent effects on self-control but something else" (p. 33).

Many researchers have emphasized the establishment of validity in qualitative research, In this regard, and since there wasn't any practical research in this field, a study was done to estimate the validity of qualitative research by Park *et al.* (2019). As the entire process of a qualitative study is carried out through the subjective eye of researchers, one of the key weaknesses of qualitative research is validity (Gibbert & Ruigrok, 2010). They analyzed 79 qualitative studies. They found that Operational measurement was the most frequently concerned validity in 79 studies. 68 of 79 (86.08%) In contrast, generalizability was less emphasized relatively. Only 30 of 79 (37.97%) were using at least one of the tactics of multiple cases and mixed methods in the research design process. This research exemplifies a meaningful impact on our understanding of validity notion in qualitative research by paradigm.

2.3.2. Empirical Finding of Non-ELT Domain

Some studies have been conducted to investigate the internal, external, and construct validity of studies, in non-English Field (e.g., Smith, 2005; Siegmund *et al*., 2015; Wilson, 2016; Orquin & Holmqvist, 2018; Patino & Ferreira, 2018; Zorlu & Sezek, 2019; Kool & Giller, 2020, Oesch , 2020).

In a study related to clinical research done by Smith (2005), he indicated that the construct validity of clinical measures is an ongoing process of discovery, Relating both to theories and the measures that exemplify them. Then he suggested five steps establishing construct validity as below: Careful theory specification, development of informative hypothesis tests, use of sound

research design, the examination of the degree to which observations confirm hypotheses, and ongoing revisions of both theory and measures. And at the end, they claimed several implications of this model for clinical research.

Siegmund *et al.* (2015) have conducted research in software engineering to find the answers to questions as, Should they focus on internal or external validity? Their findings indicated that it is an alarmingly high number of authors who do not seem to be aware of the threats to the validity of their study. Since they were one the advocator of the trade-off relationship between internal and external validity. And, many reviewers don't have any information about the trade-off between internal and external validity, but at the same time have strong opinions on maximizing one kind of validity. So the result would be the situation that getting a paper accepted is by chance rather than based on quality.

A number of studies have been conducted to find out the way for increasing the external validity of their research (e.g., Orquin & Holmqvist, 2018; Kool & Giller, 2020). For example, Orquin and Holmqvist (2018) researched eye-movement in psychology. They believed that eye movements are prepared in a laboratory environment. It may still be difficult to generalize it beyond this environment. (Even if the experiment uses an extensive range of stimuli), and since the eye-movement experiment cant under-sampling of naturalistic stimuli, so they cannot generalize anything beyond the sparse stimuli. They conclude that using a quasi-experimental design would help increase the external validity of eye-movement research. Recently Kool and Giller (2020) explored several on-farm experiments to assess how on-farm experimental studies address the scope or generalizability of their findings when based on a limited number of farms. Regarding external validity of research, they believed these researches suffer from external validity. Because for high external validity researcher should define and describe the research population and/or environment in which (they expect) the experimental finding to work. But most of them fail to do it.

A multicenter study in France (Patino & Ferreira,2018) conducted a randomized controlled experiment to test the effect of prone vs. supine positioning ventilation on mortality among patients with early, severe ARDS, and regarding increasing internal and external validity, they came to the below conclusion :

For increasing internal validity, mentioned two suggestion:

- Careful study planning: investigators should warrant careful study planning.
- Adequate quality control and implementation strategies: including adequate recruitment strategies, data collection, data analysis, and sample size.

For increasing External validity suggested 2 aspects as bellows:

- By using broad inclusion criteria that result in a study population that more closely resembles real-life patients.
- And, in the case of clinical experiment, by choosing interventions that are feasible to apply.

Several studies tried to increase internal and external validity by specific design models, or through combining two methods (e.g., Wilson, 2016; Zorlu & Sezek, 2019; Oesch, 2020). Wilson (2016) had some thought-provoking ideas about both the internal and external validity of thought experiments. About internal validity, He emphasized the importance of design for the benefit from this aspect. Wilson(2016) “experiment has internal validity only to the extent that it has been carefully and systematically designed with sufficient care to give a high degree of confidence that the results reported accurately measure the nature and the effect size of the intervention tested. Internal validity is a measure of the quality of the research design: The design of experimental trials is the subject of a massive literature” (p.4). In the field of thought experiments, the idea of internal validity is significantly complicated by the fact that thought experiments are a type of fiction, and that making judgments about cases presented in thought experiments is kind of responding to fiction. And through lack of external validity in the thought experiment he mentioned two new concepts as normative contextual variance and non-transferability of causal structures. He believed that “Thought experiments can lack external validity in at least two ways. First, if the ethical judgments that can be established as appropriate in the world of the thought experiment depend on features of the normative context that are not shared in other normative contexts. Call this normative contextual variance. Second, if the ethical judgments that can be established as appropriate in the world of the thought experiment presuppose causal structures that

are relevantly different from those that are present in other contexts. Call this non-transferability of causal structures" (p.12). Concerning to the role of design in increasing internal validity, Zorlu & Sezek (2019) used a salmon experimental design. To increase the relationship between the independent and dependent variables. Applying In this way it has increased the ability to explain the fact that changes in the independent variable lead to changes in the dependent variables using the same participants by justifying. This study determined a more powerful relationship between cause and effect variables as the application time or frequency increased. It can be said that the results performed using the Solomon research design increased the facility to generalize the application with other people (population validity), environments (ecological validity), experiments (experiment validity), and times (time validity) (Christensen, Johnson, and Turner, 2015). The Solomon research design increased the validity of the interpretations of the application's efficiency by eliminating these effects (Karasar, 2014). Besides a study was done by Oesch (2020), trying to increased internal validity by combining two experimental methods. In a factorial survey experiment and a natural experiment, as a result, combining the two experimental methods allows them to guarantee and increase internal and external validity.

2.4. Summary of Empirical Finding and Statement of the Gap

In this chapter a review of related literature is presented on the internal, external , and construct validity of researches in both ELT and non-ELT domain, empirical findings of studies related to all three types of validity in both domain .finally, this section is summarized all of the empirical findings and presented the gap at the end.

There were various kinds of studies in the ELT and non-ELT domain that tried to increase internal and external validity through the design of the study. (e.g., Christ, 2007; Ebrahimi, 2020 ; Olubela & Adebanjo, 2020 ;Wilson, 2016; Zorlu & Sezek , 2019; Orquin & Holmqvist,2018 ; Kool & Giller, 2020)and found that switching replication quasi-experimental design, concurrent and nonconcurrent multiple baseline (MB) designs, salmon experimental design , or combining two experimental methods have a great effect on increasing the internal and external validity of the studies.

Considering construct validity of studies some researchers found that (Huggins-Manley *et al*., 2019; Rahimi *et al.,* 2008; Razmjoo , 2016) defining desired constructs and their threats, a clear picture of the instruments like questionnaires or tests has a positive effect on the construct validity. In the same vein, some of them suggested (e.g., Smith, 2005; Vosgerau *et al.,* 2020) some steps in establishing construct validity.

The review of empirical studies seems to show that there is a gap in the literature. Although, there is somebody of researches to explore internal, external, and construct validity in both the ELT domain and non-ELT domain. But there is no comprehensive research related to Experimental findings in ELT researches. So preparing this kind of research is requisite, furthermore shed some light on the other researchers to have considerable efforts to establishing validity in their experiments.

CHAPTER THREE:

RESERCH METHOD

3.1. Overview

The present study aims to investigate the internal, external, and construct validity of experimental findings in ELT researches. So, to accomplish the purpose of this study, studying some valuable local and non-Iranian researches and investigate their internal, external, and construct validity and their correlation should be done. In this regard, this chapter aims at explaining the methodology which was applied to answer the research questions. It starts with participants, instruments, and the other era that are as bellows.

3.2. Research Design

Scientifics have categorized exploratory experiments in terms of what they lack: they lack direction from what has been called "local theories" of the target system or object under investigation. As Elliot (2007, p. 322) mentioned "Exploratory experiments were originally characterized merely in terms of what they lack". After that, O'Malley (2007) and Feest (2012) have provided positive characterizations, by identifying these experiments' roles in, concept formation, discovery, and the expansion of theory.

Burian (1997) proposed, the aim of the research is not testing or developing, or otherwise expressive an existing theory or hypothesis. According to his saying, exploratory experiments aim to produce significant findings of phenomena without appealing to a theory about these phenomena to focus experimental attention on a limited range of possible findings. The findings might be significant and noteworthy in a variety of aims going from the practical goal to learn how to manipulate a phenomenon to the theoretical goal to develop a theoretical framework that will help focus prospective experimental attention. According to waters (2007), exploratory experimentation is implanted within scientific inquiry that relies on a lot of theory so it's not without theory. All kinds of background theories are used to set up experiments, produce data, and conclusions. Accordingly, this study was an attempt to explore the aspects of the validity threats in the research published in the ELT domain.

According to Swedberg (2020, page.17) "Exploratory research is the soul of good research .without the ambition to say something new, research would come to a standstill. Non-exploratory

research can be definition only result in repetition of what is already known. And apart from studies that aim at replication, this will not move science forward". He believed that, two form of exploratory studies that have been the most common are the following 1) a new topic that has not been researched before 2) to yield new ideas and hypothesis an existing topic is explored, but failed to accurately verify these.

3.3. Sampling Procedure and Materials

The material of the study consisted of 40 ELT papers from Iranian and non-Iranian journals with high impact factors. To determine and locate the sample of papers, journal information in Scopus Index, and the official website of the Ministry of Science, Research and Technology were checked. Afterwards, 20 papers from TESOL quarterly, Language Learning Journal, and IRAL were selected. The Iranian papers were carefully chosen from The Journal of Teaching Language Skills (JTLS), Journal of English Language Teaching and Learning, and Applied Research on English Language. The papers were mainly selected via random sampling among the ones which reported a quantitative approach to the problem under investigation. They were published in the last ten years, since 2010. The selection was made based on the following criteria also the studies as participants of the current study are depicted in Table 3.1.

- The papers had to be written in English
- Papers had to include sufficient information for estimating their internal, external and construct validity.
- They had to be within in the ELT domain rather than other domains of applied linguistics.
- They had to enjoy experimental or quasi-experimental design only.

Table 3.1. Papers as Participants of the Study

Name of study	Authors	Type of Journal	Name of Journal
The effects of Curriculum-Based Measurement on EFL learners' achievements in grammar and reading	Tavakoli, M., & Atefi Boroujeni, S. (2012)	Iranian	Applied Research on English Language
The effects of captioning texts and caption ordering on L2 listening comprehension and vocabulary learning	Roohani, A., Rahimi Domakani, M., & Alikhani, F. (2013)	Iranian	Applied Research on English Language
The effect of reading purpose on incidental vocabulary learning and retention among elementary Iranian learners of English	Eghtesadi, A. R., & Momeni, S. (2014)	Iranian	Applied Research on English Language
The Impact of Podcasts on English Vocabulary Development in a Blended Educational Model	Mashhadi, A., & Jalilifar, A. (2016)	Iranian	Applied Research on English Language
The Effect of Argument Mapping Instruction on L2 Writing Achievement across Writing Tasks and Writing Components: A Case of Iranian EFL Learners	Malmir, A., & Khosravi, F. (2018)	Iranian	Applied Research on English Language
The Effect of Meta pragmatic Awareness, Interactive Translation, and Discussion through Video-Enhanced Input on EFL Learners' Comprehension of Implicature	Derakhshan, A., & Eslami, Z. (2020)	Iranian	Applied Research on English Language
The Impact of Gender and Task Nature on Iranian EFL Learners' Oral Corrective Feedback Preferences	Salehi, M., & Jafari Pazoki, S. (2020)	Iranian	Applied Research on English Language
The Effect of Oral Dialogue Journals on Iranian EFL Learners' Communicative Competence	Ramazanzadeh, A. (2011)	Iranian	Journal of English Language Teaching and Learning
The effect of gender and different levels of education on the relationship between students' achievement goal and their academic achievement	Rashidi, N. (2013)	Iranian	Journal of English Language Teaching and Learning
The Effect of Topic Bias on the Writing Proficiency of Extrovert/Introvert EFL Learners	Nama, S. N., & Moini, F. (2013)	Iranian	Journal of English Language Teaching and Learning
The Effect of Mnemonic Key Word Method on Vocabulary Learning and Long Term Retention	Hamzavi, R. (2014)	Iranian	Journal of English Language Teaching and Learning

The Effect of Transcribing on Beginning Learners' Phonemic Perception	Afsharrad, M., & Sadeghi Benis, A. R. (2014)	Iranian	Journal of English Language Teaching and Learning
The Effects of Direct Corrective Feedback and Metalinguistic Explanation on EFL Learners' Implicit and Explicit Knowledge of English Definite and Indefinite Articles	Rezazadeh, M., Tavakoli, M., & Eslami Rasekh, A. (2015).	Iranian	Journal of English Language Teaching and Learning
The Effect of Four Different Types of Involvement Indices on Vocabulary Learning and Retention of EFL Learners	Baleghizadeh, S., & Abbasi, M. (2013)	Iranian	The Journal of Teaching Language Skills (JTLS)
The impact of Using Computer-Aided Argument Mapping (CAAM) on The Improvement of Iranian EFL Learners Writing Self-Regulation	Pahlavani, P., & Maftoon, P. (2015).	Iranian	The Journal of Teaching Language Skills (JTLS)
The Effect of Mixed and Matched Level Dyadic Interaction on Iranian EFL Learners' Comprehension and Production of Requests and Apologies	Fakher, Z., Vahdany, F., Jafarigohar, M., & Soleimani, H. (2016)	Iranian	The Journal of Teaching Language Skills (JTLS)
The Impact of Task Complexity along Single Task Dimension on EFL Iranian Learners' Written Production: Lexical complexity	Izadpanah, S., & Shajeri, E. (2016).	Iranian	The Journal of Teaching Language Skills (JTLS)
The Effect of "Narrow Reading" on Learning Mid-Frequency Vocabulary: The Role of Genre and Author	Sotoudehnama, E., Ahmadi, M., & Asadi Zarmehri, M. (2020).	Iranian	The Journal of Teaching Language Skills (JTLS)
The Effect of Open vs. Closed Tasks on Improving Iranian EFL Learners' Oral Performance	Zohrabi, M., & Hassanpour, S. (2020)	Iranian	The Journal of Teaching Language Skills (JTLS)
The Effects of Collaborative Translation Task on the Apology Speech Act Production of Iranian EFL Learners	Kargar, A. A., Sadighi, F., & Ahmadi, A. R. (2012)	Iranian	The Journal of Teaching Language Skills (JTLS)
Language learning strategy use in context: the effects of self-efficacy and CLIL on language proficiency	Jaekel, N. (2018).	Non-Iranian	IRAL
Effects of pushed production of single word sand multiword patterns on L2 oral fluency: Some evidence from temporal measurements	Chan, H. (2019).	Non-Iranian	IRAL
Impacts of task complexity on the development of L2 oral performance over time	Kim, Y., & Payant, C. (2017)	Non-Iranian	IRAL

The effects of context and word exposure frequency on incidental vocabulary acquisition and retention through reading	Teng, F. (2019)	Non-Iranian	The Language Learning Journal
The impact of the 'writers' workshop' approach on the L2 English writing of upper-primary students in Lebanon	Al-Hroub, A., Shami, G., & Evans, M. (2019).	Non-Iranian	The Language Learning Journal
The combined effect of task repetition and post-task transcribing on L2 speaking complexity, accuracy, and fluency	Hsu, H. C. (2019).	Non-Iranian	The Language Learning Journal
The effectiveness of different explicit vocabulary-teaching strategies on learners' retention of technical and academic words	Alamri, K., & Rogers, V. (2018).	Non-Iranian	The Language Learning Journal
The effects of a guided reading intervention on reading comprehension: a study on young Chinese learners of English in Hong Kong	Nayak, G., & Sylva, K. (2013)	Non-Iranian	The Language Learning Journal
The effects of repetition and L1lexicalization on incidental vocabulary acquisition by Iranian EFL Learners	Heidari-Shahreza, M. A., & Tavakoli, M. (2016).	Non-Iranian	The Language Learning Journal
The effectiveness of group, pair and individual output tasks on learning phrasal verbs	Teng, M. F. (2020).	Non-Iranian	The Language Learning Journal
The Effectiveness of Drama as an Instructional Approach for the Development of Second Language Oral Fluency, Comprehensibility, and Accentedness	Galante, A., & Thomson, R. I. (2017).	Non-Iranian	TESOL QUARTERLY
The Effect of Focused Written Corrective Feedback and Language Aptitude on ESL Learners' Acquisition of Articles	Sheen, Y. (2017).	Non-Iranian	TESOL QUARTERLY
Effects of Distributed Retrieval Practice Over a Semester: Cumulative Tests as a Way to Facilitate Second Language Vocabulary Learning	Nakata, T., Tada, S., Mclean, S., & Kim, Y. A. (2020).	Non-Iranian	TESOL QUARTERLY
Predictive Effects of Writing Strategies for Self-Regulated Learning on Secondary School Learners' EFL Writing Proficiency	Teng, F., & Huang, J. (2019).	Non-Iranian	TESOL QUARTERLY
The Effects of Foreign Accent on Perceptions of Nonstandard Grammar: A Pilot Study	Ruivivar, J., & Collins, L. (2018).	Non-Iranian	TESOL QUARTERLY

Effects of Video-Based Interaction on the Development of Second Language Listening Comprehension Ability: A Longitudinal Study	Saito, K., & Akiyama, Y. (2018).	Non-Iranian	TESOL QUARTERLY
Effects of Instruction on Adolescent Beginners' Acquisition of Request Modification	Li, Q. (2012).	Non-Iranian	TESOL QUARTERLY
The Effect of Imagery and On-Screen Text on Foreign Language Vocabulary Learning From Audiovisual Input	Peters, E. (2019).	Non-Iranian	TESOL QUARTERLY
Effects of Pre task Modeling on Attention to Form and Question Development	Kim, Y. (2013).	Non-Iranian	TESOL QUARTERLY
Effects of the Manipulation of Cognitive Processes on EFL Writers' Text Quality	Ong, J., & Zhang, L. J. (2013).	Non-Iranian	TESOL QUARTERLY
The Effects of Explicit Instruction on the Reading Performance of Adolescent English Language Learners With Intellectual Disabilities	Reed, D. K. (2013).	Non-Iranian	TESOL QUARTERLY

3.4. Data Collection

The research began with the identification of the target journals and then the sample of research papers were collected. Based on the criteria described above, the researcher selected the journals according to their indexing information declared on the journal websites as well as the information publicized on official indexing websites such as Scopus and the Ministry of Science, Research and Technology. After the researcher had selected the journals, the papers which met the criteria were extracted from which a random sample of 40 papers, 20 papers from Iranian journals written by Iranian authors, and 20 papers from non-Iranian journals written by non-Iranian authors was selected. To collect the required data, the researcher developed a checklist of the threats presented in Table 3.2, to the internal and external validity of experimental and quasi-experimental research based on various available sources (Cook & Campbell, 1979; Finger & Rand, 2003; Hamersley, 1987; Rosenthal, 2002; Taylor, 1994; Ary *et al*., 2018) which formed the basis for evaluating the sample of papers in this study.

Table 3.2. Threats Checklist of Validity

Internal validity threats	External validity threats
History	Selection-treatment interaction
Maturation	setting-treatment interaction
Testing Effect	Pretest-treatment interaction
Instrumentation	Subject and experimenter effect
Regression	
Selection Bias	
Mortality	
Selection-Maturation Interaction	
Experimenter &Subject effect	
Diffusion	

3.5. Reliability of the Study

To have a reliable evaluation of the paper, the researcher conducted a check-list-based evaluation of the papers twice with a three-week interval. That is, two checklists were developed for each paper without the first one being consulted while the second one was being completed. To ensure the intra-rater reliability of the collected data the researcher estimated the agreement between the two sets of the checklists which equaled 97 percent. It is worth mentioning that although the special focus of the researcher was on the method section of the papers, however, the whole body of papers was considered when evaluating the papers. Finally, the results were then extracted and tabulated for further analysis.

3.6. Data Analysis

The researcher attempted to extract the frequencies, percentage of main aspects of the construct, external and internal validity which were operationally defined to be the adequate specification of construct, random sampling, and random assignment, respectively in Iranian and non-Iranian papers. And find the degree of ruling out threats of internal and external validity according to Table 3.2, After extracted the frequencies of the criteria, the chi-square goodness of fit test and chi-square independence as a statistical procedure were used to analyze data and to see how much difference exists between our observed counts in the accessible population and the ideally counts expected to be observed.

CHAPTER FOUR:

RESULTS

4.1. Overview

This chapter is going to present the results of the analyses done with regard to the research questions of the study. The main purpose here is to test the hypotheses one by one based on the data extracted from the Iranian and non-Iranian research papers in ELT domain. In order to have an orderly presentation of the results, first the research question of the study is presented and then after the presentation of the frequencies of the threats, the null hypothesis will be tested. Finally, the discussion of the findings is presented.

4.2. Internal, External and Construct Validity of ELT Papers

Research Question 1: Is there any significant difference in frequency between studies that assured internal validity through the random assignment and what was ideally expected to be observed in these papers?
Research Question 2: Is there any significant difference in frequency between studies that assured external validity through random selection and what was ideally expected to be observed in these papers?
Research Question 3: Is there any significant difference in frequency between studies that assured construct validity through adequate specification of construct and what was ideally expected to be observed in these papers?

The primary purpose of this study was to explore if the selected papers in applied linguistics yield expected degree of construct, external and internal validity. To this end, the researchers attempted to extract the frequencies of main aspects of construct, external and internal validity which were operationally defined to be adequate specification of construct, random sampling, and random assignment, respectively. Figure 4.2, illustrates the distribution of these features in the sample articles. Accordingly, it has to be mentioned that 18 papers specified their constructs, 3 papers featured random sampling and 23 papers enjoyed random assignment. Overall, it can be argued the a little over 50 percent of the articles met the basic internal validity criterion, random assignment, and a little less than 50 percent of the papers met the basic construct validity criterion,

adequate construct specification; however, very few of the papers took the basic external validity criterion, random sampling, into consideration.

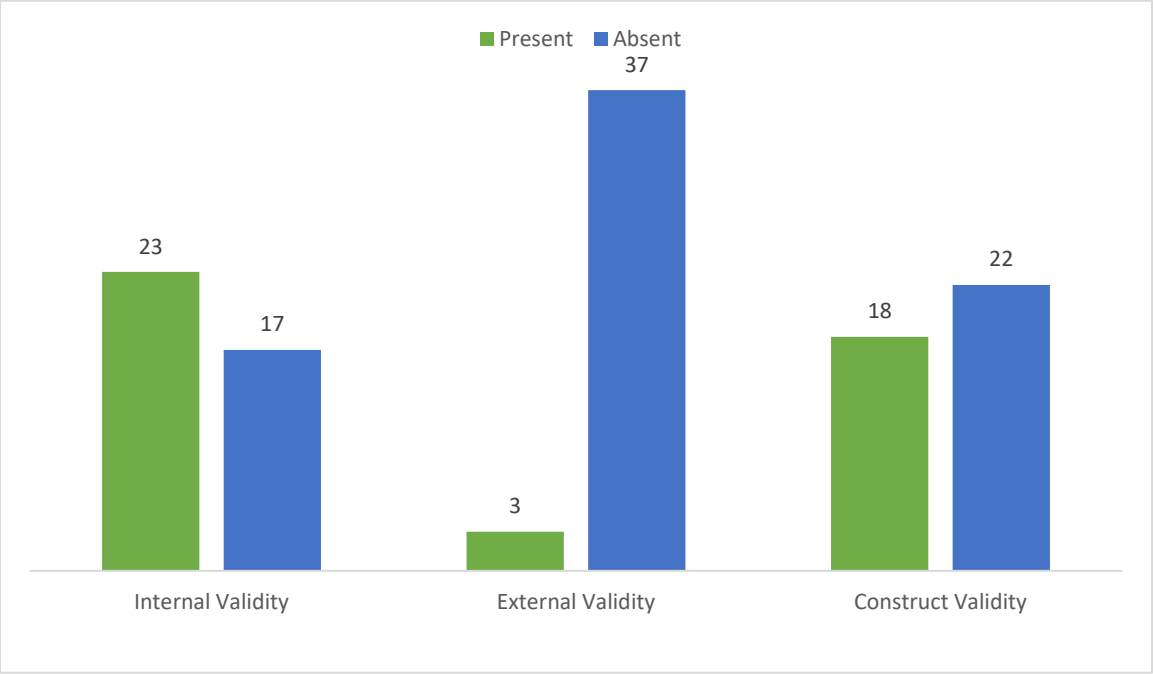

Figure4.1. Distribution of the Sample Articles in terms of their Validity

Table 4.2, shows the frequency, percentage of construct, external, and internal validity criteria in Iranian and non-Iranian papers. In addition, the cumulative frequency and percentage of the criteria in the sample were reported. Chi-square was also used to see if this distribution matches the expected distribution.

Table 4.1. Distribution of Different types of Validity in the Sample

Samples	N	Internal validity			External validity			Construct validity		
		n	Percentage	X^2	n	Percentage	X^2	n	Percentage	X^2
Sum	40	23	57%	256.17**	3	7.5%	1296.85**	18	45%	361.29**

** Significant at p < .01

With respect to the first, second and third research questions, Table 4.2, demonstrated, out of the 40 papers analyzed in this study, 23 papers (57 percent) met internal validity criterion, 3 (7.5 percent) papers specified their external validity, and 18 papers (45 percent) met construct validity. Based on the observed results, the researchers investigated if the observed results were significantly different from what was ideally expected to be observed in these papers. To this end, a chi-square test was run, the results of which are reported in Table 4.2, and the observed chi value for internal validity criterion was significant ($X^2 = 256.17, p = .00 < .01$). Accordingly, it was concluded that there was a significant difference in frequency between the sample of studies in this research and the ideal condition that assured internal validity through random assignment. The observed chi-value was also significant ($X^2 = 1296.85, p = .00 < .01$) for the distribution of external validity criterion. Accordingly, it was argued that the sample papers failed to meet the external validity criterion of random sampling. In addition, the observed chi value was significant ($X^2 = 361.29, p = .00 < .01$) for construct validity criterion. Accordingly, it was concluded that generally, the sample papers failed to assure construct validity criterion through adequate specification of construct so the null hypothesis is rejected.

4.3. Distribution of Iranian and Non-Iranian Papers In Terms of Their Validity

Research Question 4: Is there any significant difference in frequency between studies that assured internal, external validity and valid measure between Iranian and non-Iranian papers?

Regarding to answer the forth question of the study, to trace the possible differences between the papers written by Iranian authors and those by non-Iranian authors. As mentioned before, 20 papers in the sample were authored by Iranian researchers and the other 20 papers were written by non-Iranian .Table 4.3. shows the distribution of major criteria of construct validity, external validity and internal validity in these two sets of papers.

Table 4.3. Distribution in Validity Criteria in Iranian and Non-Iranian Papers

Samples	N	Internal validity				External validity				Construct validity			
		n	Percentage	X^2	*p*	n	Percentage	X^2	*p*	*n*	Percentage	X^2	*p*
Iranian papers	20	12	60	.10	.74	1	5	.36	.54	10	50	.40	.52
Non-Iranian papers	20	11	55			2	10			8	40		

** Significant at p < .01

As it was mentioned earlier, the sample contained 20 Iranian papers and the same number of non-Iranian papers. Table 4.3, Illustrate the distribution of these two sets of papers in terms of their construct, external and internal validity. As mentioned before, they were measured with regard to adequate specification of construct, random sampling, and random assignment, respectively. Accordingly, in 10 Iranian papers and 8 non-Iranian papers were adequate specification of constructs which imply meeting the basic criterion of construct validity. Moreover, two non-Iranian papers and only one iranian papers reported random sampling, the basic external validity criterion. Finally, 12 Iranian papers and 11 non-Iranian papers featured the basic internal validity criterion, random assignment. All in all, it may be argued that roughly half the Iranian and non-Iranian papers featured construct validity and internal validity. In addition, in the majority of the sample papers, the basic external validity criterion, random sampling, was violated. It may be concluded that both Iranian and non-Iranian researchers attend construct and internal validity criteria much more seriously than external validity criterion. The results of chi-square test reported in Table 4.3, shows that there is no significant difference between the Iranian and non-Iranian papers in terms of their meeting construct validity ($X^2 = .40,\ p = .52 > .05$), external validity ($X^2 = .36,\ p = .54 > .05$), and internal validity ($X^2 = .10,\ p = .74 > .05$) criteria. Accordingly, it was concluded that both Iranian papers and non-Iranian ones are similar in terms of violating different types of validities and the null hypothesis is accepted.

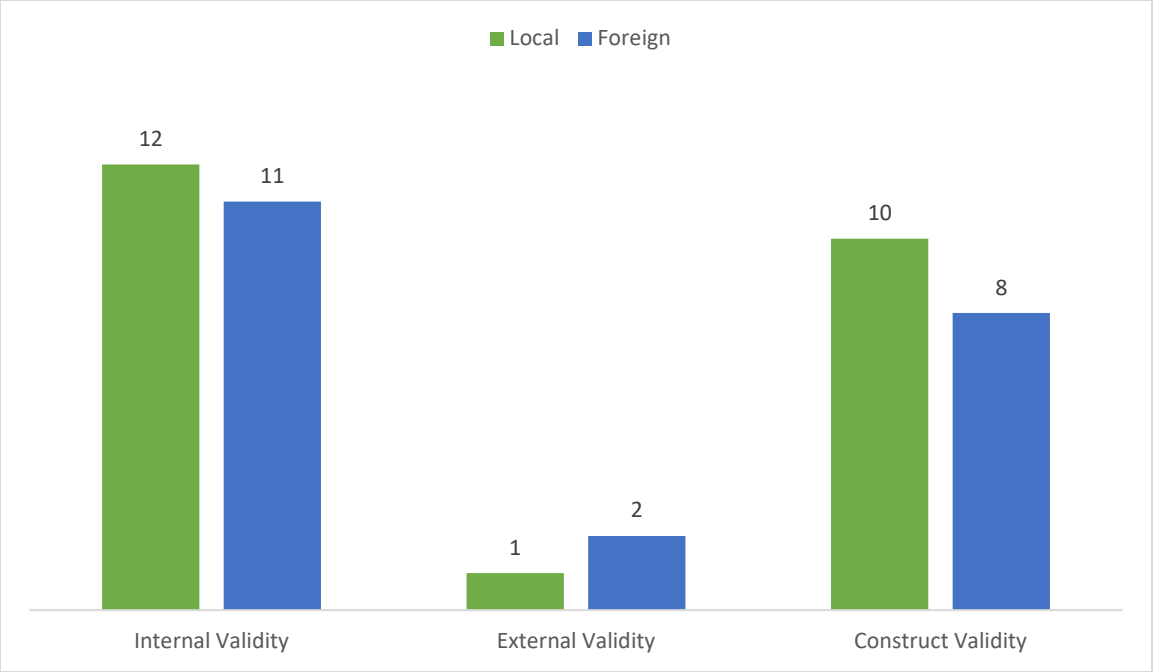

Figure 4.2. Distribution of Iranian and non-Iranian articles in terms of their validity

4.4. Threats to Internal and External Validity

4.4.1. Threats to Internal Validity

Research Question 5: Is there any significant difference in frequency between studies that addressed the threat to internal validity and what was ideally expected to be observed in these papers?

In addition, in order to have a more comprehensive analysis of the external validity and internal validity with regard to the sample papers in this study, and answer the fifth question of study, the distribution of the additional threats of previously defined as the components of these two types of validity were also considered. Table 4.4.1, shows the distribution of these threats in the sample papers in terms of frequency and percentage. In addition, chi-square test was used to compare the observed distribution with the ideal condition expected for each threat.

Table 4.4.1 Distribution of Different Internal Validity Criteria in the Sample

	Present		Absent		Not Related		
	N	Percentage	N	Percentage	N	Percentage	X^2
History	24	58	8	22	8	20	802.17**
Maturation	22	55	9	22	9	23	925.19**
Testing Effect	19	48	13	25	8	27	1252.26**
Instrumentation	36	90	4	10	-	-	81.01**
Regression	1	2.5	1	2.5	38	95	8839.20**
Selection Bias	24	60	16	40	-	-	1521.15**
Mortality	5	13	-	-	35	87	7396.75**
Selection-Maturation Interaction	-	-	12	30	28	70	5602.98**
Experimenter &Subject effect	6	15	6	15	28	70	4957.72**
Diffusion	22	55	18	45	-	-	1936.19**

** Significant at p < .01

Table 4.4.1, shows the distribution of the internal validity threat as observed in the sample papers.

Among the measured threats, instrumentation was the less frequent threat followed by selection bias, history, maturation, diffusion and testing effect. By contrast, selection-maturation interaction, regression, mortality, experimenter-subject effect were the less important ones. However, in order to probe if the distribution matched the ideal distribution of the threats, chi-square analyses were done. The results are shown in table 4.4.1, based on which, the observed chi value is significant for history ($X^2 = 802.17,\ p = .00 < .01$), maturation ($X^2 = 925.19,\ p = .00 < .01$), testing effect ($X^2 = 1252.26,\ p = .00 < .01$), instrumentation ($X^2 = 81.01,\ p = .00 < .01$), regression ($X^2 = 8839.20,\ p = .00 < .01$), selection bias ($X^2 = 1521.15,\ p = .00 < .01$), mortality ($X^2 = 7396.75,\ p = .00 < .01$), selection-maturation interaction ($X^2 = 5602.98,\ p = .00 < .01$), experimenter-subject interaction ($X^2 = 4957.72,\ p = .00 < .01$), and

diffusion ($X^2 = 1936.19, p = .00 < .01$). Accordingly, it was concluded that there were significant differences in distribution of the internal validity threats as observed in the sample and the ideal expected distribution. Accordingly, it may be argued that the findings reported in these papers are questionable in terms of meeting internal validity criteria so the null hypothesis is rejected.

4.4.2. Threats to External Validity

Research Question 6: Is there any significant difference in frequency between studies that addressed the threat to external validity and what was ideally expected to be observed in these papers?

In order to answer the sixth question of the study, external validity of the findings of the sample papers are also examined based on the threats which negatively affect the generalizability of the results. As shown in table 4.4.2, Setting-treatment interaction and Pretest-treatment interaction are the most important threat and then selection-treatment interaction and subject-experimenter effect considered in the sample papers. In addition, Based on the results reported in Table 4.4.2, the observed chi value is significant for selection-treatment interaction ($X^2 = 8281.86,\ p = .00 < .01$), setting-treatment interaction ($X^2 = 925.19,\ p = .00 < .01$), testing effect ($X^2 = 1252.26,\ p = .00 < .01$), instrumentation ($X^2 = 9409.98,\ p = .00 < .01$), pretest-treatment interaction ($X^2 = 5905.98,\ p < .01$), and subject and experimenter effect ($X^2 = 1521.15,\ p = .00 < .01$). Accordingly, it was concluded that there were significant differences in the distribution of the external validity threats in the sample and the ideal expected distribution so the null hypothesis is rejected. That is, the findings reported in the sample papers are hardly generalizable.

Table 4.4.2. Distribution of Different External Validity Criteria in the Sample

	Present		Absent		Not Related		
	N	Percentage	N	Percentage	N	Percentage	X^2
Selection-treatment interaction	3	7.5	37	92.5	-	-	8281/86**
Setting-treatment interaction	-	-	40	100	-	-	9409/98**
Pretest-treatment interaction	-	-	30	75	10	25	5905/98**
Subject and experimenter effect	13	33	9	22	18	45	2377/44**

** Significant at p < .01

4.4.3. Distribution of Iranian and Non-Iranian Papers In Terms of External Validity

In order to have a more detailed analysis of the Iranian and non-Iranian papers in terms of the distribution of threats to external and internal validity, further comparisons of the two groups of papers were made. The results are shown in table 4.4.3.

Table 4.4.3. Threats to External Validity of Iranian and Non-Iranian Papers

	Iranian Articles			Non-Iranian Articles			X^2	p
	Present	Absent	Not related	Present	Absent	Not related	.	
Selection-treatment interaction	5	95	-	10	90	-	.36	.54
Setting-treatment interaction	-	100	-	-	100	-	.00	1.00
Pretest-treatment interaction	-	75	25	-	25	75	8.12	.00
Subject and experimenter effect	10	25	65	55	20	25	9.89	.00

As demonstrated in table 4.4.3, there were significant differences between the two groups of papers in terms of pretest-treatment interaction ($X^2 = 8.12,\ p = .00 < .01$), and subject-experimenter effect ($X^2 = 9.89,\ p = .00 < .01$). However, these two groups of papers were not significantly different in terms of selection-treatment interaction and setting-interaction treatment. With regard to the frequencies reported in table 4.4.3, it was concluded that non-Iranian papers are controlling pretest-treatment interaction and subject-experimenter effect more seriously.

4.4.4 Distribution of Iranian and Non-Iranian Papers in Terms of Internal Validity

Table 4.4.4, demonstrates the comparison of the distribution of internal validity threats as observed in iranian and non-Iranian papers. The observed chi values are not significant for history ($X^2 = .15,\ p = .92 > .05$), maturation ($X^2 = .15,\ p = .92 > .05$), testing effect ($X^2 = 2.50,\ p = .28 > .05$), instrumentation ($X^2 = .00,\ p = 1.00 > .05$), regression ($X^2 = 2.10,\ p = .14 > .05$), selection bias ($X^2 = .41,\ p = .51 > .05$), mortality ($X^2 = 2.05,\ p = .15 > .05$). however, these two groups of papers were different in terms of selection-maturation interaction ($X^2 = 12.00,\ p = .00 < .01$), experimenter-subject interaction ($X^2 = 7.63,\ p = .00 < .01$), and diffusion ($X^2 = 8.12, p = .00 < .01$). Accordingly, it was argued that, with regard to the observed threats in both groups, these threats were better controlled in Non-Iranian papers.

Table 4.4.4. Threats to Internal Validity of Iranian and Non-Iranian Papers

	Iranian Articles			Non-Iranian Articles			X^2	p
	Present	Absent	Not related	Present	Absent	Not related		
History	55	25	20	60	20	20	.15	.92
Maturation	55	25	20	55	20	25	.15	.92
Testing Effect	35	25	40	60	25	15	2.50	.28
Instrumentation	90	10	-	90	10	-	.00	1.00
Regression	5	5	90	-	-	100	2.10	.14
Selection Bias	65	35	-	55	45	-	.41	.51
Mortality	5	-	95	20	-	80	2.05	.15
Selection-Maturation Interaction	-	30	70	30	-	70	12.00	.00
Experimenter &Subject effect	10	25	65	20	5	75	7.63	.02
Diffusion	35	65	-	75	25	-	8.12	.00

CHAPTER FIVE:

DISCUSSION AND CONCLUSION

5.1. Overview

This study shed light on the unique characteristic of validity within the experimental research paradigm. Although some scholars have suggested that the notion of validity and techniques used to ensure it may differ depending on the research results, to our knowledge this is the first study to highlight the unique quality of validity in experimental research by providing a systematic review of validity frameworks established in the quantitative research method. As such our comparative discussion warrants benefits to enrich our understanding of the distinct characteristic of validity in Iranian and non-Iranian papers. To this end, 40 experimental ELT studies were investigated through internal, external, and construct validity. This chapter presents a summary of the findings, a discussion of the findings, the pedagogical implication of the study, and the recommendation for future research.

5.2. Discussion and Conclusion

The result section addressed the six research questions posed in this exploratory study. Table 4.2, answers the first, second, and third questions by showing all researchers attend construct and internal validity criteria much more seriously than external validity criteria. According to our investigation, 23 of 40 articles were met internal validity (%57), In contrast, external validity was less emphasized relatively in 3 of 40 articles (%7.5). And regarding Construct Validity near half of the articles met the essential criteria, about 18 of 40(%45). It was demonstrated that the papers generally failed to meet the ideal condition that assured internal validity through random assignment, construct validity through the adequate specification of the construct, and the external validity criterion of random sampling, as a result the null hypothesis is rejected.

The result of this study shed light that both Iranian and non-Iranian scholars considered the internal validity of their experiments more important than external validity. It was anticipated many years ago. Campbell and Stanley (1963) stated that although a good study should be strong in both types of validity, internal validity is indispensable and more important, while the question of external validity is never entirely answerable. External validity is concerned with whether the same result of a given study can be observed in other situations. It seems like an inductive

implication, so, this question will never be conclusive. No matter how many new cases coincide with the previous finding, it takes just one counter-example to weaken the external validity of the study. In other words, Campbell and Stanley's statement implies that internal validity is more important and crucial than external validity.

Regarding the trade-off notion of internal and external validity, most of the papers that benefit from internal validity lacks external validity at the same time, the conclusion according to Table 4.2, is in line with (Bernstein, 2018; Lewis *et al*., 2006). Bernstein (2018) believed that, by controlling the experimental conditions perfectly, we can be assured the treatment is causing the effect and we meet the internal validity criteria. However, to have internal validity, the more we manipulate nature, the less we can generalize from our study to the larger context so we face a lack of external validity. Lewis, Perry, and Murata's (2006, p. 8) remarked that "the very qualities that suit an innovation to controlled trial may handicap it at the later stage of broad dissemination." seems there is often a tradeoff between.

Furthermore, another aspect of this study was the importance of random assignment to increase internal validity. Thus, this aspect is compatible with other scholars' views (e.g., Taylor, *et al*., 2008; Flannelly, *et al*., 2018). Taylor *et al* (2008) stated that, strong internal validity through randomizing participants to experimental conditions, guarantee True experiments designs but Quasi-experimental designs have weaker internal validity. In the same vein regarding to importance of random assignment in experimental research, flannelly *et al* (2018) believed that in educational research, instead of assigning individual students to groups to have experimental research, whole classes of students are assigned to groups, which makes the design easy to the device in educational settings and the design will be quasi-experimental. However, the design is subject to selection bias because of the lack of random assignment to groups.

On the other hand, the result of this study regarding the ignorance of external validity, presented in Table 4.2, was in contrast with other scholar's views (e.g., Briggs, 2008; Cronbach, 1982; Findley, 2020). Briggs (2008) asserted that although statistical conclusion validity and internal validity together confirms a causal effect, construct validity and external validity are still essential for generalizing a causal conclusion to other settings. In a similar vein, Cronbach (1982)

argued that if a treatment is expected to be relevant to a broader context, the causal inference must go further than the specific conditions. If the study lacks external validity and generalizability, then the so-called internally valid causal effect is useless to decision-makers. As Findley (2020) asserted, there is not any prior reason to believe that internal validity is more important and more significant than external validity. Ignoring external validity can potentially be as harmful as ignoring internal validity. “Even the gold standard randomized experiment faces challenges such as attrition, noncompliance, and spillover, so there are often exist inferential challenges that limit the applicability of an inference. A similar dynamic characterizes external validity, and the task is, therefore, to make credible, rather than universally applicable, inferences about external validity” (p. 21).

Although we expected non-Iranian papers benefited strongly from all validity types in comparison to Iranian papers, but according to the results obtained for the fourth research question in Table 4.3, it was demonstrated that all papers were not different in terms of the ideal condition that assured internal validity through random assignment, construct validity through the adequate specification of the construct, and the external validity criterion of random sampling.

Finally, results obtained for the fifth and sixth research questions through threats to internal and external validity, the null hypothesis is rejected. Because according to Tables 4.4.1, and 4.4.2, the papers generally failed to meet the needy criteria. Pretest-treatment interaction and Setting-treatment were the most important threat and then selection-treatment interaction and Subject and experimenter effect in the sample papers as the threats to external validity and diffusion, Selection bias, selection-treatment interaction, testing effect, maturation, history, subject and experimenter effect, instrumentation, regression, and mortality as the threats to the internal validity. The results imply that the current papers published in the domain of ELT, no matter it is published in Iranian or non-Iranian journal, are far from the ideal criteria, or in simple words, the standards agreed upon as construct, external and internal validity (Cook & Campbell, 1979; Finger & Rand, 2003; Hammersley,1987; Rosenthal, 2002; Taylor, 1994). Because the experimental and quasi-experimental designs are originally borrowed from the neighboring field of psychology and later adopted by the experts of the ELT domain from the broader field of education, it is necessary to reconsider the epistemological aspects of the method and see if it is inherently possible to apply

the experimental or quasi-experimental method in ELT domain as defined in primary sources of experimental research design. That is, the reason why these research papers were significantly, far away from the criteria held by experimental design ,because this design is an inherently strange bedfellow for most research inquiries done in the domain of ELT.

5.3. Implications for Practice

The findings of this study implied that the papers published in both Iranian and non-Iranian journals are comparable in terms of construct, external, and internal validity and are significantly different from the ideal experimental or quasi-experimental criteria of validity depicted in the key sources of research design. This may have several implications for:

- The journal officials. First, the journals publishing papers in the ELT domain are required to reconsider the issue of construct, external and internal validity more rigorously. To this end, it is recommended that they provide authors with explicit guidelines on what are they expected to provide as the evidence of construct, external and internal validity in their papers if they adopt an experimental or quasi-experimental design in their inquiries. The same procedure has to be followed when the paper is refereed.

- Reviewers, so, a clear statement of the list of required proofs has to be provided in the checklist for the reviewers so that a reliable assessment of the different aspects of validity is done when the paper is reviewed and sound comments are delivered to the authors about the construct, external and internal validity of their findings. This has to be more impeccably done by the Iranian journals especially about their observed poorer condition in terms of observing pretest-treatment interaction, subject-experimenter interaction, selection-maturation interaction, experimenter and subject effect, and diffusion which were significantly less attended in Iranian papers.

- academicians are recommended to more scrupulously attend the validity debate and its criteria in experimental and quasi-experimental design while presenting the related issues

in research methodology courses, especially, in master's and Ph.D. levels so that the would-be researchers graduating at these levels can have a better understanding of the threats to construct, external and internal validity and consider them when they are researching ELT domain.

- Policymakers, since it helps them make informed and knowledgeable decisions based on the accurate exploring of forty authentic papers through validity criteria, rather than a general frame of one qualitative or quantitative paper.

5.4. Suggestion for Further Research

The current study represents a meaningful contribution to the understanding of internal, external, and constructs validity notion of the experimental research paradigm. By reviewing the prior discussion of validity, this study attempt to expand our sight for considering all three kinds of validity important. Further research in this area is critical, not only to enrich understanding of validity but also to investigate all four aspects of design validity with more Iranian and non-Iranian samples. Considering the limitations of the study, especially, in terms of the number of journals and papers included in the sample of the study, it seems necessary that further research is done on a larger body of papers published in Iranian and non-Iranian journals since the replication of this study on a larger scale may entail more credible results. Besides, it is suggested that the validity of the papers published in applied linguistics journals, both Iranian and non-Iranian ones, is investigated concerning the non-ELT domains, as well.

References

American Psychological Association. (1954). Technical recommendations for psychological test and diagnostic techniques. *Psychology Bulletin Supplement, 51*, 1–38.

American Psychological Association. (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association

American Psychological Association. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Anastasi A. (1954). *Psychological Testing*. New York, NY: Mac-Millan.

Ary, D., Jacobs, L. C., Irvine, C. K. S., & Walker, D. (2018). *Introduction to research in education*. Cengage Learning.

Barbour, R. S. (2001). Checklists for improving rigour in qualitative research: a case of the tail wagging the dog? . *Bmj*, *322*(7294), 1115-1117.

Bernstein, J. L. (2018). Unifying SoTL methodology: Internal and external validity. *Teaching & Learning Inquiry*, *6*(2), 115-126.

Bornhöft, G., Maxion-Bergemann, S., Wolf, U., Kienle, G. S., Michalsen, A., Vollmar, H. C. & Matthiessen, P. F. (2006). Checklist for the qualitative evaluation of clinical studies with particular focus on external validity and model validity. *BMC Medical Research Methodology*, *6*(1), 56-77.

Borsboom, D., G.J. Mellenberg, & J. van Herden. (2004). the concept of validity. *Psychological Review, 111* (4), 1061–71

Briggs, D. C. (2008). Comments on Slavin: Synthesizing causal inferences. *Educational Researcher*, *37*(1), 15-22.

Burian R., 1997, 'Exploratory Experimentation and the Role of Histochemical Techniques in the Work of Jean Brachet, 1938–1952', *History and Philosophy of the Life Sciences*, 19: 27–45.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*, 297–312.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multi trait-multimethod matrix. *Psychological bulletin, 56*(2), 81.

Campbell, D. T., & Stanley, J. C. (1963).*Experimental and quasi-experimental designs for research*. Skokie, IL: Rand McNally.

Chatterji, M. (2007). Grades of evidence: Variability in quality of findings in effectiveness studies of complex field interventions. *American Journal of Evaluation, 28*(3), 239-255.

Christ, T. J. (2007). Experimental control and threats to internal validity of concurrent and non-concurrent multiple baseline designs. *Psychology in the Schools, 44*(5), 451-459.

Christensen, L. B., Johnson, R. B., & Turner, L. A. (2015). *Research methods: Design and analysis*. Ankara: Anı.

Churchill, G. A., Jr. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research, 16*, 64-73

Cook. T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis for field studies.* Skokie, IL: Rand McNally.

Crack, T. F., Gieves, R., & Lown, M. G. (2011). Put your best food forward: A pre-submission checklist for journal articles. Retrieved 17 January, 2021 from http://jfe.rochester.edu/checklist.pdf

Cronbach, L.J. (1971). *Test validation in educational measurement*. Washington, DC: American Council on Education.

Cronbach, L. J. (1982). In Praise of Uncertainty. *New directions for program evaluation*.

Cronbach, L. J. (1988). Five perspectives on validity argument. *Test validity*, 3-17.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin, 52*(4), 281.

Crooks, T.J., K.T. Kane, and A.S. Cohen. 1996. Threats to the valid use of assessments. *Assessment in Education: Principles, Policy & Practice 3* (3), 265–85.

Derntl, M. (2014). Basics of research paper writing and publishing. *International Journal of Technology Enhanced Learning, 6*(2), 105-123.

Donovan, S. K. (2007). The importance of resubmitting rejected papers. *Journal of Scholarly Publishing*, *38*(3), 151-155

Durant, R. H. (1994). Checklist for the evaluation of research articles. *Journal of Adolescent Health*, *15*, 4-8.

Ebrahimi, S., Afraz, S., & Samimi, F. (2020). Validation of a Preliminary Model of Cultural Identity for Iranian Advanced EFL Learners: A Structural Equation Modeling Approach. *International Journal of foreign Language Teaching and Research*, *8*(30), 115-137.

Elliott K (2007) Varieties of exploratory experimentation in nanotoxicology. *Historical Philosophy of Life Science 28* (3), 313–336.

Feest, U. (2012). Exploratory experiments, concept formation, and theory construction in psychology. *Scientific concepts and investigative practice*, *3*, 167-189.

Findley, M. G., Kikuta, K., & Denly, M. (2020). External Validity. *Annual Review of Political Science forthcomin,* 1-51.

Finger, M. S., & Rand, K. L. (2003). Addressing validity concerns in clinical psychology research. In M. C. Roberts & S. S. Ilardi (Eds.), *Handbook of research methods in clinical psychology* (pp. 13–30). Malden, MA: Blackwell.

Flannelly, K. J., Flannelly, L. T., & Jankowski, K. R. (2018). Threats to the internal validity of experimental and quasi-experimental research in healthcare. *Journal of health care chaplaincy*, *24*(3), 107-130.pared in a research study.

Gay, L.R., & Airasian, P.W. (2000).*Educational research: Competencies for analysis and application* (6th ed.). Englewood Cliffs, N.J.: Prentice Hall.

Gersten, R., & Baker, S. (2002). The relevance of Messick's four faces for understanding the validity of high-stakes assessments. *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation*, 49-66.

Gibbert, M., & Ruigrok, W. (2010). The "what" and "how" of case study rigor: Three strategies based on published work. *Organizational Research Methods*, *13*(4), 710-737.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and psychological measurement*, *6*(4), 427-438.

Hammersley, M. (1987). Some notes on the terms 'validity' and 'reliability. *British Educational Research Journal*, *13*(1), 73-81.

Hammersley, M. (1991). A note on Campbell's distinction between internal and external validity. *Quality and Quantity*, *25*(4), 381-387.

Heeler, R. M., & Ray, M. L. (1972). Measure validation in marketing. *Journal of Marketing Research, 9*, 361-370.

Henningsen, A. (2015). Checklist for manuscripts to be submitted to scientific journals. Retrieved, 20 January, 2021 from https://files.itslearning.com/data/ku/103018/teaching/checklistmanuscri pts.pdf.

Huggins-Manley, A. C., Beal, C. R., D'Mello, S. K., Leite, W. L., Cetin-Berber, D. D., Kim, D., & McNamara, D. S. (2019). A commentary on construct validity when using operational virtual learning environment data in effectiveness studies. *Journal of Research on Educational Effectiveness*, *12*(4), 750-759.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological bulletin*, *112*(3), 527.

Kane, M. T. (2006). Validation. *Educational measurement*, *4*(2), 17-64.

Karasar, N. (2014). *Scientific research method*. Ankara: Nobel.

Kool, H., Andersson, J. A., & Giller, K. E. (2020). Reproducibility and external validity of on-farm experimental research in Africa. *Experimental Agriculture*, 1-21.

Koopman, p. (1997). How to write an abstract. Carnegie Mellon University. Retrieved 15 January, 2021 from http://www.ece.cmu.edu/~koopman/essays/abstract.html

Letts, L., Wilkins, S., Law, M., Stewart, D., Bosch, J., & Westmorland, M. (2007). Guidelines for critical review form: Qualitative studies (Version 2.0). Retrieved 20 January, 2021, from http://www.srsmcmaster.ca/Portals/20/pdf/ebp/qualguidelines_version2.0.pdf.

Lewis, C., Perry, R., & Murata, A. (2006). How should research contribute to instructional improvement? *Educational Researcher, 35*(3), 3-14

Lovejoy, T.I., Revenson, T.A., & France, C.R. (2011). Reviewing manuscripts for peer-reviewed journals: a primer for novice and seasoned reviewers. *Annals of Behavioral Medicine, 42*, 1-13.

Mahboob, A., Paltridge, B., Phakiti, A., Wagner, E., Starfield, S., Burns, A& De Costa, P. I. (2016). TESOL Quarterly research guidelines. *TESOL Quarterly*, *50*(1), 42-65.

Maxwell, J. A. (1992). Understanding and validity in qualitative research. *Harvard Educational Review*, 62(3), 279-300.

McMil1an, J. H. (2007). Randomized field trials and internal validity: Not so fast my friend. *Practical Assessment, Research, and Evaluation*, *12*(1), 15.

Mehrens, W. A. (1997). The Consequences of Consequential Validity. *Educational Measurement: Issues and Practice*, *16*(2), 16-18.

Messick, S., & Linn, R. (1989). Validity: Educational Measurement. New York: American Council on Education.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, *50*(9), 741.

Moss, P.A. (1998). The role of consequences in validity theory. *Educational Measurement*, 17 (2), 6–12.

Oesch, D. (2020). Discrimination in the hiring of older jobseekers: Combining a survey experiment with a natural experiment in Switzerland. *Research in Social Stratification and Mobility*, *65*, 100441.

Olubela, A., & Adebanjo, A. (2020). Learning Styles and Gender Effects on Secondary School Students' Learning Outcomes in Ijebu-Ode Community, Ogun State, Nigeria. *KIU Journal of Humanities*, *4*(4), 155-169.

O'Malley, M. A. (2007). Exploratory experimentation and scientific practice: Metagenomics and the proteorhodopsin case. *History and Philosophy of the Life Sciences*, 337-360.

Orquin, J. L., & Holmqvist, K. (2018). Threats to the validity of eye-movement research in psychology. *Behavior research methods*, *50*(4), 1645-1656

Park, W. W., Han, H., Hyeong Kim, S., Yoon, S., & Yu, H. (2019). An integrative review and theoretical framework of validity in qualitative research: Reflections on the Academy of Management Journal for 2000 to 2016. *Seoul Journal of Industrial Relations*, *30*, 17-39

Patino, C. M., & Ferreira, J. C. (2018). Internal and external validity: can you apply research study results to your patients?. *Jornal Brasileiro de Pneumologia*, *44*(3), 183-183.

Peter, J. P. (1981). Construct validity: A review of basic issues and marketing practices. *Journal of Marketing Research, 18*, 133-145.

Popham, W.J. (1997). Consequential validity: Right concern – wrong concept. *Educational Measurement: Issues and Practice16* (2), 9–13.

Rahimi, M., Riazi, A., & Saif, S. (2008). An investigation into the factors affecting the use of language learning strategies by Persian EFL learners. *Canadian Journal of Applied Linguistics*, *11*(2), 31-60.

Razmjoo, S. A. (2016). A Putative Evaluation Scheme for Critical Appraisal of ELT Papers. *Journal of Modern Research in English Language Studies*, *3*(3), 18-1.

Rosenthal, R. (2002). Covert communication in classrooms, clinics, courtrooms, and cubicles. *American Psychologist, 57*, 839–849.

Shadish, W. R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin

Shepard, L.A. 1993. Evaluating test validity. Review of Research in Education19: 405–50.

Siegmund, J., Siegmund, N., & Apel, S. (2015, May). Views on internal and external validity in empirical software engineering. In *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering* (Vol. 1, pp. 9-19).

Smith, G. T. (2005). On construct validity: issues of method and measurement. *Psychological assessment*, *17*(4), 396.

Steckler, A., & McLeroy, K. R. (2008). The importance of external validity.

Stobart, G. 2001. The validity of national curriculum assessment. *British Journal ofEducational Studies*, *49*(1), 26–39.

Swedberg, R. (2020). Exploratory research. *The production of knowledge: Enhancing progress in social science*, 17-41.

Taylor, S. (1994). The overprediction of fear: Is it a form of regression toward the mean? *Behaviour Research and Therapy, 32*, 753–757.

Taylor, S., & Asmundson, G. J. (2008). Internal and external validity in clinical research. *Handbook of research methods in abnormal and clinical psychology*, 23-34.

Törnkvist, B., & Henriksson, W. (2006). Validity issues concerning repeated test taking of the SweSAT. *Educational Measurement*, 2(56), 33-64.

Vosgerau, J., Scopelliti, I., & Huh, Y. E. (2020). Exerting self-control≠ sacrificing pleasure. *Journal of Consumer Psychology*, *30*(1), 181-200.

Waters, C. K. (2007). The nature and context of exploratory experimentation: An introduction to three case studies of exploratory research. *History and Philosophy of the Life Sciences*, 275-284.

Wikström, C. (2006). Classroom assessment and grading–validity issues in the process of selection to higher education. In *NCME-conference, April* (pp. 8-10).

Wilson, J. (2016, July). VII—Internal and external validity in thought experiments. In *Proceedings of the Aristotelian Society* (pp. 127-152). Oxford University Press.

Winter, G. (2000). A comparative discussion of the notion of' validity in qualitative and quantitative research. *The qualitative report*, *4*(3), 1-14.

Wolming, S. (1998). Validitet. Ett traditionellt begrepp i modern tillämpning [Validity: A modernapproach to a traditional concept]. *Pedagogisk Forskning i Sverige3* ++(2), 81–103.

Wolming, S. 1999. Validity issues in higher education selection: A Swedish example. *Studies in Educational Evaluation 25*(4), 335–51

Yu, C. H., & Ohlund, B. (2010). Threats to validity of research design. *Retrieved January*, *12*, 2012.

Zorlu, F., & Sezek, F. (2019). Effectiveness of Applying the Learning Together Method at Different Intervals in Teaching Science. *Acta Didactica Napocensia*, *12*(2), 195-208.

چکیده

اعتبار یافته های تحقیق همواره موضوعی قابل بحث در تحقیقات آموزشی به طور کلی و در تحقیقات آموزش زبان انگلیسی به طور خاص بوده است. چنین موضوع مهمی خصوصا اززمانی که بحث طراحی آزمایشی و شبه آزمایشی درتحقیقات آموزش زبان انگلیسی مطرح شده است بیشتر مورد بحث قرار گرفته است. با توجه به این نکته ما به عنوان محقق شروع به بررسی اعتبارساختار ،خارجی وداخلی۴۰ مقاله ،۲۰ مقاله از نشریات خارجی و ۲۰ مقاله از نشریات داخلی درحوزه آموزش زبان انگلیسی کردیم.نمونه ها جهت بررسی به صورت تصادفی از مجموع نشریات منتشرشده ده سال اخیر منتهی به سال۲۰۲۰ انتخاب شدند.با اتخاذ رویکردی اکتشافی به نمونه ها ,به تجزیه و تحلیل دقیق آنها از نظر کنترل تهدیدهای اعتبار داخلی،خارجی و ساختاری پرداختیم .پس از استخراج فراوانی معیارها از آزمون های آماری" کای دو "جهت تجزیه و تحلیل داده ها استفاده شده است.طبق نتایج تحقیقات, مقالات از نظر اعتبارسنچش ،خارجی وداخلی از حالت ایده آل فاصله دارند .علاوه بر این پس از مقایسه تمامی مقالات از جنبه تهدیدهای داخلی و خارجی به این نتیجه رسیدیم که ابزار سنجش و پس از آن سوگیری در انتخاب،تاریخچه،بلوغ و انتشار به خوبی کنترل شدند وتاثیر تست و تعامل انتخاب و بلوغ،رگرسیون،کاهش نمونه ها و تاثیر محقق و نمونه اهمیت کمتری داشتند.و در زمینه تهدیهای خارجی تعامل محیط،تعامل تست اولیه ،تعامل انتخاب و اثر محقق و نمونه به ترتیب از درجه کنترل پایینی برخوردار بودند. یافته های این پژوهش کاربردهایی برای دانشگاهیان نیزدارد.

کلمات کلیدی: اعتبار ساختار ، آموزش زبان انگلیسی،اعتباردلخلی،اعتبارخارجی

گروه زبان انگلیسی

پایان نامه کارشناسی ارشد آموزش زبان انگلیسی

بررسی اعتباراداخلی، اعتبار خارجی و اعتبار سازه در تحقیقات کمی در حوزه آموزش زبان انگلیسی

نگارنده:

حبیبه حکیمی

استاد راهنما

دکتر سید علی استوار نامقی

بهمن ۱۳۹۹