مجموعه کتابهای جامع تخصصی آزمون دکتری

آمـوزش زبان انگلیسـی

KEY CONCEPTS IN ELT

RESEARCH METHODS

Masoume Ahmadi
Naser Sabourian Zadeh

In The Name Of God

KEY CONCEPTS IN ELT RESEARCH METHODS

Compiled by:

Masoume Ahmadi
Naser Sabourian Zadeh

Contents

chapter **3**

Validity and Reliability

Validity

After spending a great deal of time and effort designing a study, we want to make sure that the results of our study are valid. That is, we want them to reflect what we believe they reflect and that they are meaningful in the sense that they have significance not only to the population that was tested, but, at least for most experimental research, to a broader, relevant population. The validity of a scale refers to the degree to which it measures what it is supposed to measure. Unfortunately, there is no one clear-cut indicator of a scale's validity.

The validation of a scale involves the collection of empirical evidence concerning its use. There are many types of validity, including ***content***, ***face***, ***construct***, ***criterion-related***, and ***predictive*** *validity*.

Content Validity

Content validity refers to the representativeness of our measurement regarding the phenomenon about which we want information. So, content validity refers to the adequacy with which a measure or scale has sampled from the intended universe or domain of content. If we are interested in the acquisition of relative clauses in general and plan to present learners with an acceptability judgment task, we need to make sure that all relative clause types are included. For example, if our test consists only of sentences such as *"The boy who is running is my friend,"* we do not have content validity because we have not included other relative clause types such as *"The dog that the boy loves is beautiful."* In the first sentence the relative pronoun *who* is the subject of its clause, whereas in the second sentence the relative pronoun *that* is the object. Thus, our testing instrument is not sensitive to the full range of relative clause types, and we can say that it lacks content validity.

Face Validity

Face validity is closely related to the notion of content validity and refers to the familiarity of our instrument and how easy it is to convince others that there is content validity to it. If, for example, learners are presented with reasoning tasks to carry out in an experiment and are already familiar with these sorts of tasks because they have carried them out in their classrooms, we can say that the task has face validity for the learners. Face validity thus hinges on the participants' perceptions of the research treatments and tests. If the participants do not perceive a connection between the research activities and other educational or second language activities, they may be less likely to take the experiment seriously.

Construct Validity

This is perhaps the most complex of the validity types discussed so far. Construct validity is an essential topic in second language acquisition research precisely because many of the variables investigated are not easily or directly defined. In second language research, variables such as language proficiency, aptitude, exposure to input, and linguistic representations are of interest. However, these constructs are not directly measurable in the way that height, weight, or age are. In research, construct validity refers to the degree to which the research adequately captures the construct of interest. Construct validity can be enhanced when multiple estimates of a construct are used. Construct validity involves testing a scale not against a single criterion but in terms of theoretically derived hypotheses concerning the nature of the underlying variable or construct. The construct validity is explored by investigating its relationship with other constructs both related (***convergent validity***) and unrelated (***discriminant validity***).

Criterion-Related Validity

Criterion-related validity refers to the extent to which tests used in a research study are comparable to other well-established tests of the construct in question. For example, many language programs attempt to measure global proficiency either for placement into their own program or to determine the extent to which a student might meet a particular language requirement. For the sake of convenience, these programs often develop their own internal tests, but there may be little external evidence that these tests are measuring what the programs assume they are measuring. One could measure the performance of a group of students on the local test and a well-established test (e.g., TOEFL in the case of English, or in the case of other languages, another recognized standard test). Should there be a good correlation, one can then say that the local test has been demonstrated to have criterion-related validity.

Predictive Validity

Predictive validity deals with the use that one might eventually want to make of a particular measure. Does it predict performance on some other measure?

Elaborating on the most common types of validity, we now turn to the two main types of validity that are important in conducting research: *internal validity* and *external validity*.

Internal Validity

Internal validity refers to the extent to which the results of a study are a function of the factor that the researcher intends. In other words, to what extent are the differences that have been found for the dependent variable directly related to the independent variable? A researcher must control for (i.e., rule out) all other possible factors that could potentially account for the results. For example, if we wanted to observe reaction times to a set of grammatical and ungrammatical sentences, we might devise a computer program that presents sentences on a computer screen one at a time, with learners responding to the acceptability/unacceptability of each sentence by pressing a button on the computer. To make the task easier for the participants in the study, we could tape the letter *A* for "acceptable" over the letter *t* on the keyboard and tape the letter *U* for "unacceptable" over the *y* key on the keyboard. After we have completed the study, someone might ask us if we checked for handedness of the participants. In other words, could it be the case that for those who are left handed, the *A* key ("acceptable") might be faster not because it is faster to respond to acceptable as opposed to unacceptable sentences (part of our hypothesis), but because left hands on left-handed people react faster. Our results would then have been compromised.

We would have to conclude that there was little internal validity. It is important to think through a design carefully to eliminate or at least minimize threats to internal validity. There are many ways that internal validity can be compromised, some of the most common and important of which include *participant characteristics, participant mortality* (dropout rate), *participant inattention and attitude, participant maturation, data collection* (location and collector), instrumentation, and *test effects.*

Participant Characteristics

The example provided in the previous section concerning handedness is a participant characteristic. Clearly, not all elicitation techniques will require controlling for handedness. In other words, there may be elements of the research questions and/or elicitation technique that require a careful selection of one characteristic or another. Let us consider some relevant participant characteristics for second language research: *language background, language learning experience,* and *proficiency level.*

Language Background

In many studies, researchers want to compare one group of students with another group based on different treatments. It would be important that each group of students be relatively homogeneous. Were they not homogeneous, one could not be sure about the source of the results.

Language Learning Experience

Participants come to a language learning situation with a wide range of past experiences. In some instances, these experiences may have importance for research. For example, many students in an ESL setting have had prior English instruction in their home countries, and this prior instruction may differ from one country to another.

History

Empirical research does not take place in a vacuum arid therefore we might be subject to the effects of unanticipated events while the study is in progress. Such events are outside the research study, yet they can alter the participants' performance. The best we can do at times like this is to document the impact of the events so that later we may neutralize it by using some kind of statistical control.

chapter 7

Classroom Action Research

Classroom Research

Although classrooms constitute a distinct *context* for research, many of the methodological practices and data collection techniques associated with classroom research are not unique to classroom settings, and some are also discussed elsewhere in this book. For example, we discuss diary studies as part of qualitative research methods, and in the current chapter where we focus exclusively on diary use by learners and teachers in second and foreign language classroom contexts. We begin the chapter with a discussion of the nature of classroom research.

Classroom Research Contexts

Traditionally second language researchers have distinguished between *classroom-based research* and research conducted in controlled laboratory contexts. Typical *laboratory-based research* has the advantage of allowing the researcher to tightly control the experimental variables, randomly assign subjects to treatment groups, and employ control groups—all of which are difficult, and sometimes impossible, to implement in classroom- based research contexts. Such concerns regarding classroom research have led some second language researchers to claim that although laboratory settings are more abstract, the benefits connected with being better able to control and manipulate intervening variables may be worth the potential costs of abstraction.

Whether research carried out in the laboratory can (or cannot) be generalized to the L2 classroom is an empirical question. In any case, in light of the complementary strengths and limitations of laboratory and classroom studies, second language researchers are increasingly recognizing that studies must be carried out in different contexts and that a range of different approaches must be used to gain a deeper understanding of the complexity of second language learning. Thus, whereas classroom research can enhance our understanding of how to implement effective ways of improving learners' second language skills, laboratory studies can provide more tightly controlled environments in which to test specific theories about second language development.

Combined approaches to classroom research—that is, those involving a range of different approaches, including both experimental and observational techniques—are also gaining popularity. Increasingly it appears, second language classroom researchers are calling for judicious selection and combined approaches rather than rigid adherence to one approach over another.

Common Techniques for Data Collection in Classroom Research

Observations

Observational data are common in second language research and observations are a useful means for gathering in-depth information about such phenomena as the types of language, activities, interactions, instruction, and events that occur in second and foreign language classrooms. Additionally, observations can allow the study of a behavior at close range with many important contextual variables present. Here, we focus on the particular concerns that can arise when carrying out observations in intact classrooms, as well as providing information about the different types of observation schemes that have been developed by second language classroom researchers.

Obtrusive observers

Any observer in the classroom runs the risk of being an obtrusive observer, which can be problematic for research. An obtrusive observer's presence may be felt in the classroom to the extent that the events observed cannot be said to be fully representative of the class in its typical behavior, and therefore the observation data may have limited validity. An obtrusive observer may also be problematic for the instructor and students in terms of compromising the quality of the lesson, preventing instructors from delivering the lesson to the best of their ability and, consequently, preventing the students from learning to the best of theirs.

The Hawthorne effect

The presence of observers may result in changed behavior due to the fact that those being observed feel positive about being included in a study.

Debriefing the instructor

It is also important as part of the negotiation surrounding the observation process to debrief the instructor about the research findings or the content of the observation notes or scheme. Timing is also an important consideration here. For example, researchers might provide instructors with a copy of their notes after each lesson or arrange a time to meet in order to discuss the research. By keeping the observation process as transparent and interactive as possible, researchers can often establish a more trusting and cooperative relationship with instructors Of course, in some cases, the instructors may be the focus of the research, or it may unduly influence the research if they are kept continually debriefed. In these cases, it may be preferable to make such contact after the project has been completed.

Introspective Methods in Classroom Research

Introspective methods—or data-elicitation techniques that encourage learners to communicate their internal processing and perspectives about language learning experiences—can afford researchers access to information unavailable from observational approaches. In second language research, a range of introspective methods have been employed. These methods vary with respect to the practicality of their application to classroom research. Uptake sheets, for example, described in the next section, allow researchers to investigate learners' perceptions about what they are learning. Stimulated recalls may yield insights into a learner's thought processes during learning experiences, whereas diaries can present a more comprehensive view of the learning context from a participant's viewpoint.

Uptake Sheets

One way to elicit learners' perspectives on second language classroom events is through the use of uptake sheets. Uptake sheets were initially developed as a method of data collection following Allwright's (1984a, 1984b, 1987) interest in learners' perceptions about what they learned in their language classes. He collected learners' reports about their learning, which he termed *uptake* or "whatever it is that learners get from all the language learning opportunities language lessons make available to them". In classroom research, uptake sheets are often distributed at the beginning of the lesson, and learners are asked to mark or note things on which the researcher or teacher is focusing. Whether used to uncover information about learning, noticing, attitudes, or a range of other interesting phenomena, uptake sheets can allow researchers to compare their own observations and other triangulated data with information obtained from the learners, and they create a more detailed picture of classroom events in the process.

Action Research

Definitions

Action research is basically a way of reflecting on your teaching by systematically collecting data on your everyday practice and analyzing it in order to come to some decisions about what your future practice should be. In this view, action research is a mode of inquiry undertaken by teachers and is more oriented to instructor and learner development than it is to theory building, although it can be used for the latter. Action research does not imply any particular theory or consistent methodology of research. Action research exemplifies the following features:

1. *Action research is contextual, small-scale and localized—it identifies and investigates problems within a specific situation.*
2. *It is evaluative and reflective as it aims to bring about change and improvement in practice.*
3. *It is participatory as it provides for collaborative investigation by teams of colleagues, practitioners and researchers.*
4. *Changes in practice are based on the collection of information or data which provides the impetus for change.*

There are several features of this definition that are important to highlight. First, action research, as the name implies, involves *action* in that it seeks to bring about change, specifically in local educational contexts. It is also *research* because it entails the collection and analysis of data. Finally, it is *participatory and collaborative* in that teachers work together to examine their own classrooms. The concept of action research developed out of the progressive education movement of the early 20th century when educators like John Dewey challenged the prevalent reliance on scientific research methods. Dewey believed it was essential for researchers, practitioners, and others in explained by the independent (group) variable. SPSS does not provide eta squared values for t-tests.

It can, however, be calculated using the information provided in the output. The procedure for calculating eta squared is provided below.

The formula for eta squared is as follows:

$$\text{Eta squared} = \frac{t^2}{t^2 + (N1 + N2 - 2)}$$

Replacing with the appropriate values from the example above:

$$\text{Eta squared} = \frac{1.62^2}{1.62^2 + (184 + 252 - 2)}$$

$$\text{Eta squared} = .006$$

The guidelines (proposed by Cohen, 1988) for interpreting this value are: .01=small effect, .06=moderate effect, .14=large effect. For our current example you can see that the effect size of .006 is very small.

The results of the analysis could be presented as follows:

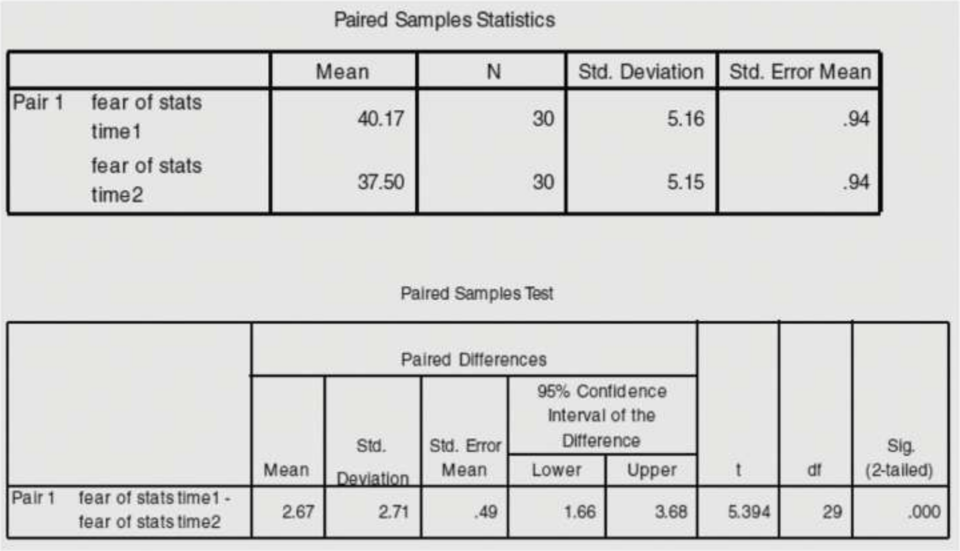
An independent-samples t-test was conducted to compare the self-esteem scores for males and females. There was no significant difference in scores for males (M=34.02, SD=4.91) and females [M=33.17, SD=5.71; t (434)=1.62, p=.11]. The magnitude of the differences in the means was very small (eta squared=.006).

Presenting the Results from Paired-samples T-test Analysis

Research Question: Is there a significant change in participants' fear of statistics scores following participation in an intervention designed to increase students' confidence in their ability to successfully complete a statistics course?

Does the intervention have an impact on participants' fear of statistics scores?

The output generated from this analysis is shown below.



Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	fear of stats time1	40.17	30	5.16	.94
	fear of stats time2	37.50	30	5.15	.94

Paired Samples Test

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	fear of stats time1 - fear of stats time2	2.67	2.71	.49	1.66	3.68	5.394	29	.000

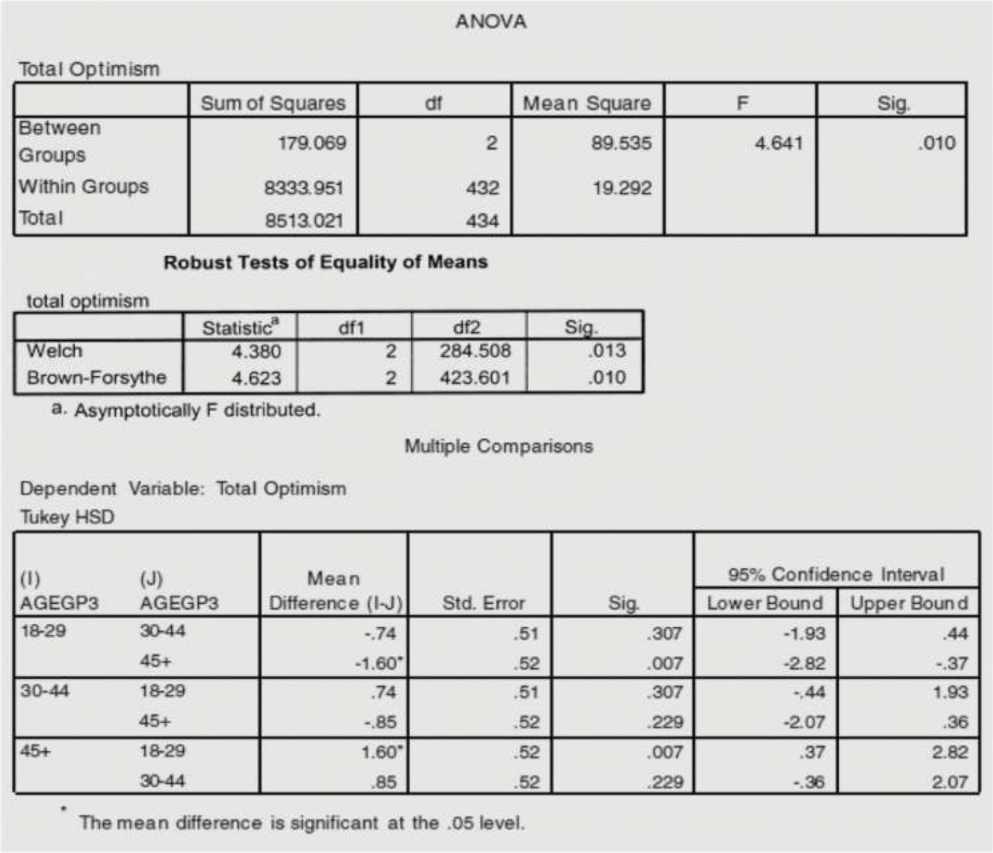
The key details that need to be presented are the name of the test, the purpose of the test, the t-value, the degrees of freedom (df), the probability value, and the means and standard deviations for each of the groups or administrations. It is also a good idea to present an effect size statistic (e.g. eta squared). The results of the analysis conducted above could be presented as follows:

A paired-samples t-test was conducted to evaluate the impact of the intervention on students' scores on the Fear of Statistics Test (FOST). There was a statistically significant decrease in FOST scores from Time 1 (M=40.17, SD=5.16) to Time 2 [M=37.5, SD=5.15, t(29)=5.39, p<.0005]. The eta squared statistic (.50) indicated a large effect size.

Presenting the Results from One-way Analysis of Variance

Research Question: Is there a difference in optimism scores for young, middle-aged and old subjects?

The output generated from this analysis is shown below.



ANOVA

Total Optimism

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	179.069	2	89.535	4.641	.010
Within Groups	8333.951	432	19.292		
Total	8513.021	434			

Robust Tests of Equality of Means

total optimism

	Statistic[a]	df1	df2	Sig.
Welch	4.380	2	284.508	.013
Brown-Forsythe	4.623	2	423.601	.010

a. Asymptotically F distributed.

Multiple Comparisons

Dependent Variable: Total Optimism

Tukey HSD

(I) AGEGP3	(J) AGEGP3	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
18-29	30-44	-.74	.51	.307	-1.93	.44
	45+	-1.60*	.52	.007	-2.82	-.37
30-44	18-29	.74	.51	.307	-.44	1.93
	45+	-.85	.52	.229	-2.07	.36
45+	18-29	1.60*	.52	.007	.37	2.82
	30-44	.85	.52	.229	-.36	2.07

* The mean difference is significant at the .05 level.

Calculating effect size for One-way Analysis of Variance

Although SPSS does not generate it for this analysis, it is possible to determine the effect size for this result (see the introduction to Part Five for a discussion on effect sizes). The information you need to calculate eta squared, one of the most common effect size statistics, is provided in the ANOVA table (a calculator would be useful here). The formula is:

$$\text{Eta squared} = \frac{\text{Sum of squares between-groups}}{\text{Total sum of squares}}$$

In this example all you need to do is to divide the Sum of squares for between-groups (179.07) by the Total sum of squares (8513.02). The resulting eta squared value is .02, which in Cohen's (1988) terms would be considered a small effect size. Cohen classifies .01 as a small effect, .06 as a medium effect and .14 as a large effect.

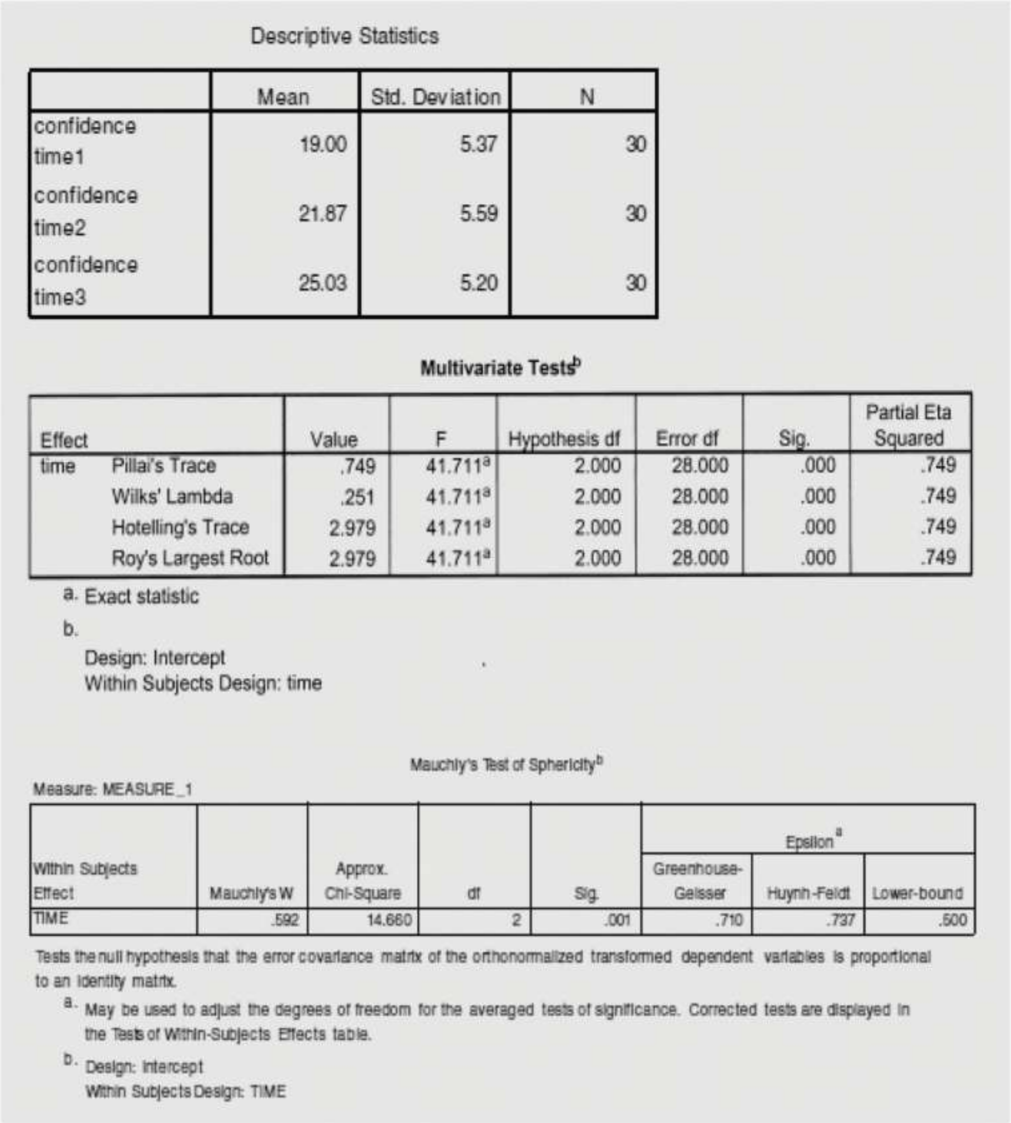
The results of the one-way between-groups analysis of variance with post-hoc tests could be presented as follows:

A one-way between-groups analysis of variance was conducted to explore the impact of age on levels of optimism, as measured by the Life Orientation test (LOT). Subjects were divided into three groups according to their age (Group 1: 29 or less; Group 2: 30 to 44; Group 3: 45 and above). There was a statistically significant difference at the p<.05 level in LOT scores for the three age groups [F(2, 432)=4.6, p=.01]. Despite reaching statistical significance, the actual difference in mean scores between the groups was quite small. The effect size, calculated using eta squared, was .02. Post-hoc comparisons using the Tukey HSD test indicated that the mean score for Group 1 (M=21.36, SD=4.55) was significantly different from Group 3 (M=22.96, SD=4.49). Group 2 (M=22.10, SD=4.15) did not differ significantly from either Group 1 or 3.

Presenting the Results from One-way Repeated Measures ANOVA

Research Question: Is there a change in confidence scores over the three time periods?

The output generated from this analysis is shown below.



Descriptive Statistics

	Mean	Std. Deviation	N
confidence time1	19.00	5.37	30
confidence time2	21.87	5.59	30
confidence time3	25.03	5.20	30

Multivariate Tests[b]

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
time	Pillai's Trace	.749	41.711[a]	2.000	28.000	.000	.749
	Wilks' Lambda	.251	41.711[a]	2.000	28.000	.000	.749
	Hotelling's Trace	2.979	41.711[a]	2.000	28.000	.000	.749
	Roy's Largest Root	2.979	41.711[a]	2.000	28.000	.000	.749

a. Exact statistic

b.

Design: Intercept
Within Subjects Design: time

Mauchly's Test of Sphericity[b]

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[a]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
TIME	.592	14.660	2	.001	.710	.737	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

b. Design: Intercept
Within Subjects Design: TIME

The results of a one-way repeated measures ANOVA could be presented as follows:

A one-way repeated measures ANOVA was conducted to compare scores on the Confidence in Coping with Statistics test at Time 1 (prior to the intervention), Time 2 (following the intervention) and Time 3 (three-month follow-up). The means and standard deviations are presented in Table XX. There was a significant effect for time [Wilks' Lambda=.25, F(2, 28)=41.17, p<.0005, multivariate partial eta squared=.75.]