Journal of Psychological and
Educational Research
JPER - 2021, 29 (1), May, 72-96

INTERPRETATION OF STATISTICAL CONCEPTS IN PSYCHOLOGY: KNOWLEDGE LEVEL AND EFFECTS OF AN INSTRUCTIONAL DESIGN IN PERUVIAN UNIVERSITY STUDENTS

Arnold Alejandro Tafur-Mendoza *

Brian Norman Peña-Calero

Connie Daniela Aliaga Guanilo

National University of San Marcos, Peru

Christian Alexis Moreau Almaraz

Francesca Cecilia Ramírez-Bontá

Jackeline Edith Garcia-Serna

National University of San Marcos, Peru

Oscar Esteban Meza-Chahuara

National University of San Marcos, Peru

Abstract

Statistics is considered a basic instrument for information analysis. Therefore, it is teaching in Psychology is of the utmost importance. However, there are difficulties in the interpretation of statistical concepts by university students. Consequently, the present study seeks to use the Psychometrics Group Instrument to compare scores obtained by a group of Psychology students attending a teaching program based on the Merrill's instructional design concerning what was found in three previous studies, and to analyse the effects produced by the teaching program for the improvement of the interpretation of statistical concepts. The participants were Psychology undergraduate students from a public university in Lima, Peru. The results indicated that the sample presents a low knowledge level in some statistical concepts, before the teaching program, similar to the three comparison investigations. On the other hand, the teaching program generated an improvement in the interpretation of statistical concepts presented in it. Based on the

Correspondence concerning this paper should be addressed to:

* BSc., National University of San Marcos, Grupo de Estudios Avances en Medición Psicológica, Peru. Address: 375 German Amezaga Street, 15081, Lima, Peru. E-mail: *aa.tafurm@up.edu.pe*

evidence found on Merrill's instructional design, it is recommended to proof each of its principles in the development of learning sessions on Statistics in social, health and behavioural sciences careers. All materials, code and data are publicly accessible via the Open Science Framework (OSF) at https://osf.io/pxbcs/.

Keywords: statistical concepts; Merrill's instructional design; psychological statistics; teaching-statistics

Introduction

Statistics has become more relevant in Psychology, being a necessary instrument for the analysis of information (Osorio, 2012). Likewise, the knowledge of statistical concepts has been considered essential for an adequate interpretation and discussion of the results and elaboration of the research findings (Ato & Vallejo, 2015; Larson-Hall & Plonsky, 2015; Repišti, 2015). Because it was found that making mistakes in the interpretation of statistical concepts leads to producing unreliable results and, consequently, to obtaining distorted conclusions (Bakker & Wicherts, 2011; Caperos, Olmos, & Pardo, 2016; Matamoros & Ceballos, 2017; Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016).

Previous studies indicated that Psychology students make mistakes in the interpretation of statistical concepts, such as hypothesis testing, *p*-value, effect size, correlation, statistical power, among others (Badenes-Ribera & Frías-Navarro, 2017; Badenes-Ribera, Frías-Navarro, & Pascual-Soler, 2015; Castro, Vanhoof, Van den Noortgate, & Onghena, 2007). Unfortunately, these mistakes have also been repeated by teachers, professionals and researchers (Badenes-Ribera, Frías-Navarro, & Bonilla-Campos, 2017a, 2017b; Badenes-Ribera, Frias-Navarro, Iotti, Bonilla-Campos, & Longobardi, 2018; Badenes-Ribera, Frías-Navarro, Iotti, Bonilla-Campos, & Longobardi, 2016; Badenes-Ribera, Frías-Navarro, Monterde-i-Bort, & Pascual-Soler, 2015; Badenes-Ribera, Frías-Navarro, Pascual-Soler, & Monterde-i-Bort, 2016).

On the other hand, in studies where the Psychometrics Group Instrument (Mittag, 1999) was used, the conclusions were similar, also finding problems in understanding the statistical tests (Gordon, 2001; Mittag & Thompson, 2000; Monterde-i-Bort, Frías-Navarro, & Pascual-Llobell, 2010). In Peru, although no studies were obtained that demonstrate this problem in psychology students, it has

been found that students of related careers showed the same difficulties when interpreting basic concepts in Statistics (Osorio, 2012; Rivera, 2010).

The difficulties for correctly interpreting statistical concepts have been attributed to various factors (students, curricula, didactic materials, among others), one of them is the inadequate teaching (Osorio, 2012; Rivera, 2010). For that reason, it is necessary to look for the most appropriate teaching strategy that guides the student to a better understanding of the conceptual contents (Rojas & Ovejero, 2014). A solution to this problem has been found in the implementation of instructional designs for the teaching of Statistics in Psychology, which provides principles based on theories of instruction and learning for the consolidation of the latter (Centeno, Gonzáles-Tablas, López, & Mateos, 2016).

Based on the above, Merrill's instructional design would allow the student to understand and interpret basic statistical concepts, due to the demonstration shown in the learning of other areas (Gardner, 2011; Mendenhall, 2012; Truong, Elen, & Clarebout, 2019). This design has been the result of a review of several models, which have five principles that, under appropriate conditions and independently of the methods and models that a theory has, they show the property of being used in theories of different approaches (Merrill, 2002, 2007, 2009). These principles are described in Figure 1 (Merrill, 2013).

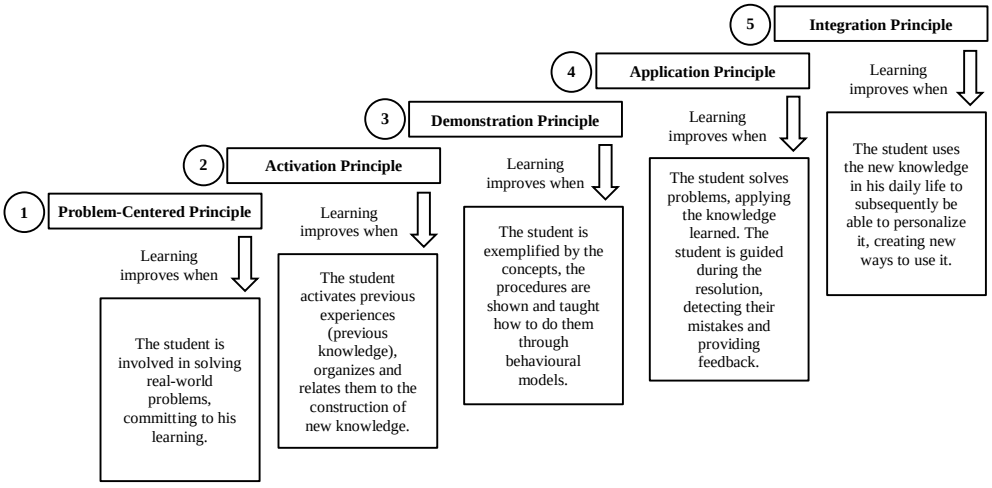

Figure 1. Merrill's instructional design principles

Otherwise, the quantitative methodology has played an important role in Psychology, currently supporting it to base its work through empirical evidence. The characteristic of this methodology is the underlying positivist epistemological paradigm, the measurement of human traits or social phenomena and the statistical analysis of quantitative data (Wang, Watts, Anderson, & Little, 2013). Consequently, the Psychology curricula of various universities have included subjects related to applied statistics, psychometrics or quantitative research methods. The basic statistical concepts for the teaching of Statistics in Psychology are detailed in Table 1.

Table 1. Definitions of basic statistical concepts

Concept	Author	Definition
Null hypothesis significance testing (NHST)	Cumming (2014), and Grissom and Kim (2005)	Procedure in which a *p*-value is calculated under the assumption that the null hypothesis is true. This value will be used to decide whether or not to reject the null hypothesis at a certain significance level, commonly .05. A statistically significant result (p <.05) and one that is not can differ slightly, other indicators need to be examined.
p-value	Altman & Krzywinski (2017), and Wasserstein and Lazar (2016)	Probability of observing statistical values of the test applied (for example, Student's *t*, ANOVA, etc.) as or more extreme than those observed, from the assumption that the null hypothesis is true.
Effect size	Castillo-Blanco and Alegre-Bravo (2015), and Cohen (1988)	Measurement of the degree to which a phenomenon studied (the relationship between variables, group differences, etc.) is presented in a population or sample of interest.
Confidence intervals	Morey, Hoekstra, Rouder, Lee, and Wagenmakers (2016)	They imply that, at a 95% level, if an infinite (or very large) number of samples were taken and confidence intervals were calculated, 95% of the intervals would contain the population parameter.
Statistical power	Bono and Arnau (1995)	It is the probability that a statistical test to reject a false null hypothesis or the probability of not committing the type II error.
General Linear Model (GLM)	American Psychological Association (2014)	A broad set of statistical techniques that describe the relationship between a dependent variable and one or more independent variables, for example, regression, variance or correlation analysis.

Table 1. Definitions of basic statistical concepts - *continued*

Concept	Author	Definition
Reliability	American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014)	The degree to which test scores for a particular sample are consistent through repeated applications. The degree to which scores are free of random measurement errors for a particular sample.
Type I error	Kirk (2008)	It implies concluding that a study supports the research hypothesis when in reality it is false. In terms of the null hypothesis, it involves rejecting this when it is true.
Type II error	Aron, Coups, and Aron (2013)	It occurs when the research hypothesis is true but the *p*-value is not so extreme as to reject the null hypothesis. In other words, it implies not rejecting this when it is false.
Stepwise analysis	Huberty (1989)	Develop a sequence of linear models and at each step, under certain criteria, add or delete an independent variable. These criteria will depend on the type of analysis (regression or discriminant), the purpose of analysis (prediction, construction of a model, etc.) and the judgment of the investigator.

Objectives

The purpose of this study is to use the Psychometrics Group Instrument (Mittag, 1999) to compare scores obtained by a group of Psychology students attending a teaching program based on the Merrill's instructional design concerning what was found in the studies of Mittag and Thompson (2000), Gordon (2001) and Monterde-i-Bort et al. (2010). Likewise, to analyse the effects produced by the teaching program for the improvement of the interpretation of statistical concepts.

Method

Participants

Sampling was non-probabilistic of an intentional type (Kerlinger & Lee, 2000). The participants were Psychology undergraduate students from a public university in Lima, Peru, attending the teaching program. The sample for the first objective was the students who completed only the pretest assessment, while, for

the second objective, the sample was composed of the students who answered the pretest and posttest.

For the first objective, the sample was 16 students, ages between 18 and 28 years (M=22.50, SD=3.12), 10 were women and were in the first to the sixth year of study. For the second objective, the sample consisted of nine students aged between 18 and 24 years (M=20.90, SD=2.26). Regarding sex, six were women and, considering the year of studies, two students were in the first year, two in second, four in third, and one in fourth.

Instruments

Inferential Statistics teaching program with R. The teaching program consisted of seven sessions, applied in seven weeks, at intervals of once a week and lasting three hours per session. The general objective of the program was to know the management of open-source R software in the application of inferential statistics in Psychology, while the specific objectives sought to allow participants to (1) learn about the concepts of inferential statistics most used in the data analysis in Psychology; (2) use R for the inferential statistics in the data analysis in Psychology; and (3) recognize the use of inferential statistics in correspondence to the type of variables with which one is working.

Psychometrics Group Instrument (Mittag, 1999). Provides scores that represent perceptions about statistical hypothesis tests and other statistical issues (Appendix A). It is made up of 29 items distributed in nine topics (perceptions): (1) general; (2) about the GLM; (3) on the stepwise analysis; (4) on the score reliability; (5) on type I and II errors; (6) on the sample size influences; (7) of the *p*-value as an effect size measure; (8) of the *p*-value as direct measures of the importance of the result; and (9) of the *p*-value as replicability evidence.

The items were answered through a five-point scale (1=disagree, 2=disagree somewhat, 3=neutral, 4=agree somewhat, and 5= agree). After the application, 14 items, whose statements are false, were recoded to invert their response scales, with the aim that the score obtained expresses the degree of correctness (instead of grade according to the statement expressed in the reagent), that is, the degree of knowledge about the subject studied.

Design

According to Ato, López, and Benavente (2013), the study was empirical research. Respecting the first objective, the strategy was an associative,

comparative type, with a cross-cultural design (XCUD). About the second objective, the strategy was manipulative, quasi-experimental type, following a pretest-posttest design (PPD), which has a single group and measures before (pretest) and after (posttest) the teaching program, using only within-subject comparisons.

Procedure

The data collection began with the request to the participants to apply the selected instrument, giving them informed consent, the content of which specified the research objective. Subsequently, the instrument was applied in an environment that had the necessary conditions to guarantee a standardized evaluation. The application was made in two moments: the first, before the start of the first session, and the second, after the seventh session. Finally, the handling of missing data was done through the pairwise method. It should be noted that, throughout the development of the study, international ethical standards were respected (American Psychological Association, 2016).

Data analysis

For the analysis of the first objective, the means of the 29 items from the pretest were calculated and the means belonging to the studies of Mittag and Thompson (2000), Gordon (2001) and Monterde-i-Bort et al. (2010) were used. Before the analysis of the second objective, 12 items were chosen whose contents were worked on the teaching program. Descriptive measures were the mean (*M*) and standard deviation (*SD*). For the comparison pretest-posttest, it was worked with non-parametric statistics, recommended for small sample size, this being the Wilcoxon signed-rank test (*W*). Respecting the effect size, the matched-pairs rank-biserial correlation was used, r_C (Kerby, 2014), considering small, medium and large effects corresponding to .10, .30, and .50, respectively (Cohen, 1988).

The analyses were performed in the R software, version 4.0.4 (R Core Team, 2021), using the packages: base, tidyverse version 1.3.0 (Wickham et al., 2019), here version 1.0.1 (Müller, 2020), psych version 2.0.12 (Revelle, 2020) and extrafont version 0.17 (Chang, 2014).

Results

The results of the first objective are presented in Figure 2. About general perceptions, in item 5, the present study and Monterde-i-Bort et al. (2010),

demonstrated a greater understanding of the relationship between statistical significance and rejection of the null hypothesis. In the remaining items, controversies regarding the use of NHST, the use of the term "statistical significance", the low statistical power of the majority of the study, and the *p*-value ban, the present research was overcome by the researches of Gordon (2001), and Mittag and Thompson (2000).

Concerning the perceptions of GLM, in item 12, the current study showed a better understanding of the correlational nature of all statistical analyses. On the other hand, item 26 showed a greater clarity on the uses of the regression in the study of Mittag and Thompson (2000).

In respect of the perceptions of the stepwise method, in items 13 and 20 it was observed that the study of Gordon (2001) had a better understanding of the problematic use of the mentioned method. In item 20, the Peruvian sample obtained the lowest average.

On the perceptions about the score reliability, items 7 and 19, provide the current definition of reliability and the utility of checking the significance of a reliability or validity coefficient, respectively, being better understood in both cases by the study of Monterde-i-Bort et al. (2010). In items 23 and 28, focused on the importance of having reliable scores, the present investigation obtained the highest averages.

About the perceptions of type I and II errors, in item 9, the current study was superior to the others in the definition of the type II error. In item 17 dealing with the definition of type I error, the study by Monterde-i-Bort et al. (2010) had a better performance. In items 22 and 29 dealing with the possibility of committing both errors, as well as the frequency of type II error in the scientific literature, the participants of the current research showed low knowledge.

In the perceptions about the sample size influences, in item 16, the present investigation showed a low knowledge of the relationship between sample size and the rejection of the null hypothesis. In item 10, on the importance of statistically significant results when the sample size is small, the study by Monterde-i-Bort et al. (2010) obtained the highest average. Otherwise, in item 25, the Peruvian sample showed a better understanding of the relationship between large sample size and obtaining statistically significant results.

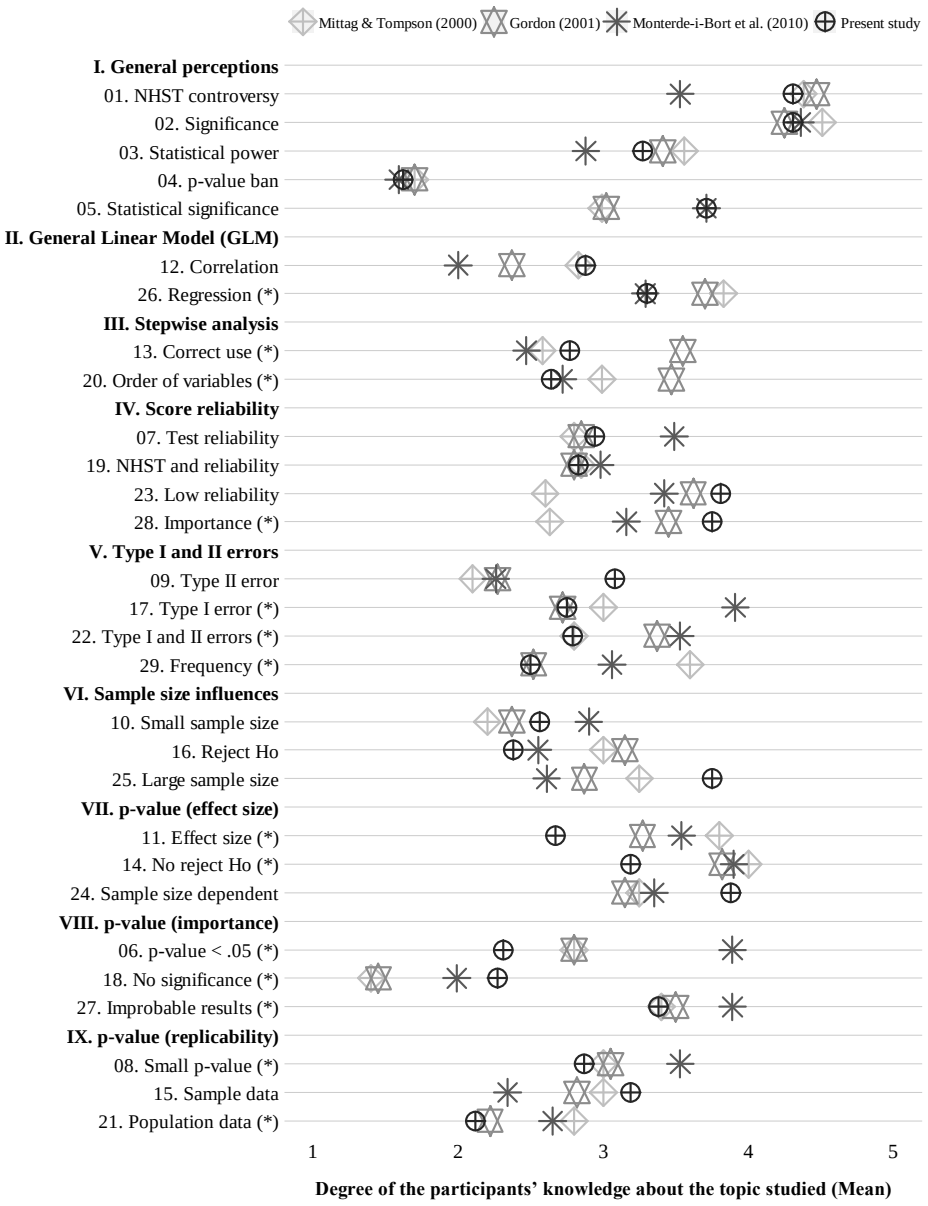

Figure 2. Comparison between the studies of Mittag and Thompson (2000), Gordon (2001), Monterde-i-Bort et al. (2010), and the present study. (*) Items considered false where the response scale was reversed

Finally, respecting the perception of the *p*-value as replicability evidence, in item 8, the sample of Monterde-i-Bort et al. (2010) was more clear that the size of the *p*-value does not influence the replication of results in future studies. In item 15, the present investigation obtained the highest average, considering that, the *p*-values obtained in a study imply the probability that the results will occur in the sample but not in the population. However, this affirmation probably generated confusion when believing that the test of significance predicts the probability of replicating the results of a sample to the population, so the Peruvian sample obtained the lowest average in item 21.

About the second objective, the results are presented in Table 2. The effect size was small (r_C>.10) in most cases. Nevertheless, in item 22 the effect size was large (r_C>.50), where the posttest average was higher than the pretest. Only in items 5, 9, and 18, the means difference between the pretest-posttest was favourable to the first group, in the remaining items the highest averages belonged to the posttest. On the other hand, none of the items found a statistically significant difference (*p*-value <.05).

Table 2. Statistical analysis of differences in pretest-posttest

		Pretest		Posttest				
Item	*n*	*M*	*SD*	*M*	*SD*	*W*	*p*	r_C
01. NHST controversy.	9	4.22	0.83	4.56	0.73	6.00	.149	.133
02. Significance.	9	4.11	1.36	4.89	0.33	10.00	.089	.222
04. *p*-value ban.	9	1.78	0.67	2.44	1.13	17.00	.202	.289
05. Statistical significance.	7	3.43	1.27	2.86	1.46	7.00	.526	.250
06. p-value < .05.	9	2.67	1.32	3.11	1.69	22.00	.621	.178
08. Small p-value.	7	3.00	0.82	3.43	1.27	10.50	.490	.214
09. Type II error.	7	3.00	1.00	2.71	1.60	5.00	.572	.179
11. Effect size.	8	2.62	0.92	3.12	1.55	14.00	.518	.194
14. No reject Ho.	9	3.33	1.22	3.67	1.32	13.50	.595	.133
17. Type I error.	7	2.57	0.79	3.29	1.70	14.50	.457	.286
18. No significance.	8	2.25	0.71	1.88	0.99	3.00	.233	.250
22. Type I and II errors.	8	2.88	0.99	4.00	1.60	25.50	.058	.639

Note: In italics are the items considered false (the response scale was reversed).

Finally, Figure 3 shows the average scores (pretest-posttest) of the 12 items that were related to the topics developed in the teaching program. In nine items the means in the posttest exceeded those of the pretest.

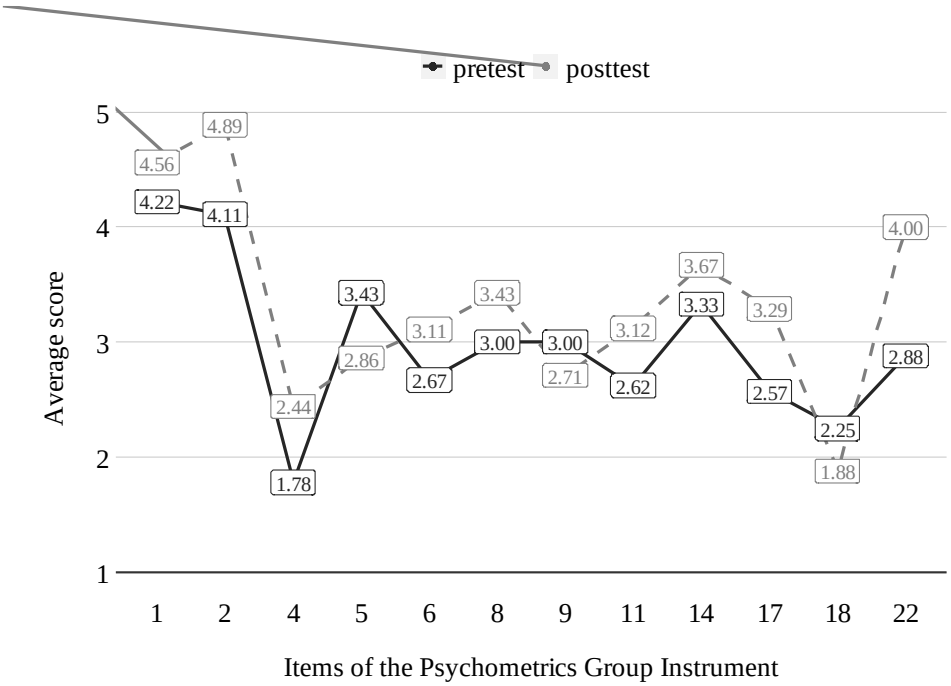

Figure 3. Average scores in the pretest-posttest

Discussion

The objectives of this study were to use the Psychometrics Group Instrument to compare scores obtained by a group of Psychology students attending a teaching program based on the Merrill's instructional design concerning what was found in the studies of Mittag and Thompson (2000), Gordon (2001) and Monterde-i-Bort et al. (2010), and to analyse the effects produced by the teaching program for the improvement of the interpretation of statistical concepts. The Peruvian sample presented better results in at least one item in eight topics of the instrument (general perceptions, GLM, score reliability, type I and II errors, sample size influences, *p*-value as an effect size measure, as direct measures of the importance of the result and as replicability evidence). However, low scores were observed in the items belonging to the perception of stepwise analysis.

In respect of the comparison of averages, the sample of Mittag and Thompson (2000) was made up of members of the American Educational Research Association (AERA), the research of Gordon (2001) by members of the American Vocational Education Research Association (AVERA), and the study of Monterde-i-Bort et al. (2010) by psychologists, teachers and researchers, from Spanish universities. On the other hand, the Peruvian sample was made up of Psychology undergraduate students, however, the latter obtained a higher score in at least one item in eight topics.

The first topic covered general perceptions, focused mainly on the NHST, where this study with the research of Monterde-i-Bort et al. (2010) obtained the highest average in the item describing its operation. Being an item that demands basic theoretical knowledge, the superiority of the current research score could be influenced by previous learning of students in the university or other spaces. Likewise, low averages of the professionals are likely based on the predominance or majority (incorrect) use of the NHST.

Concerning perceptions of the GLM, opposite behaviour is observed between the two items that make up the GLM. While the Peruvian sample has the highest score to correlation, it also has one of the lowest scores to regression. In this sense, the contents of courses on applied statistics in psychological research tend to focus their hypotheses and analyses on a correlational framework, and rarely explain or develop the logic of regression analysis. In Psychology, more than 80% of thesis can be correlational (Mamani, 2018).

In the topic that deals with score reliability, the Peruvian sample achieved a score above the rest of the studies in two indicators. This can be explained by the increase and dissemination of documents focused on reliability. Given this, the term “reliability” was searched in the Web of Science (WOS) database between 1900 and 2018, where 15,552 results were obtained in the Psychology area. Respecting the number of publications per year, in the year 2000, 543 publications were found; in 2001, there were 606; in 2006, 800; and in 2018, 1801 were found, which implies a greater amount of bibliography available to the reader.

On the topic of type I and II errors, the highest score in the Peruvian sample corresponds to an item that refers exclusively to the possibility of making the type II error. Although this score differs between 0.81 and 0.98 points from previous research, it is close to a neutral position in terms of the degree of agreement with the statement. In recent years there has been greater dissemination of the concepts

associated with NHST, compared to 10 or 20 years ago (Trafimow & Marks, 2015).

About the topic on the sample size influences, the present research obtained the highest average in the item referred to the possibility of predicting the sample size from the results of a statistical hypothesis test. Answering this question implies not only knowing theoretical aspects about sample size but also about the NHST, being one of the concepts in which the Peruvian sample denotes greater dominance. Concerning the average obtained by the other investigations, it is observed that, in all three items, the averages tend to be neutral, indicating a lack of knowledge or a tendency not to question the results provided by the NHST.

In the topic that addresses the perception of the *p*-value as an effect size measure, the Peruvian sample understands that *p*-values of different investigations cannot be directly compared because they depend on the sample sizes used. This item could be adequately answered based on theoretical knowledge about both the NHST and the influence of the sample size. When searching for the word "*p*-value" in WOS, it produced 24 results for the Psychology area. Regarding the years of publication, one publication was found in 2000, two were found in 2001, no publications were registered in 2006, and six results were found in 2018. It is important to highlight that in 2017, 10 publications were found.

Regarding the *p*-value as direct measures of the importance of the result, a higher score was observed in the Peruvian sample than in previous research on the importance of non-statistically significant studies. This may imply that over the years, the severity or requirement of a necessarily positive *p*-value in NHST has decreased. Despite this, the score of 2.27 points reached to the Peruvian students implies an average degree of disagreement with the indicated statement, being taken as controversial by the participants.

Concerning the perception of the *p*-value as replicability evidence, the present study understood better that, the values measure only the probability that the results will occur in a sample but not in the population. However, it is observed that the four studies tend to obtain neutral scores, indicating a lack of *p*-value about replicability, despite the existence of various documents, and in different years, that criticize these beliefs (Amrhein, Trafimow, & Greenland, 2019; Cohen, 1990, 1994; Pascual-Llobell, García, & Frías-Navarro, 2000; Verdam, Oort, & Sprangers, 2014).

On the other hand, respect of the results of the pretest and posttest, no statistically significant difference was found in any of the items, given this, the

relationship between the sample size and *p*-value should be considered, since in small samples the latter will be affected (Spence & Stanley, 2018). Besides, type II error increases, resulting in a decrease in statistical power which implies a lower probability of revealing differences or effects (Cumming et al., 2007).

Therefore, it is necessary to resort to other indicators before concluding that an effect or difference does not occur, justifying only a *p*-value (Altman & Krzywinski, 2017; Amrhein, Greenland, & McShane, 2019; Wasserstein & Lazar, 2016). Among the indicators that provide additional information for adequate statistical interpretation are the effect size and exploratory graphs (Finch et al., 2004; Kirk, 2001; Trafimow & Marks, 2015). In this way, the effect size report warns of the presence of changes in the indicators in favour of the development of the teaching program, with small effects on eight indicators and a large effect on item 22, referring to the conceptions of type I and II errors.

The positive changes found in the posttest (75%) assume that the development of the teaching program under the Merrill's instructional design generated favourable changes in the learning of the participants, translated into an increase in the information and interpretation of the basic statistical terms represented in the topics specified above. However, three indicators show a negative change, higher pretest score. These three indicators are framed within an interpretive response, that is, for the person to adequately answer the question posed, it is necessary to use the information learned and interpret it in a specific situation.

The disagreement presented can be understood under the premise that a successful response of the indicators represents a greater complexity in the learning process. This process may be affected if the student does not have sufficient knowledge of the subject to which they were exposed, and therefore cannot satisfactorily assimilate it (Acharya, 2017). Similarly, understanding the usefulness and potential of learning in the teaching of statistical aspects seems to improve student performance, and therefore their learning (Acee & Weinstein, 2010).

A global assessment of the indicators reported allows affirming that the majority changes support the effectiveness, although partial, of the teaching program. According to Merrill, Li, and Jones (1991), people achieve a better knowledge and registration of information from the use and integration of mental representations of thematic content, in structures that interrelate with other structures. These new representations are interrelation of previous and new

information, leaving aside isolated representations. Therefore, the integration principle happens progressively and at a different rate in each student.

Based on the evidence found on the Merrill's instructional design, it is recommended to proof each of its principles in the development of learning sessions on Statistics in social, health and behavioural sciences careers. Likewise, understanding and criticism of statistical methods should be promoted, avoiding teaching only formulas or calculations.

Therefore, publications related to the use of Statistics in Psychology must include in their content the present situation and the debate about different statistical procedures and concepts. Currently, there are good examples that incorporate the recommendations indicated (Cassidy, Dimova, Giguère, Spence, & Stanley, 2019; Funder & Ozer, 2019; Greenland et al., 2016; Makin & de Xivry, 2019; Sarafoglou, Hoogeveen, Matzke, & Wagenmakers, 2019; Trafimow, 2019; Wilkinson, 1999).

Among the limitations of the study, it is observed that the sample size used affects the representativeness of the results and the research design does not allow definitive conclusions regarding the effectiveness of the instructional design. It is recommended to use a larger sample size in future replications, appropriate to have sufficient statistical power and achieve a correct generalization of the results (Perugini, Gallucci, & Costantini, 2018).

Additionally, it is suggested to use an experimental research design, where the participants are randomly selected. Thus, attributing direct and causal changes to the development of the teaching program. On the other hand, it is appropriate to replicate the research in other universities, not only in Lima but also in other cities of Peru, as well as the review of the degree to which Psychology students present erroneous interpretations of various statistical concepts or methods.

This research shows the problem of low level of understanding in statistical concepts considered basic in several sciences including Psychology. The theoretical contribution of this study lies in the presentation and discussion basic statistical concepts in the teaching of Statistics in Psychology, useful to be reviewed by students, researchers and teachers. Regarding practical implications, the effectiveness of Merrill's instructional design was presented as a didactic strategy to solve the problem studied. Likewise, it is important to indicate that this study needs to be replicated, for which purpose the materials of the teaching program were consigned and thus facilitate future research to address these issues and that professionals can contribute to the teaching of statistical concepts.

Conclusions

The present study constitutes the first approach to this type of topic in the Peruvian context and should be interpreted with the pertinent considerations of the methodological characteristics presented. The findings indicated that, in the study sample, at least one item in eight of the nine existing topics presented a higher score than those reported in studies conducted in the United States and Spain. However, the rest of the indicators (18 of 27) presented a lower score compared to the studies. Regarding the post-application measurements of the teaching program, it was observed that, despite not showing statistically significant differences, 75% of the items (nine of 12) showed differences in favour of the posttest (higher scores). Specifically, eight of the items presented a small magnitude difference, while item 22 (about type I and II errors) showed a large magnitude difference.

The results allow concluding that the sample of students showed a low level of knowledge in some statistical concepts related to general perceptions, GLM, stepwise analysis, score reliability, type I and II errors, the sample size influences, *p*-value as an effect size measure, direct measures of the importance of the results, and replicability evidence. The persistence of these deficiencies over time and in different spaces has as a possible cause the teaching of statistics at the undergraduate level. Therefore, the use of instructional designs with empirical evidence is necessary to improve the understanding of statistical concepts in psychology education. In this sense, the use of the teaching program based on Merrill's instructional design achieved a significant improvement in nine of the 12 indicators considered for evaluation. Finally, it is important to mention that this study is pending replication.

References

Acee, T. W., & Weinstein, C. E. (2010). Effects of a value-reappraisal intervention on statistics students' motivation and performance. *The Journal of Experimental Education*, *78*(4), 487-512.

Acharya, B. R. (2017). Factors affecting difficulties in learning mathematics by mathematics learners. *International Journal of Elementary Education*, *6*(2), 8-15. https://doi.org/10.11648/j.ijeedu.20170602.11

Altman, N. S., & Krzywinski, M. (2017). Points of significance: Interpreting p values. *Nature Methods*, *14*(3), 213-214. https://doi.org/10.1038/nmeth.4210

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Psychological Association. (2014). *APA dictionary of statistics and research methods*. (S. Zedeck, Ed.). Washington, DC: American Psychological Association. https://doi.org/10.1037/14336-000

American Psychological Association. (2016). Revision of Ethical Standard 3.04 of the "Ethical Principles of Psychologists and Code of Conduct" (2002, as amended 2010). *American Psychologist*, *71*(9), 900. https://doi.org/10.1037/amp0000102

Amrhein, V., Greenland, S., & McShane, B. B. (2019). Statistical significance gives bias a free pass. *European Journal of Clinical Investigation*, *49*(12), e13176. https://doi.org/10.1111/eci.13176

Amrhein, V., Trafimow, D., & Greenland, S. (2019). Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician*, *73*(sup1), 262-270.

Aron, A., Coups, E. J., & Aron, E. N. (2013). *Statistics for Psychology* (6th ed.). Upper Saddle River, NJ: Pearson.

Ato, M., López, J. J., & Benavente, A. (2013). A classification system for research designs in psychology. *Anales de Psicología*, *29*(3), 1038-1059. https://doi.org/10.6018/analesps.29.3.178511

Ato, M., & Vallejo, G. (2015). *Diseños de investigación en Psicología*. Madrid: Pirámide.

Badenes-Ribera, L., & Frías-Navarro, D. (2017). Falacias sobre el valor p compartidas por profesores y estudiantes universitarios. *Universitas Psychologica*, *16*(3), 1-10. https://doi.org/10.11144/Javeriana.upsy16-3.fvcp

Badenes-Ribera, L., Frías-Navarro, D., & Bonilla-Campos, A. (2017a). Errores de interpretación de los valores p entre psicólogos profesionales españoles: un estudio exploratorio. *International Journal of Developmental and Educational Psychology*, *2*(1), 551-560.

Badenes-Ribera, L., Frías-Navarro, D., & Bonilla-Campos, A. (2017b). Un estudio exploratorio sobre el nivel de conocimiento sobre el tamaño del efecto y meta-análisis en psicólogos profesionales españoles. *European Journal of Investigation in Health, Psychology and Education*, *7*(2), 111-122. https://doi.org/10.30552/ejihpe.v7i2.200

Badenes-Ribera, L., Frías-Navarro, D., Iotti, B., Bonilla-Campos, A., & Longobardi, C. (2016). Misconceptions of the p-value among Chilean and Italian academic psychologists. *Frontiers in Psychology, 7*, 1247. https://doi.org/10.3389/fpsyg.2016.01247

Badenes-Ribera, L., Frias-Navarro, D., Iotti, N. O., Bonilla-Campos, A., & Longobardi, C. (2018). Perceived statistical knowledge level and self-reported statistical practice among academic psychologists. *Frontiers in Psychology*, *9*, 996. https://doi.org/10.3389/fpsyg.2018.00996

Badenes-Ribera, L., Frías-Navarro, D., Monterde-i-Bort, H., & Pascual-Soler, M. (2015). Interpretation of the p value: A national survey study in academic psychologists from Spain. *Psicothema, 27*(3), 290-295.

Badenes-Ribera, L., Frías-Navarro, D., & Pascual-Soler, M. (2015). Errors d'interpretació dels valors p en estudiants universitaris de psicologia. *Anuari de Psicologia*, *16*(2), 15-31.

Badenes-Ribera, L., Frías-Navarro, D., Pascual-Soler, M., & Monterde-i-Bort, H. (2016). Knowledge level of effect size statistics, confidence intervals and meta-analysis in Spanish academic psychologists. *Psicothema*, *28*(4), 448-456. https://doi.org/10.7334/psicothema2016.24

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*(3), 666–678. https://doi.org/10.3758/s13428-011-0089-5

Bono, R., & Arnau, J. (1995). Consideraciones generales en torno a los estudios de potencia. *Anales de Psicología*, *11*(2), 193-202.

Caperos, J. M., Olmos, R., & Pardo, A. (2016). Incosistencies in reported p-values in Spanish journals of psychology: The case of correlation coefficients. *Methodology*, *12*(2), 44-51. https://doi.org/10.1027/1614-2241/a000107

Cassidy, S. A., Dimova, R., Giguère, B., Spence, J. R., & Stanley, D. J. (2019). Failing grade: 89% of introduction-to-Psychology textbooks that define or explain statistical significance do so incorrectly. *Advances in Methods and Practices in Psychological Science*, *2*(3), 233-239.

Castillo-Blanco, R., & Alegre-Bravo, A. (2015). Importancia del tamaño del efecto en el análisis de datos de investigación en psicología. *Persona*, *18*, 137-148. https://doi.org/10.26439/persona2015.n018.503

Castro, A. E., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review*,

2(2), 98-113. https://doi.org/10.1016/j.edurev.2007.04.001

Centeno, A. V., Gonzáles-Tablas, M., López, M. E. E., & Mateos, P. M. (2016). Una experiencia de aprendizaje combinado en Estadística para estudiantes de Psicología usando la evaluación como herramienta de aprendizaje. *Education in the Knowledge Society, 17*(1), 65-85.

Chang, W. (2014). *Extrafont: Tools for using fonts*. Retrieved from https://cran.r-project.org/package=extrafont

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

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*(12), 1304-1312. https://doi.org/10.1037/0003-066X.45.12.1304

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

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7-29. https://doi.org/10.1177/0956797613504966

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., … Wilson, S. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science*, *18*(3), 230-232.

Finch, S., Cumming, G., Williams, J., Palmer, L., Griffith, E., Alders, C., … Goodman, O. (2004). Reform of statistical inference in psychology: The case of Memory & Cognition. *Behavior Research Methods, Instruments, & Computers, 36*(2), 312-324. https://doi.org/10.3758/BF03195577

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156-168.

Gardner, J. L. (2011). *Testing the efficacy of Merrill's first principles of instruction in improving student performance in introductory biology courses* (Utah State University). Retrieved from https://digitalcommons.usu.edu/etd/885

Gordon, H. R. D. (2001). American Vocational Education Research Association members' perceptions of statistical significance tests and other statistical controversies. *Journal of Vocational Education Research, 26*(2), 244-271. https://doi.org/10.5328/JVER26.2.244

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, p values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology, 31*(4), 337-350. https://doi.org/10.1007/s10654-016-0149-3

Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Huberty, C. J. (1989). Problems with stepwise methods-better alternatives. *Advances in Social Science Methodology, 1*, 43-70.

Kerby, D. S. (2014). The simple difference formula: An approach to teaching nonparametric correlation. *Comprehensive Psychology*, *3*(1), 1-9. https://doi.org/10.2466/11.IT.3.1

Kerlinger, F. N., & Lee, H. B. (2000). *Foundations of behavioral research* (4th ed.). Fort Worth, TX: Harcourt College Publishers.

Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement, 61*(2), 213-218. https://doi.org/10.1177/00131640121971185

Kirk, R. E. (2008). *Statistics: An introduction* (5th ed.). Belmont, CA: Thomson Wadsworth.

Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning*, *65*(S1), 127-159. https://doi.org/10.1111/lang.12115

Makin, T. R., & de Xivry, J. J. O. (2019). Ten common statistical mistakes to watch out for when writing or reviewing a manuscript. *ELife*, *8*, 1-13. https://doi.org/10.7554/eLife.48175

Mamani, O. J. (2018). Methodological quality and characteristics of the undergraduate psychology theses of a private university of Peru. *Propositos y Representaciones*, *6*(2), 321-338. https://doi.org/10.20511/pyr2018.v6n2.224

Matamoros, R. A., & Ceballos, A. (2017). Errores conceptuales de estadística más comunes en publicaciones científicas. *Revista CES Medicina Veterinaria y Zootecnia*, *12*(3), 211-229. https://doi.org/10.21615/cesmvz.12.3.4

Mendenhall, A. M. (2012). *Examining the use of first principles of instruction by instructional designers in a short-term, high volume, rapid production of online K-12 teacher professional development modules* (Florida State University). Retrieved from http://purl.flvc.org/fsu/fd/FSU_migr_etd-5402

Merrill, M. D. (2002). First principles of instruction. *Educational Technology Research and Development, 50*(3), 43-59.

Merrill, M. D. (2007). First principles of instruction: A synthesis. In R. A. Reiser & J. V. Dempsey (Eds.), *Trends and issues in instructional design and technology* (2nd ed., pp. 62-71). Upper Saddle River, NJ: Prentice Hall.

Merrill, M. D. (2009). First principles of instruction. In C. M. Reigeluth & A. A.

Carr-Chellman (Eds.), *Instructional-design theories and models: Building a common knowledge base* (Vol. 3, pp. 41-56). New York, NY: Routledge.

Merrill, M. D. (2013). *First principles of instruction: Identifying and designing effective, efficient, and engaging instruction*. San Francisco, CA: Pfeiffer.

Merrill, M. D., Li, Z., & Jones, M. K. (1991). Instructional transaction Theory: An Introduction. *Educational Technology, 31*(6), 7-12.

Mittag, K. C. (1999). *The psychometrics group instrument: Attitudes about contemporary statistical controversies*. Texas, TX: University of Texas at San Antonio.

Mittag, K. C., & Thompson, B. (2000). A National survey of AERA members' perceptions of statistcal significane tests and other statistical issues. *Educational Researcher, 29*(4), 14-20. https://doi.org/10.2307/1176454

Monterde-i-Bort, H., Frías-Navarro, D., & Pascual-Llobell, J. (2010). Uses and abuses of statistical significance tests and other statistical resources: A comparative study. *European Journal of Psychology of Education*, *25*(4), 429-447. https://doi.org/10.1007/s10212-010-0021-x

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103-123.

Müller, K. (2020). *Here: A simpler way to find your files*. Retrieved from https://cran.r-project.org/package=here

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985-2013). *Behavior Research Methods*, *48*(4), 1205-1226. https://doi.org/10.3758/s13428-015-0664-2

Osorio, A. R. (2012). *Análisis de la idoneidad de un proceso de instrucción para la introducción del concepto de probabilidad en la enseñanza superior* (Pontificia Universidad Católica del Perú). Retrieved from http://hdl.handle.net/20.500.12404/4658

Pascual-Llobell, J., García, J. F., & Frías-Navarro, D. (2000). Significación estadística, importancia del efecto y replicabilidad de los datos. *Psicothema, 12*(S2), 408-412.

Perugini, M., Gallucci, M., & Costantini, G. (2018). A practical primer to power analysis for simple experimental designs. *International Review of Social Psychology*, *31*(1), 1-23. https://doi.org/10.5334/irsp.181

R Core Team. (2021). *R: A language and environment for statistical computing*.

Vienna: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org

Repišti, S. (2015). Some common mistakes of data analysis, their interpretation, and presentation in biomedical sciences. *Istraživanje Matematičkog Obrazovanja (IMO)*, *7*(12), 37-46.

Revelle, W. (2020). *Psych: Procedures for psychological, psychometric, and personality research*. Evanston, IL: Northwestern University. Retrieved from https://cran.r-project.org/package=psych

Rivera, L. M. (2010). El aprendizaje experiencial de la estadística en base a los estilos de aprendizaje del estudiante universitario. *UCV-SCIENTIA*, *2*(2), 111-117.

Rojas, I. R., & Ovejero, D. (2014). Errores cometidos por los alumnos en la asignatura estadística y biometría, de la carrera de ingeniería agronómica, Universidad Nacional de Catamarca (2012). *Biología En Agronomía*, *4*(1), 156-167.

Sarafoglou, A., Hoogeveen, S., Matzke, D., & Wagenmakers, E. J. (2019). Teaching good research practices: Protocol of a research master course. *Psychology Learning and Teaching*. https://doi.org/10.1177/1475725719858807

Spence, J. R., & Stanley, D. J. (2018). Concise, simple, and not wrong: In search of a short-hand interpretation of statistical significance. *Frontiers in Psychology*, *9*, 2185. https://doi.org/10.3389/fpsyg.2018.02185

Trafimow, D. (2019). A frequentist alternative to significance testing, p-values, and confidence intervals. *Econometrics*, *7*, 26. https: // doi.org / 10.3390 / econometrics7020026

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*(1), 1-2. https://doi.org/10.1080/01973533.2015.1012991

Truong, M. T., Elen, J., & Clarebout, G. (2019). Implementing Merrill's first principles of instruction: Practice and identification. *Journal of Educational and Instructional Studies in the World*, *9*(2), 14-28.

Verdam, M. G. E., Oort, F. J., & Sprangers, M. A. G. (2014). Significance, truth and proof of p values: Reminders about common misconceptions regarding null hypothesis significance testing. *Quality of Life Research*, *23*(1), 5-7. https://doi.org/10.1007/s11136-013-0437-2

Wang, L. L., Watts, A. S., Anderson, R. A., & Little, T. D. (2013). Common fallacies in quantitative research methodology. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods* (Vol. 2, pp. 718-758). New York,

NY: Oxford University Press.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129-133. https://doi.org/10.1080/00031305.2016.1154108

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., … Woo, K. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594-604. https://doi.org/10.1037/0003-066X.54.8.594

Received September 1, 2020
Revision February 4, 2021
Accepted February 22, 2021

Appendix A

Psychometrics Group Instrument (Mittag, 1999).

01 Controversies regarding the use of significance tests have existed for many years in the past, and will doubtless continue for many years in the future.
02 It would be better if everyone used the phrase, "statistically significant," rather than "significant", to describe the results when the null hypothesis is rejected.
03 Most studies are conducted with insufficient statistical power against Type II error.
04 Science would progress more rapidly if tests of significance were banned from journal articles.
05 All that significance means is that the researcher rejected the null hypothesis.
06 *Finding that $p < .05$ is one indication that the results are important.*
07 On its face, the statement, "the reliability of the test," asserts an untruth, since reliability is not a characteristic of a given test.
08 *Smaller and smaller values for the calculated p indicate that the results are more and more likely to be replicated in future research.*
09 A Type II error is impossible if the results are statistically significant.
10 Statistically significant results are more noteworthy when sample sizes are small.
11 *Smaller p values provide direct evidence that study effects were larger.*
12 All statistical analyses (e.g., *t*-tests, ANOVA, *r, R*) are correlational.
13 *In regression and other analyses, stepwise analyses can reasonably be used to identify the best subset of predictors of a given subset size.*
14 *If a dozen different researchers investigated the same phenomenon using the same null hypothesis, and none of the studies yielded statistically significant results, this means that the effects being investigated were not noteworthy or important.*
15 The p values that are calculated in a given study test the probability of the results occurring in the sample, and not the probability of results occurring in the population.
16 Every null hypothesis will eventually be rejected at some sample size.
17 *Type I errors may be a concern when the null hypothesis is not rejected.*
18 *Studies with non-significant results can still be very important.*
19 Testing the significance of a reliability or a validity coefficient with a null hypothesis that $r^2=0$ is not useful or productive.
20 *When researchers do stepwise analyses, the order of the entry of the variables (1st, 2nd, etc.) provides one useful indication of the importance of the variables.*
21 *Significance tests evaluate the probability that the results for the sample are the same in the population.*
22 *It is possible to make both a Type I and a Type II error in a given study.*
23 Poor reliability of data in a given study will tend to lower or attenuate the effect sizes that are detected.
24 The values reported in different studies cannot be readily compared, because these values are confounded with the different sample sizes across studies.

25 Significance tests are partly a test of whether the researcher had a large sample.
26 *It is not possible to use regression to statistically test the null that means of different groups are equal.*
27 *Unlikely results are generally more important or noteworthy.*
28 *Reliability does not directly affect the likelihood of obtaining significance in a given study.*
29 *Type II errors are probably fairly common within published research.*

Note: In italics are the items considered false.