Citar como:

Tafur-Mendoza, A. A., Peña-Calero, B. N., Aliaga, C. D., Moreau, C. A., Ramírez-Bontá, F. C., Garcia-Serna, J. E., & Meza-Chahuara, O. E. (2021). Interpretation of statistical concepts in Psychology: Knowledge level and effects of an instructional design in Peruvian university students. *Journal of Psychological and Educational Research*, *29*(1), 72-96.

Interpretación de conceptos estadísticos en Psicología: nivel de conocimientos y efectos de un diseño instruccional en estudiantes universitarios peruanos

Resumen

La estadística es considerada un instrumento básico para el análisis de la información. Por tanto, su enseñanza en Psicología es de suma importancia. No obstante, se evidencian dificultades en la interpretación de conceptos estadísticos por parte de estudiantes universitarios. Por ello, el presente estudio busca utilizar el Psychometrics Group Instrument para comparar las puntuaciones obtenidas por un grupo de estudiantes de Psicología asistentes a un programa de enseñanza basado en el diseño instruccional de Merrill con respecto a lo hallado en tres estudios anteriores, y analizar los efectos producidos por el programa de enseñanza para la mejora de la interpretación de conceptos estadísticos. Los participantes fueron estudiantes de pregrado de Psicología de una universidad pública en Lima, Perú. Los resultados indicaron que, la muestra presenta un nivel bajo de conocimientos en algunos conceptos estadísticos, antes del programa de enseñanza, similar a las tres investigaciones de comparación. Por otro lado, el programa de enseñanza generó una mejora en la interpretación de los conceptos estadísticos presentados en este. A partir de la evidencia encontrada sobre el diseño instruccional de Merrill, se recomienda poner a prueba cada uno de sus principios en el desarrollo de sesiones de aprendizaje sobre Estadística en carreras de ciencias sociales, de la salud y del comportamiento. Todos los materiales, códigos y datos son de acceso público a través del Open Science Framework (OSF) en https://osf.io/pxbcs/.

Palabras clave: conceptos estadísticos; diseño instruccional de Merrill; Estadística en Psicología; enseñanza de la Estadística

Introducción

La Estadística ha cobrado mayor relevancia en Psicología, siendo un instrumento necesario para el análisis de información (Osorio, 2012). Asimismo, se ha considerado fundamental el conocimiento de conceptos estadísticos para una adecuada interpretación y discusión de los resultados y elaboración de los hallazgos de la investigación (Ato & Vallejo, 2015; Larson-Hall & Plonsky, 2015; Repišti, 2015). De esta manera, se ha encontrado que cometer errores en la interpretación de conceptos estadísticos conlleva a producir resultados no confiables y, por consiguiente, obtener conclusiones distorsionadas (Bakker & Wicherts, 2011; Caperos, Olmos, & Pardo, 2016; Matamoros & Ceballos, 2017; Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2016).

Estudios previos señalaron que, los estudiantes de Psicología cometen errores en la interpretación de conceptos estadísticos, tales como la prueba de hipótesis, valor p, tamaño del efecto, correlación, potencia estadística, entre otros (Badenes-Ribera & Frías-Navarro, 2017; Badenes-Ribera, Frías-Navarro, & Pascual-Soler, 2015; Castro, Vanhoof, Van den Noortgate, & Onghena, 2007). Desafortunadamente, estos errores también han sido repetidos por docentes, profesionales e investigadores (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).

Por otro lado, en estudios donde se empleó el Psychometrics Group Instrument (Mittag, 1999), las conclusiones fueron similares, encontrando además problemas en la comprensión de las pruebas estadísticas (Gordon, 2001; Mittag & Thompson, 2000; Monterde-i-Bort, Frías-Navarro, & Pascual-Llobell, 2010). En Perú, si bien no se han obtenido estudios que evidencien esta problemática en estudiantes de Psicología, se ha encontrado que, estudiantes de carreras afines presentan las mismas dificultades al momento de interpretar conceptos básicos en Estadística (Osorio, 2012; Rivera, 2010).

Las dificultades para interpretar correctamente conceptos estadísticos han sido atribuidas a diversos factores (estudiantes, currículo, materiales instruccionales, entre otros), siendo uno de ellos la inadecuada enseñanza (Osorio, 2012; Rivera, 2010). Por ello, es necesario buscar la estrategia didáctica más adecuada que guíe al estudiante a una mejor comprensión de contenidos conceptuales (Rojas & Ovejero, 2014). Una solución a esta problemática ha sido hallada en la incorporación de diseños

instruccionales para la enseñanza de la Estadística en Psicología, que brinda principios basados en teorías de la instrucción y del aprendizaje para la consolidación de este último (Centeno, Gonzáles-Tablas, López, & Mateos, 2016).

Con base en lo anterior, el diseño instruccional de Merrill permitiría al estudiante comprender e interpretar adecuadamente conceptos estadísticos básicos, debido a la efectividad mostrada en el aprendizaje de otras áreas (Gardner, 2011; Mendenhall, 2012; Truong, Elen, & Clarebout, 2019). Este diseño ha sido el resultado de una revisión de diversos modelos que cuenta con cinco principios que, bajo apropiadas condiciones e independientemente de los métodos y modelos que posee una teoría, tienen la propiedad de ser usados en teorías de distintos enfoques (Merrill, 2002, 2007, 2009). En la Figura 1 se describen estos principios (Merrill, 2013).

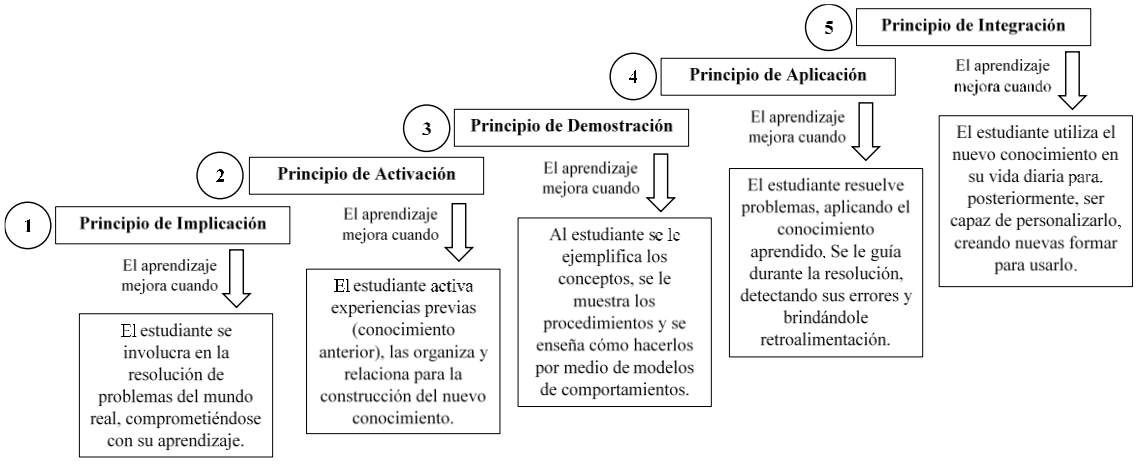

Figura 1. Principios del diseño instruccional de Merrill.

Por otro lado, la metodología cuantitativa ha jugado un rol importante en Psicología, actualmente apoyando a esta a fundamentar su trabajo mediante evidencias empíricas. Lo característico de esta metodología es el paradigma epistemológico positivista subyacente, la medición de rasgos humanos o fenómenos sociales y el análisis estadístico de datos cuantitativos (Wang, Watts, Anderson, & Little, 2013). Por ello, los currículos de Psicología de diversas universidades han incluido materias relacionadas a Estadística aplicada, psicometría o métodos de investigación cuantitativa. Los conceptos estadísticos considerados básicos para la enseñanza de la Estadística en Psicología son detallados en la Tabla 1.

Tabla 1

Definiciones de conceptos estadísticos básicos

Concepto	Autor	Definición
Test de Significancia de la Hipótesis Nula (NHST)	Cumming (2014), y Grissom y Kim (2005)	Procedimiento en el que se calcula un valor p bajo el supuesto de que la hipótesis nula es verdadera. Este valor se usará para decidir si se rechaza o no la hipótesis nula a un determinado nivel de significancia, comúnmente .05. Un resultado estadísticamente significativo ($p < .05$) y uno que no lo es pueden diferir poco, se necesita examinar otros indicadores.
Valor *p*	Altman y Krzywinski (2017) y Wasserstein y Lazar (2016)	Probabilidad de observar valores estadísticos de la prueba aplicada (por ejemplo, t de Student, ANOVA, etc.) tan o más extremos que los observados, desde la suposición de que la hipótesis nula es verdadera.
Tamaño del efecto	Castillo-Blanco y Alegre-Bravo (2015) y Cohen (1988)	Medida del grado en que un fenómeno estudiado (relación entre variables, diferencias de grupos, etc.) se presenta en una población o muestra de interés.
Intervalos de confianza	Morey, Hoekstra, Rouder, Lee y Wagenmakers (2016)	Implican que, a un nivel del 95%, si se tomara un número infinito (o muy numeroso) de muestras y se calcularan los intervalos de confianza, el 95% de los intervalos contendrían el parámetro poblacional.
Poder estadístico	Bono y Arnau (1995)	Es la probabilidad que utiliza una prueba estadística para rechazar una hipótesis nula falsa o la probabilidad de no cometer el error de Tipo II.
Modelo Lineal General (MLG)	American Psychological Association (2014)	Conjunto amplio de técnicas estadísticas que describen la relación entre una variable dependiente y una o más variables independientes, por ejemplo, análisis de regresión, varianza o correlación.
Fiabilidad	American Educational Research Association, American Psychological Association y National Council on Measurement in Education (2014)	Grado en que las puntuaciones de un test para una muestra particular son consistentes a través de aplicaciones repetidas. Grado en que los puntajes están libres de errores aleatorios de medición para una muestra particular.
Error Tipo I	Kirk (2008)	Implica concluir que un estudio apoya la hipótesis de investigación cuando en realidad esta es falsa. En términos de la hipótesis nula, involucra rechazar esta cuando en realidad es verdadera.
Error Tipo II	Aron, Coups y Aron (2013)	Se produce cuando la hipótesis de investigación es verdadera pero el valor p no es tan extremo como para rechazar la hipótesis nula. En otras palabras, implica no rechazar esta cuando en realidad es falsa.
Análisis paso a paso	Huberty (1989)	Desarrolla una secuencia de modelos lineales y en cada paso, bajo ciertos criterios, agrega o elimina una variable independiente. Estos criterios dependerán del tipo de análisis (regresión o discriminante), propósito de análisis (predicción, construcción de un modelo, etc.) y del juicio del investigador.

Objetivos del estudio

La propuesta de este estudio es utilizar el Psychometrics Group Instrument (Mittag, 1999) para comparar las puntuaciones obtenidas por un grupo de estudiantes de Psicología asistentes a un programa de enseñanza basado en el diseño instruccional de Merrill con respecto a lo hallado en los estudios de Mittag y Thompson (2000), Gordon (2001) y Monterde-i-Bort et al. (2010). Asimismo, analizar los efectos producidos por el programa de enseñanza para la mejora de la interpretación de conceptos estadísticos.

Método

Participantes

El muestreo fue no probabilístico de tipo intencional (Kerlinger & Lee, 2000). Los participantes fueron estudiantes de pregrado de Psicología de una universidad pública en Lima, Perú, asistentes al programa de enseñanza. La muestra para el primer objetivo fueron los estudiantes que completaron únicamente la evaluación pretest, en tanto que, para el segundo objetivo, la muestra estuvo compuesta por los estudiantes que contestaron el pretest y postest.

Para el primer objetivo, la muestra fue 16 estudiantes, con edades entre 18 y 28 años (M = 22.50, DE = 3.12), 10 fueron mujeres y cursaban del primer al sexto año de estudios. Para el segundo objetivo, la muestra estuvo conformada por nueve estudiantes con edades entre 18 y 24 años (M = 20.90, DE = 2.26). Respecto al sexo, seis fueron mujeres y, considerando el año de estudios, dos estudiantes fueron de primer año, dos de segundo, cuatro de tercero, y uno de cuarto.

Instrumentos

Programa de enseñanza de Estadística inferencial con R

El programa de enseñanza estuvo constituido por siete sesiones, aplicadas durante siete semanas, en intervalos de una vez por semana y con una duración de tres horas por sesión. El objetivo general del programa fue conocer el manejo del software libre R en la aplicación de la Estadística inferencial en Psicología, en tanto que, los objetivos específicos buscaron que los participantes logren (1) conocer los conceptos de la Estadística inferencial más empleados en el análisis de datos en Psicología; (2) utilizar R para el empleo de la estadística inferencial en el análisis de datos en Psicología; y (3) reconocer el uso de los estadísticos inferenciales en correspondencia al tipo de variables con las que se está trabajando.

Psychometrics Group Instrument (Mittag, 1999).

Proporciona puntuaciones que representan las percepciones sobre las pruebas de significancia estadística y otras cuestiones estadísticas (Anexo A). Está conformado por 29 ítems distribuidos en nueve tópicos (percepciones): (1) generales; (2) sobre el MLG; (3) sobre los procedimientos de análisis paso a paso; (4) sobre la fiabilidad de las puntuaciones; (5) sobre errores tipo I y tipo II; (6) sobre la influencia del tamaño de muestra; (7) del valor p como medida del tamaño del efecto; (8) del valor p como medida de la importancia del resultado; y (9) del valor p como evidencia de replicabilidad.

Los ítems fueron respondidos a través de una escala de cinco puntos (1 = desacuerdo, 2 = algo en desacuerdo, 3 = neutral, 4 = algo de acuerdo, y 5 = acuerdo). Luego de la aplicación, 14 ítems, cuyas afirmaciones son falsas, fueron recodificados para invertir sus escalas de respuesta, con el objetivo de que la puntuación obtenida exprese el grado de acierto (en lugar de grado de acuerdo con la afirmación expresada en el reactivo), es decir, el grado de conocimientos sobre el tema estudiado.

Diseño

Según Ato, López y Benavente (2013) el estudio fue una investigación empírica. Respecto al primer objetivo, la estrategia fue asociativa, de tipo comparativo, con un diseño transcultural (DTC). En relación con el segundo objetivo, la estrategia fue manipulativa, de tipo cuasiexperimental, siguiendo un diseño pretest-postest (DPP), que cuenta con un grupo único y con medidas antes (pretest) y después (postest) del programa de enseñanza, empleando solamente comparaciones intrasujetos.

Procedimiento

La recolección de datos inició con la solicitud a los participantes para aplicar el instrumento seleccionado, brindándoles el consentimiento informado, cuyo contenido especificaba el objetivo del estudio. Posteriormente, se aplicó el instrumento en un ambiente que contó con las condiciones necesarias para garantizar una evaluación estandarizada. La aplicación se realizó en dos momentos: el primero, antes del inicio de la primera sesión, y el segundo, después de la séptima sesión. Finalmente, el manejo de datos faltantes en la base de datos se realizó a través del método pairwise. Es preciso señalar que, durante todo el desarrollo del estudio, se respetaron las normas éticas internacionales (American Psychological Association, 2016).

Análisis de datos

Para el análisis estadístico del primer objetivo, se calcularon los promedios de los 29 ítems en la evaluación pretest y se tomaron las medias de los estudios de Mittag y Thompson (2000), Gordon (2001) y Monterde-i-Bort et al. (2010). Previo al análisis del segundo objetivo, se eligieron 12 ítems cuyos contenidos fueron trabajados en el programa de enseñanza. Las medidas descriptivas fueron la media (*M*) y desviación estándar (*DE*). Para la comparación entre el pretest y postest, se trabajó con estadística no paramétrica, recomendable cuando el tamaño de muestra es pequeño, siendo esta la prueba de rangos con signos de Wilcoxon (*W*). Respecto al tamaño del efecto, se empleó el coeficiente de correlación rango biserial para muestras pareadas, r_C (Kerby, 2014), considerándose efectos pequeño, mediano y grande los que corresponden a .10, .30 y .50, respectivamente (Cohen, 1988).

Los análisis se realizaron en el software R, versión 4.0.4 (R Core Team, 2021), empleando los paquetes: base, tidyverse versión 1.3.0 (Wickham et al., 2019), here versión 1.0.1 (Müller, 2020), psych versión 2.0.12 (Revelle, 2020) y extrafont versión 0.17 (Chang, 2014).

Resultados

En la Figura 2 se presentan los resultados del primer objetivo. Respecto a las percepciones generales, en el ítem 5, el estudio actual y el de Monterde-i-Bort et al. (2010), demostraron una mayor comprensión de la relación entre la significancia estadística y el rechazo de la hipótesis nula. En los ítems restantes, referentes a la controversia del uso del NHST, el empleo del término “significancia estadística”, el bajo poder estadístico de la mayoría de estudio y la prohibición del valor p, el presente estudio fue superado por las investigaciones de Gordon (2001) y Mittag y Thompson (2000).

En cuanto a las percepciones sobre el MLG, en el ítem 12, el presente estudio mostró una mejor comprensión sobre la naturaleza correlacional de todos los análisis estadísticos. Por otro lado, en el ítem 26 se evidenció una mayor claridad sobre el uso de la regresión en el estudio de Mittag y Thompson (2000).

En relación con las percepciones sobre el método paso a paso, en los ítems 13 y 20 se observó que, el estudio de Gordon (2001) tuvo una mejor comprensión sobre el uso problemático del método mencionado. En el ítem 20, la muestra peruana obtuvo el promedio más bajo.

Sobre las percepciones acerca de la fiabilidad de las puntuaciones, los ítems 7 y 19, brindan la definición actual de fiabilidad y la utilidad de comprobar la significancia de un coeficiente de fiabilidad o validez, respectivamente, siendo mejor comprendidos en ambos casos por el estudio de Monterde-i-Bort et al. (2010). En los ítems 23 y 28, enfocados en la importancia de tener puntuaciones fiables, el presente estudio obtuvo los promedios más altos.

Respecto a las percepciones sobre el error tipo I y II, en el ítem 9, el estudio actual fue superior a los demás respecto a la definición del error tipo II. En el ítem 17 que trata sobre la definición del error tipo I, el estudio de Monterde-i-Bort et al. (2010) tuvo el mejor rendimiento. En los ítems 22 y 29 que tratan sobre la posibilidad de cometer ambos errores, así como la frecuencia del error tipo II en la literatura científica, los participantes del actual estudio mostraron un bajo conocimiento.

En las percepciones sobre la influencia del tamaño de muestra, en el ítem 16, el estudio actual demostró un bajo conocimiento de la relación entre tamaño de muestra y el rechazo de la hipótesis nula. En el ítem 10, sobre la importancia de los resultados estadísticamente significativos cuando la muestra es pequeña, el estudio de Monterde-i-Bort et al. (2010) obtuvo la media más alta. Por otro lado, en el ítem 25, la muestra peruana mostró un mejor entendimiento de la relación entre un tamaño de muestra grande y la obtención de resultados estadísticamente significativos.

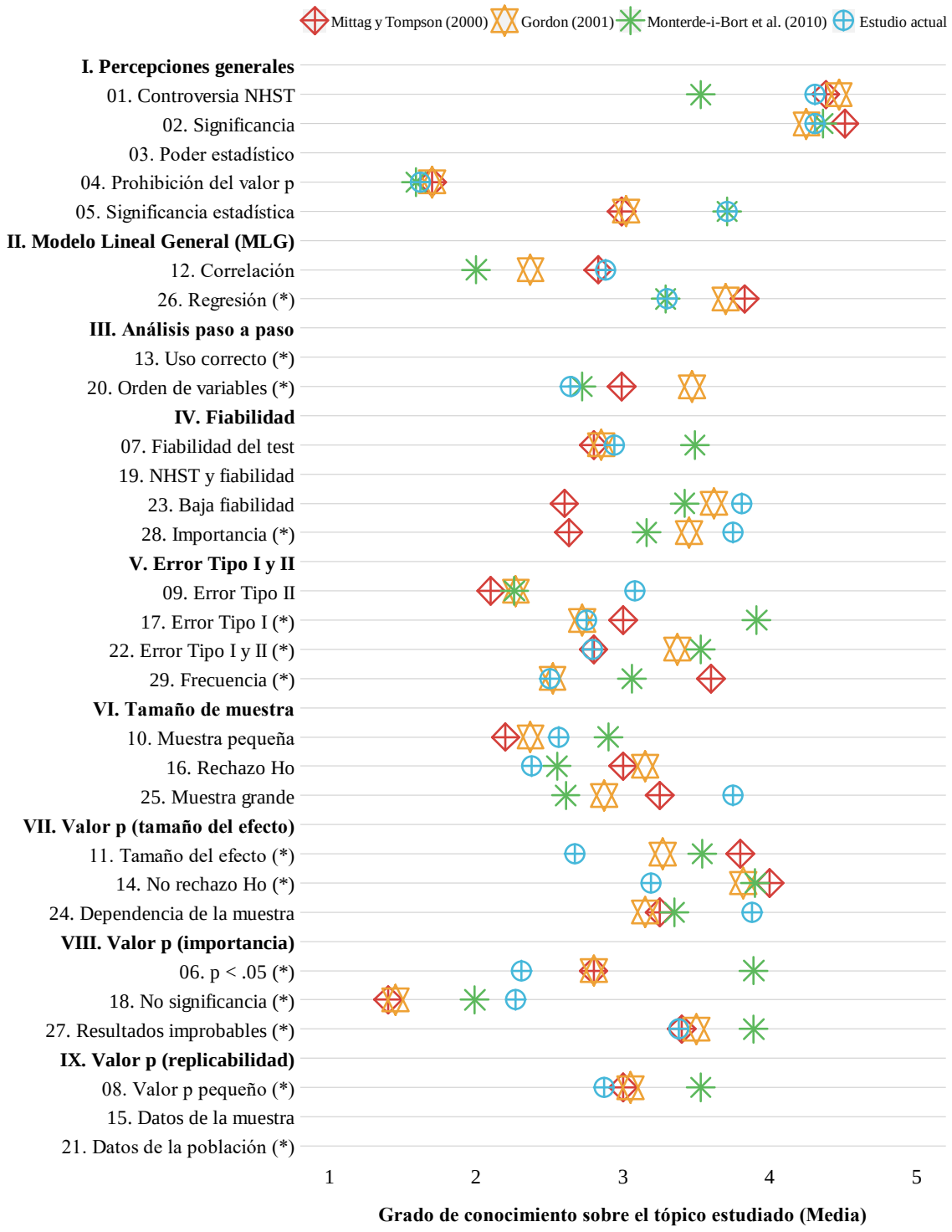

Figura 2. Comparación entre los estudios de Mittag y Thompson (2000), Gordon (2001), Monterde-i-Bort et al. (2010) y el estudio actual. (*) Ítems considerados falsos donde la escala de respuesta fue invertida.

Por último, en cuanto a la percepción del valor p como evidencia de replicabilidad, en el ítem 8, la muestra de Monterde-i-Bort et al. (2010) tuvo mayor claridad acerca de que el tamaño del valor p no influye en la replicación de resultados

en estudios futuros. En el ítem 15, la presente investigación obtuvo el promedio más alto, considerando que, los valores p obtenidos en un estudio implican la probabilidad de que los resultados ocurran en la muestra más no en la población. Sin embargo, probablemente dicha afirmación generó una confusión al creer que la prueba de significación predice la probabilidad de replicar los resultados de una muestra a la población por lo que, la muestra peruana obtuvo el promedio más bajo en el ítem 21.

En relación al segundo objetivo, los resultados son presentados en la Tabla 2. El tamaño del efecto fue pequeño (r_C > .10) en la mayoría de los casos. Sin embargo, en el ítem 22 el tamaño del efecto fue grande (r_C > .50), donde el promedio en el postest fue superior al pretest. Únicamente en los ítems 5, 9 y 18, la diferencia entre las medias del pretest y postest fueron favorables al primer grupo, en los restantes ítems los promedios más altos pertenecieron al postest. Por otro lado, en ninguno de los ítems se encontró una diferencia estadísticamente significativa (p < .05).

Tabla 2

Análisis estadístico de las diferencias en el pretest y postest

		Pretest		Postest				
Ítem	*n*	*M*	*DE*	*M*	*DE*	*W*	*p*	r_C
01. Controversia NHST.	9	4.22	0.83	4.56	0.73	6.00	.149	.133
02. Significancia.	9	4.11	1.36	4.89	0.33	10.00	.089	.222
04. Prohibición del valor p.	9	1.78	0.67	2.44	1.13	17.00	.202	.289
05. Significancia estadística.	7	3.43	1.27	2.86	1.46	7.00	.526	.250
06. p < .05.	9	2.67	1.32	3.11	1.69	22.00	.621	.178
08. Valor p pequeño.	7	3.00	0.82	3.43	1.27	10.50	.490	.214
09. Error Tipo II.	7	3.00	1.00	2.71	1.60	5.00	.572	.179
11. Tamaño del efecto.	8	2.62	0.92	3.12	1.55	14.00	.518	.194
14. No rechazo Ho.	9	3.33	1.22	3.67	1.32	13.50	.595	.133
17. Error Tipo I.	7	2.57	0.79	3.29	1.70	14.50	.457	.286
18. No significancia.	8	2.25	0.71	1.88	0.99	3.00	.233	.250
22. Error Tipo I y II.	8	2.88	0.99	4.00	1.60	25.50	.058	.639

Nota: En cursiva se encuentran los ítems considerados falsos (la escala de respuesta fue invertida).

Finalmente, en la Figura 3 se muestran los puntajes promedios (pretest-postest) de los 12 ítems que se encontraban en relación con los tópicos desarrollados en el programa de enseñanza. En nueve ítems las medias en el postest superaron a las del pretest.

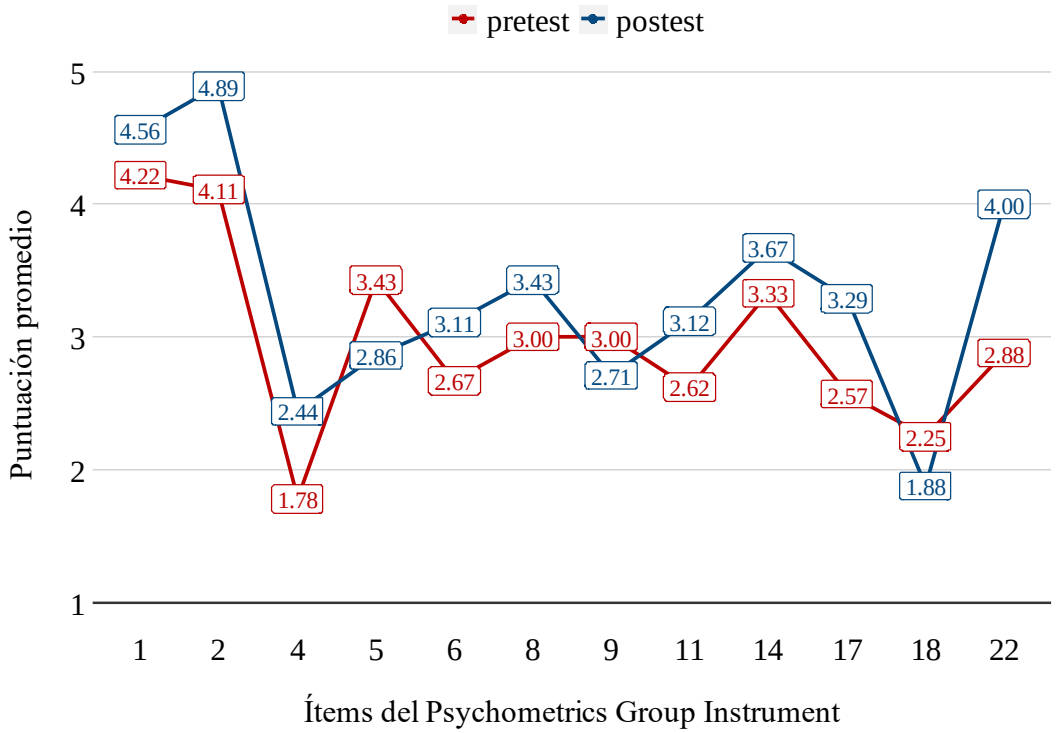

Figura 3. Puntuación promedio en el pretest y postest.

Discusión

Los objetivos del estudio fueron utilizar el Psychometrics Group Instrument para comparar las puntuaciones obtenidas por un grupo de estudiantes de Psicología asistentes a un programa de enseñanza basado en el diseño instruccional de Merrill con respecto a lo hallado en los estudios de Mittag y Thompson (2000), Gordon (2001) y Monterde-i-Bort et al. (2010) y analizar los efectos producidos por el programa de enseñanza para la mejora de la interpretación de conceptos estadísticos. La muestra peruana presentó mejores resultados en al menos un ítem en ocho tópicos del instrumento (percepciones generales, MLG, fiabilidad de las puntuaciones, errores tipo I y II, influencia del tamaño de muestra, valor p como medida del tamaño del efecto, como medida de la importancia del resultado y como evidencia de replicabilidad). No obstante, se observaron puntajes bajos en los ítems pertenecientes a la percepción del análisis paso a paso.

Respecto a la comparación de promedios, la muestra de Mittag y Thompson (2000), estuvo conformada por miembros de la American Educational Research Association (AERA), la investigación de Gordon (2001) por miembros de la American Vocational Education Research Association (AVERA) y el estudio de Monterde-i-Bort

et al. (2010) por psicólogos, docentes e investigadores, de universidades españolas. Por otro lado, la muestra peruana estuvo conformada por estudiantes de Psicología de pregrado, sin embargo, estos últimos obtuvieron un mayor puntaje en al menos un ítem en ocho tópicos.

El primer tópico abarcó sobre percepciones generales, enfocado mayormente en el NHST, donde el presente estudio junto con la investigación de Monterde-i-Bort et al. (2010) obtuvieron el mayor promedio en el ítem que describe su funcionamiento. Al ser un ítem que demanda conocimiento teórico básico, la superioridad del puntaje de la investigación actual pudo estar influenciado por aprendizaje previo de los estudiantes en la universidad u otros espacios. Asimismo, los promedios bajos de los profesionales son probablemente basados en el predominio o uso mayoritario (incorrecto) del NHST.

En cuanto a las percepciones del MLG, se observa un comportamiento opuesto entre los dos ítems que lo componen. Mientras que la muestra peruana tiene la puntuación más alta en correlación, también tiene una de las puntuaciones más bajas en regresión. En este sentido, los contenidos de los cursos de estadística aplicada en la investigación psicológica tienden a centrar sus hipótesis y análisis en un marco correlacional, y rara vez explican o desarrollan la lógica del análisis de regresión. En Psicología, más del 80% de las tesis pueden ser correlacionales (Mamani, 2018).

En el tópico que aborda la fiabilidad de las puntuaciones, la muestra peruana logró un puntaje por encima del resto de estudios en dos indicadores. Esto se puede explicar por el incremento y difusión de documentos enfocados en la fiabilidad. Ante ello, una búsqueda en la base de datos Web of Science (WOS) del término “reliability” entre los años 1900 y 2018, donde se obtuvieron 15,552 resultados en el área de Psicología. Respecto a la cantidad de publicaciones por año, en el año 2000 se encontraron 543 publicaciones; en 2001, fueron 606; en 2006, 800; y en 2018, se hallaron 1801, que implica una mayor cantidad de bibliografía disponible para el lector.

En el tópico de los errores de tipo I y II, la puntuación más alta en la muestra peruana corresponde a un ítem que se refiere exclusivamente a la posibilidad de cometer el error de tipo II. Si bien este puntaje difiere entre 0.81 y 0.98 puntos de los estudios anteriores, se acerca a una posición neutral en cuanto al grado de acuerdo con la afirmación. En los últimos años se ha producido una mayor difusión de los conceptos asociados al NHST, en comparación con hace 10 o 20 años (Trafimow & Marks, 2015).

Respecto al tópico sobre la influencia del tamaño de muestra, el presente estudio obtuvo el mayor promedio en el ítem referido a la posibilidad de predecir el tamaño de

la muestra a partir de los resultados de una prueba de significación estadística. Responder esta pregunta implica no solamente conocer aspectos teóricos sobre tamaño de muestra, sino también sobre el NHST, siendo uno de los conceptos en el que la muestra peruana denota un mayor dominio. Respecto al promedio obtenido por las demás investigaciones, se observa que, en los tres ítems, las medias tienden a la categoría neutra, indicando una falta de conocimiento o una tendencia a no cuestionar los resultados brindados por el NHST.

En el tópico que aborda la percepción del valor p como medida del tamaño del efecto, la muestra peruana comprende que, valores p de diferentes investigaciones no se pueden comparar directamente porque estos dependen de los tamaños de muestra utilizados. Este ítem se podría responder adecuadamente a partir de un conocimiento teórico tanto sobre el NHST como de la influencia del tamaño de muestra. Al buscar la palabra "p-value" en WOS, este arrojó 24 resultados para el área de Psicología. En cuanto a los años de publicación, en 2000 se encontró una publicación, en 2001 se encontraron dos, en 2006 no se registraron publicaciones, y en 2018 se hallaron seis resultados. Es importante resaltar que en 2017 se encontraron 10 publicaciones.

En cuanto al valor p como medida de la importancia del resultado, se observó una puntuación más alta en la muestra peruana que en las investigaciones anteriores sobre la importancia de los estudios no estadísticamente significativos. Esto puede implicar que, con el paso de los años, ha disminuido la importancia o la exigencia de un valor p necesariamente positivo en la NHST. A pesar de ello, la puntuación de 2.27 puntos alcanzada por los estudiantes peruanos implica un grado medio de desacuerdo con la afirmación indicada, siendo tomada como controvertida por los participantes.

En el tópico sobre la percepción del valor p como evidencia de replicabilidad, el presente estudio comprendió mejor que, los valores p miden únicamente la probabilidad de que los resultados ocurran en una muestra mas no en la población. No obstante, se observa que, los cuatro estudios tienden a obtener puntuaciones neutras, indicando un desconocimiento del valor p en relación a la replicabilidad, a pesar de la existencia de diversos documentos, y en diferentes años, que critican estas creencias (Amrhein, Trafimow, & Greenland, 2019; Cohen, 1990, 1994; Pascual-Llobell, García, & Frías-Navarro, 2000; Verdam, Oort, & Sprangers, 2014).

Por otro lado, respecto a los resultados del pretest y postest, en ninguno de los ítems se encontró una diferencia estadísticamente significativa, ante esto, se debe considerar la relación entre el tamaño de muestra y valor p, ya que en muestras

pequeñas este último se verá afectado (Spence & Stanley, 2018). Además, el error tipo II se incrementa, produciéndose un decremento en el poder estadístico lo que implica una menor probabilidad de revelar diferencias o efectos (Cumming et al., 2007).

Por ello, es necesario recurrir a otros indicadores antes de concluir que un efecto o diferencia no se produce, justificándose únicamente en un valor p (Altman & Krzywinski, 2017; Amrhein, Greenland, & McShane, 2019; Wasserstein & Lazar, 2016). Entre los indicadores que otorgan información adicional para una adecuada interpretación estadística se encuentran, el tamaño del efecto y los gráficos exploratorios (Finch et al., 2004; Kirk, 2001; Trafimow & Marks, 2015). De esta forma, el reporte del tamaño del efecto advierte la presencia de cambios en los indicadores a favor del desarrollo del programa de enseñanza, con efectos pequeños en ocho indicadores y un efecto grande en el ítem 22, referido a las concepciones sobre el error tipo I y II.

Los cambios positivos encontrados en la evaluación postest (75%) suponen que el desarrollo del programa de enseñanza bajo el diseño instruccional de Merrill generó cambios favorables en el aprendizaje de los participantes, traducido en un incremento de la información e interpretación de los términos estadísticos básicos representados en los tópicos especificados anteriormente. Sin embargo, tres indicadores presentan un cambio negativo, mayor puntaje en el pretest. Estos tres indicadores se encuentran enmarcados dentro de una respuesta interpretativa, es decir, para que la persona pueda responder adecuadamente a la pregunta planteada es necesario que utilice la información aprendida y lo interprete en una situación específica.

La discordancia presentada, puede entenderse bajo la premisa de que una respuesta acertada de los indicadores representa una mayor complejidad en el proceso de aprendizaje. Este proceso puede verse afectado si es que el estudiante no tiene conocimiento suficiente de la temática a la que se encuentra expuesto, y por consiguiente no puede asimilarlo de forma satisfactoria (Acharya, 2017). De igual manera, entender la utilidad y potencialidad del aprendizaje en la enseñanza de aspectos estadísticos parece mejorar el desempeño de los estudiantes, y por consiguiente su aprendizaje (Acee & Weinstein, 2010).

Una valoración global de los indicadores reportados permite afirmar que, los cambios mayoritarios apoyan la efectividad, aunque parcial, del programa de enseñanza. De acuerdo con Merrill, Li y Jones (1991), las personas logran un mejor conocimiento y registro de la información a partir del uso e integración de las representaciones mentales de un contenido temático, en estructuras que se interrelacionan con otras estructuras.

Estas nuevas representaciones son una interrelación de información previa y nueva, dejando de lado las representaciones aisladas. Por ello, el principio de integración sucede de forma progresiva y a un ritmo distinto en cada estudiante.

A partir de las evidencias encontradas sobre diseño instruccional de Merrill, se recomienda poner a prueba cada uno de sus principios en el desarrollo de sesiones de aprendizaje sobre Estadística en carreras de ciencias sociales, de la salud y del comportamiento. Asimismo, se debe promover el entendimiento y la crítica de los métodos estadísticos, evitando enseñar únicamente fórmulas o cálculos.

Por tanto, publicaciones relacionadas al uso de la Estadística en Psicología deben incluir en su contenido la situación presente y el debate acerca de distintos procedimientos y conceptos estadísticos. Actualmente existen buenos ejemplos que incorporan las recomendaciones señaladas (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).

Entre las limitaciones del estudio, se observa que el tamaño de muestra utilizado afecta a la representatividad de los resultados y el diseño de investigación no permite sacar conclusiones definitivas respecto a la efectividad del diseño instruccional. Se recomienda emplear un mayor tamaño de muestra en futuras replicaciones, apropiado para tener una suficiente potencia estadística y lograr una correcta generalización de los resultados (Perugini, Gallucci, & Costantini, 2018).

Adicionalmente, se sugiere utilizar un diseño de investigación experimental, donde los participantes sean seleccionados de forma aleatoria. De ese modo, atribuir cambios directos y causales al desarrollo del programa de enseñanza. Por otra parte, es apropiado replicar la investigación en otras universidades, no solamente de Lima sino también en otras ciudades de Perú, así como la revisión del grado en el que los estudiantes de Psicología presentan interpretaciones erróneas de diversos conceptos o métodos estadísticos.

Esta investigación muestra el problema del bajo nivel de comprensión en los conceptos estadísticos considerados básicos en diversas ciencias incluyendo la Psicología. La contribución teórica del estudio radica en la presentación y discusión de conceptos estadísticos básicos en la enseñanza de la Estadística en Psicología, útiles para ser revisados por estudiantes, investigadores y docentes. En cuanto a las implicaciones prácticas, se presentó la efectividad del diseño instruccional de Merrill

como estrategia didáctica para resolver el problema estudiado. Asimismo, es importante indicar que el estudio necesita ser replicado, por lo que se consignaron los materiales del programa de enseñanza y así facilitar futuras investigaciones que aborden estos temas y que los profesionales puedan contribuir a la enseñanza de los conceptos estadísticos.

Conclusiones

El presente estudio constituye una primera aproximación sobre este tipo de temáticas en el contexto peruano y debe ser interpretado con las consideraciones pertinentes a las características metodológicas presentadas. Los hallazgos indican que, en la muestra de estudio, al menos un ítem en ocho de los nueve tópicos existentes presentó un mayor puntaje que los reportados en estudios realizados en Estados Unidos y España. Sin embargo, el resto de los indicadores (18 de 27) presentaron una puntuación menor en comparación a los estudios. Respecto a las mediciones posteriores a la aplicación del programa, se observa que, a pesar de no evidenciarse diferencias estadísticamente significativas, en el 75% de los ítems (nueve de 12) se observaron diferencias a favor del postest (puntuaciones mayores). En específico, ocho de los ítems presentaron una diferencia de magnitud pequeña, mientras que el ítem 22 (acerca de los errores tipo I y II) mostró una diferencia de magnitud grande.

Los resultados permiten concluir que la muestra de estudiantes mostró un nivel bajo de conocimientos en algunos conceptos estadísticos relacionados a percepciones generales, MLG, análisis paso a paso, fiabilidad de las puntuaciones, errores tipo I y tipo II, influencia del tamaño de muestra, valor p como medida del tamaño del efecto, como medida directa de la importancia de los resultados y como evidencia de replicabilidad. La persistencia de estas deficiencias a lo largo del tiempo y en distintos espacios tiene como posible causa la enseñanza de estadística en pregrado. Por tanto, es necesario el uso de diseños instruccionales con evidencia empírica para mejorar la comprensión de los conceptos estadísticos en la formación en Psicología. En este sentido, el uso del programa de enseñanza basado en el diseño instruccional de Merrill logró una mejora significativa en nueve de los 12 indicadores considerados para su evaluación. Finalmente, es importante indicar que, este estudio se encuentra pendiente de replicación.

Referencias

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. https://doi.org/10.1080/00220970903352753

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. https://doi.org/10.1080/00031305.2018.1543137

Aron, A., Coups, E. J., & Aron, E. N. (2013). *Statistics for Psychology* (6a 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. https://doi.org/10.17060/ijodaep.2017.n1.v2.870

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. https://doi.org/10.7334/psicothema2014.283

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. https://doi.org/10.7203/anuari.Psicologia.16.2.15

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. https://doi.org/10.1177/2515245919858072

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. https://doi.org/10.14201/eks20161716585

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2a 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. https://doi.org/10.1111/j.1467-9280.2007.01881.x

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. https://doi.org/10.1177/2515245919847202

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). Recuperado de 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* (4a 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* (5a 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). Recuperado de 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. https://doi.org/10.1007/BF02505024

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

Merrill, M. D. (2009). First principles of instruction. En 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. https://doi.org/10.3758/s13423-015-0947-8

Müller, K. (2020). *Here: A simpler way to find your files*. Recuperado de 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). Recuperado de 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. Recuperado de 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. Recuperado de 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. En 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

Anexo A

Psychometrics Group Instrument (Mittag, 1999).

01 Las controversias sobre el uso de las pruebas de significación estadística existen desde hace mucho tiempo y, con toda seguridad, seguirán todavía muchos años más.
02 Sería mejor usar la frase "estadísticamente significativo" en lugar de "significativo" para describir los resultados en los que la hipótesis nula es rechazada.
03 La mayoría de los estudios de investigación tienen insuficiente potencia estadística para evitar el error Tipo II.
04 La ciencia progresaría más rápidamente si las pruebas de significación estadística fueran suprimidas de los artículos publicados.
05 La significación estadística informa que el investigador rechazó la hipótesis nula.
06 *Un resultado con una $p < 0.05$ indica que ese resultado es importante.*
07 La expresión «la Fiabilidad del Test» constituye una falsedad, ya que la fiabilidad no es una característica de ningún test en sí.
08 *Cuanto más pequeños son los valores de "p" más frecuentemente resultarán replicados dichos hallazgos en el futuro.*
09 Cometer error Tipo II es imposible cuando los resultados obtenidos son estadísticamente significativos.
10 Resultados estadísticamente significativos resultan más destacables cuando el tamaño de la muestra es pequeño.
11 *Valores de "p" pequeños ofrecen evidencia directa de que los efectos del tratamiento han sido grandes.*
12 Todos los análisis estadísticos ("t" de Student, "r" de Pearson, ANOVA…) son correlacionales.
13 *En regresión y otros análisis, el método "paso a paso" (stepwise) puede razonablemente ser usado para identificar el mejor subgrupo de predictores de entre un grupo dado.*
14 *Si una docena de investigadores estudiaron el mismo fenómeno usando la misma hipótesis nula, y ninguno de sus estudios arrojó resultados estadísticamente significativos, significa que los efectos investigados no son destacables ni importantes.*
15 Los valores de "p" obtenidos en una investigación miden la probabilidad de que dichos resultados ocurran en la muestra, pero no la probabilidad de que ocurran en la población.
16 Toda hipótesis nula será finalmente rechazada con un tamaño de muestra determinado.
17 *El error Tipo I se puede cometer cuando la hipótesis nula NO es rechazada.*
18 *Los estudios con resultados no-significativos pueden ser aún muy importantes.*
19 Comprobar la significación de un coeficiente de fiabilidad o de validez con $r^2 = 0$ como hipótesis nula no es útil ni productivo.
20 *Cuando los investigadores utilizan el método de análisis "paso a paso" (stepwise), el orden de entrada de las variables constituye un indicador muy útil de la importancia de cada variable introducida.*
21 *Las pruebas de significación estadística indican la probabilidad de que los resultados obtenidos con la muestra utilizada se den también en la población.*
22 *Es posible cometer a la vez error Tipo I y Tipo II en una prueba estadística.*
23 La utilización de datos poco fiables provoca una disminución o atenuación de los efectos que van a ser estadísticamente comprobados.
24 Los valores de "p" obtenidos en diferentes pruebas estadísticas no pueden ser directamente comparados, porque estos valores dependen de los tamaños de muestra utilizados en cada prueba.
25 Las pruebas de significación estadística indican, en parte, si el investigador ha trabajado o no con una muestra grande.
26 *No es posible usar regresión para comprobar estadísticamente la hipótesis de nula de que las medias de diferentes grupos son iguales.*
27 *Los resultados improbables son generalmente los más importantes y destacables.*
28 *La fiabilidad no afecta directamente a la probabilidad de obtener significación en un estudio concreto.*
29 *El cometer error Tipo II es bastante común en las investigaciones publicadas.*

Nota: En cursiva se encuentran los ítems considerados falsos.