Fiabilidad Test-Retest de Cuatro Tareas Cognitivas en una Muestra no muy WEIRD

Jose A. Rodas[a,b*], Ciara M. Greene[a]

[a] *School of Psychology, University College Dublin, Ireland*
[b] *Facultad de Ciencias Psicológicas, Universidad de Guayaquil, Ecuador*

Resumen

La mayoría de las tareas cognitivas han sido diseñadas y validadas en poblaciones occidentales, educadas, industrializadas, ricas y democráticas (WEIRD), lo que plantea interrogantes sobre la generalización de sus propiedades psicométricas a contextos culturalmente distintos. Este estudio examinó la fiabilidad test–retest de cuatro tareas cognitivas ampliamente utilizadas —la Tarea de Stroop de palabras y colores, la Tarea de Atención Sostenida a la Respuesta (SART), la Tarea de Señal de Parada y la Tarea de Actualización de Memoria de Letras— en una muestra ecuatoriana de estudiantes universitarios. Los datos se recogieron en dos sesiones con una separación de una semana y se analizaron mediante coeficientes de correlación de Pearson y coeficientes de correlación intraclase (ICC). Los resultados indican que la mayoría de las tareas muestran niveles de fiabilidad comparables a los observados en poblaciones WEIRD, con puntuaciones que oscilan entre moderadas y buenas. Sin embargo, se identificaron limitaciones en la fiabilidad de algunas medidas derivadas, como el índice de inhibición de Stroop, posiblemente influidas por factores culturales o metodológicos. Este estudio sugiere que las tareas cognitivas evaluadas son en gran medida fiables en contextos culturales no WEIRD, lo que respalda su aplicabilidad transcultural. Los hallazgos también subrayan la importancia de evaluar la fiabilidad en poblaciones diversas para fortalecer la validez ecológica de estas herramientas.

Palabras clave: *fiabilidad, Stroop, señal de parada, población WEIRD, SART, memoria de letras.*

Se ha observado que la mayor parte de la investigación psicológica publicada en revistas científicas utiliza muestras procedentes de sociedades occidentales, educadas, industrializadas, ricas y democráticas (WEIRD; Henrich et al., 2010), las cuales representan solo una fracción muy pequeña y específica de la población mundial. Además, muchos de estos estudios emplean muestras de estudiantes universitarios por su fácil accesibilidad. Dado que en numerosos países occidentales con altos ingresos las tasas universitarias suelen ser elevadas, estas muestras presentan características aún más particulares. Esta población no solo cuenta con un nivel educativo elevado, sino también con los recursos económicos, intelectuales o personales necesarios —por ejemplo, para obtener una beca— que les permiten matricularse en la universidad. Las características específicas de esta población han generado inquietudes entre investigadores sobre el grado en que estos resultados pueden generalizarse a otras poblaciones (Rad et al., 2018), especialmente si se considera la amplia variabilidad cultural existente entre países y el hecho de que estas diferencias culturales pueden ejercer efectos psicológicos relevantes en los individuos (Oyserman & Lee, 2008). Aunque muchas de las variaciones observadas entre muestras de distintos países podrían atribuirse a factores sociales y culturales, características compartidas dentro de las muestras no WEIRD también pueden influir en la cognición. Por ejemplo, la pobreza sostenida puede generar dificultades en el desarrollo neurológico infantil (Duncan & Brooks-Gunn, 2000; McLoyd, 1998), lo que puede derivar en deficiencias cognitivas. Asimismo, se han documentado diferencias cognitivas relacionadas con la cultura; por ejemplo, Masuda y Nisbett (2001) encontraron que los participantes japoneses prestaban más atención a las características contextuales de los estímulos que los participantes estadounidenses. También se han observado diferencias cognitivas adicionales entre culturas de Asia Oriental y culturas occidentales (Nisbett, 2003), lo que plantea interrogantes acerca de la estabilidad de los procesos cognitivos y de las propiedades psicométricas de los instrumentos utilizados para medirlos a través de fronteras nacionales y culturales.

A pesar de estas observaciones, persiste entre muchos psicólogos la idea de que estas diferencias podrían no ser

*La correspondencia sobre este artículo debe dirigirse a Jose A. Rodas. E-mail: jose.rodas@ucd.ie
Recibido el 16 ago. 2024; Aceptado el 28 oct. 2025; Disponible en línea el 9 dic. 2025.

tan relevantes o profundas como parecen, como se refleja en el grado de generalización atribuido a los resultados de numerosos estudios publicados. Esto puede ser particularmente frecuente en la psicología cognitiva, donde se estudian funciones básicas como la atención, la memoria y la percepción. En un comentario al artículo de Henrich et al. (2010), Gaertner et al. (2010) sostienen que el nivel de análisis desempeña un papel fundamental al determinar si las diferencias culturales pueden conducir a discrepancias en los fenómenos psicológicos o en el comportamiento entre poblaciones. Los autores presentan los comportamientos alimentarios como ejemplo: personas de diversas culturas siguen dietas específicas —como la exclusión de ciertos tipos de carne—, pero, si se analizan estas conductas en un nivel más básico, se observa que todos los seres humanos requieren similares niveles de sustento y nutrición. De forma análoga, aunque la función básica de la atención se observa en toda la humanidad, la manera en que un proceso atencional se emplea en situaciones concretas —por ejemplo, en interacciones sociales— puede diferir entre poblaciones. Una limitación de este planteamiento es que no se conoce con claridad el grado en que la cultura puede influir en la cognición y en sus procesos.

Un cuerpo creciente de investigación ha examinado la estabilidad y la capacidad de las tareas cognitivas para medir los procesos que fueron diseñadas para evaluar, especialmente en poblaciones poco representadas como las de América Latina. Las funciones ejecutivas, que incluyen el control atencional, la memoria de trabajo, la inhibición y la flexibilidad cognitiva, son fundamentales para la conducta orientada a objetivos (Lezak et al., 2012; Stuss, 2011). Estas funciones se desarrollan durante la infancia y la adolescencia en paralelo a la maduración cerebral, particularmente en los lóbulos frontales (Petersen & Posner, 2012). Estudios como el de Dias et al. (2015) muestran la aplicabilidad transcultural de los modelos de funciones ejecutivas. Por ejemplo, el modelo tripartito propuesto por Miyake et al. (2000), compuesto por memoria de trabajo, inhibición y flexibilidad cognitiva, se replicó en una muestra brasileña, lo que pone de relieve la necesidad de que las tareas cognitivas mantengan estabilidad y validez en contextos culturales diversos.

Sin embargo, muchos estudios procedentes de América Latina se publican en español o portugués, lo que limita su accesibilidad para la comunidad académica internacional. Esto resulta especialmente problemático en el caso de los datos normativos sobre tareas de funciones ejecutivas en poblaciones no WEIRD. Investigaciones como las de Fierro Bósquez et al. (2024) y Rivera et al. (2015) sobre funciones ejecutivas y tareas atencionales en poblaciones latinoamericanas muestran tanto patrones cognitivos universales como influencias culturales. Sesgos contextuales similares a los observados por Masuda y Nisbett (2001) en participantes japoneses podrían afectar también el rendimiento en tareas en poblaciones latinoamericanas. Estos hallazgos señalan la importancia de realizar evaluaciones psicométricas rigurosas que garanticen la fiabilidad y la validez entre distintas culturas. El presente estudio examina la fiabilidad test–retest de cuatro tareas ampliamente utilizadas para evaluar funciones ejecutivas —la Tarea Stroop de palabras y colores, la Tarea de Atención Sostenida a la Respuesta (SART), la Tarea de Actualización de Memoria de Letras y la Tarea de Señal de Parada— en una muestra ecuatoriana. Esta población ofrece un contexto singular, con diferencias sustanciales respecto a poblaciones WEIRD en términos culturales, lingüísticos y socioeconómicos. Al abordar estas lagunas, nuestro objetivo es avanzar en la comprensión de la estabilidad de las tareas cognitivas y de su aplicabilidad transcultural.

El papel de los factores socioculturales en la configuración de procesos cognitivos como la inhibición, la actualización y la atención sostenida sigue poco explorado, especialmente en poblaciones subrepresentadas. Funciones fundamentales como la atención suelen asumirse como universales debido a su centralidad en el procesamiento de la información, aunque la evidencia sugiere que pueden verse influidas por la cultura. Nisbett (2003) y Masuda y Nisbett (2001) documentaron diferencias marcadas en el foco atencional y en los estilos cognitivos entre culturas occidentales y de Asia Oriental. De manera similar, la investigación en América Latina muestra cómo los contextos culturales y lingüísticos afectan el desempeño en medidas de funciones ejecutivas (Fierro Bósquez et al., 2024; Rivera et al., 2015). Estos hallazgos ponen de relieve la necesidad de examinar cómo los factores socioculturales influyen en el rendimiento en tareas de funciones ejecutivas, particularmente en poblaciones no WEIRD. Las tareas desarrolladas en contextos occidentales pueden no contemplar patrones o sesgos cognitivos culturalmente específicos, lo que subraya la necesidad de evaluaciones psicométricas sólidas y culturalmente pertinentes para poblaciones diversas.

Una población no tan WEIRD

Los países latinoamericanos constituyen ejemplos pertinentes de muestras con características muy distintas de aquellas comúnmente utilizadas en la investigación con poblaciones WEIRD. El número medio de años de

escolaridad suele aproximarse más al promedio mundial, los problemas relacionados con procedimientos democráticos son frecuentemente reportados, el ingreso per cápita tiende a ser bajo y las poblaciones presentan una composición demográfica más heterogénea. El presente estudio se llevó a cabo en Ecuador, donde el 72% de la población se identifica como una mezcla de ascendencia indígena, africana y europea; el 23% vive con menos de 85 dólares mensuales, la mitad del ingreso considerado necesario para que una persona pueda vivir en el país (Instituto Nacional de Estadística y Censos, 2018); y el número medio de años de escolaridad es 8.7, en comparación con 13.4 en Estados Unidos, 12.9 en el Reino Unido y 8.4 a escala global (United Nations Development Programme, 2018). Según el informe del Programa de las Naciones Unidas para el Desarrollo de 2018, el ingreso nacional bruto per cápita en Ecuador fue de 10 347[1] dólares en 2017, mientras que el ingreso mundial medio per cápita fue de 15 295 dólares y de 10 055 dólares para los países en desarrollo en ese mismo año. En comparación, el ingreso per cápita en Estados Unidos y el Reino Unido en 2017 fue de 54 941 y 39 116 dólares, respectivamente (United Nations Development Programme, 2018). Para un análisis más profundo de los aspectos sociales y culturales de la población ecuatoriana, véase Capella et al. (2019).

La muestra de este estudio está conformada por estudiantes de grado en Psicología de la Universidad de Guayaquil, una universidad pública en Ecuador donde la educación se ofrece de manera gratuita. La legislación ecuatoriana impide que la universidad cobre tasas a los estudiantes y les prohíbe exigirles gastos directamente vinculados con sus obligaciones académicas. Guayaquil es la ciudad más grande y económicamente activa del país, una situación que ha contribuido a que la Universidad de Guayaquil se convierta en la más grande del Ecuador, con más de 60 000 estudiantes provenientes de distintas regiones del país en 2016 (Universidad de Guayaquil, 2016). Otro aspecto relevante de esta muestra es su escaso contacto previo con las tareas cognitivas empleadas en este estudio. Aunque en la universidad se realiza investigación psicológica, los estudiantes tienen una exposición muy limitada a ella, ya que la formación en Psicología se centra por completo en la práctica profesional (Capella & Andrade, 2017). Esto ha tenido diversas consecuencias en la titulación, como la ausencia de asignaturas orientadas a la investigación, entre ellas la psicología cognitiva. Los participantes en este estudio eran, por tanto, completamente neófitos respecto a los métodos y procedimientos utilizados en esta investigación.

Fiabilidad test–retest

La fiabilidad constituye una de las dos propiedades psicométricas principales de un instrumento psicológico y puede dividirse en fiabilidad interna y externa. La fiabilidad interna se entiende como la consistencia interna de un instrumento, es decir, el grado en que cada ítem se relaciona con los demás ítems que miden la misma variable. El método más utilizado para su estimación es probablemente el alfa de Cronbach (Cronbach, 1951; Tavakol & Dennick, 2011). La fiabilidad externa se refiere a la estabilidad de las puntuaciones obtenidas en diferentes mediciones realizadas en un periodo breve. Aunque ambas formas de fiabilidad evalúan propiedades psicométricas distintas, es habitual que solo se informe del alfa de Cronbach en estudios de fiabilidad. Diversos autores han recomendado la fiabilidad externa por encima de la consistencia interna, en particular mediante el método test–retest (Leppink & Pérez-Fuster, 2017; McCrae et al., 2011). Esto adquiere especial relevancia en las tareas cognitivas, dado que en la mayoría de los casos los ítems dentro de cada tarea (ensayos) son muy similares o incluso idénticos entre sí, lo que hace que la consistencia interna resulte menos pertinente que la estabilidad de las puntuaciones entre sesiones de evaluación.

Como indica su nombre, el método test–retest requiere administrar la tarea en al menos dos ocasiones distintas dentro de un periodo de tiempo breve, tras lo cual se comparan las puntuaciones. La estimación del coeficiente de correlación interclase (r de Pearson) es uno de los procedimientos más utilizados para comparar las puntuaciones, ya que refleja su correlación. Una limitación de emplear únicamente r de Pearson es que no permite detectar diferencias sistemáticas o sesgos entre puntuaciones. Por ejemplo, si se comparan dos mediciones, A y B, es posible que las puntuaciones de la medición B sean sistemáticamente superiores a las de A, aun produciendo un valor r muy alto. Por este motivo, suele resultar útil comprobar si existen diferencias significativas entre ambas mediciones mediante pruebas de hipótesis como t de Student o ANOVA. Este tipo de análisis también permite una inspección del acuerdo entre puntuaciones, es decir, una evaluación de la capacidad del instrumento para producir exactamente la misma

[1] Se utilizaron dólares PPA de 2011 en el informe. PPA significa Paridad del Poder Adquisitivo y es una medida común en macroeconomía para comparar el poder adquisitivo de un país con el de otros, ya que no solo refleja los tipos de cambio, sino que también compara el costo de una canasta común de bienes entre países. Se basa en el poder adquisitivo de un dólar dentro de los Estados Unidos durante un período específico, en este caso, el año 2011.

puntuación en dos ocasiones si no se ha producido ningún cambio en el constructo evaluado (Berchtold, 2016).

Se ha propuesto un método alternativo para estimar la fiabilidad test–retest que permite evaluar simultáneamente la correlación y el acuerdo dentro de un único índice: el coeficiente de correlación intraclase (ICC; McGraw & Wong, 1996). Una diferencia importante entre la correlación interclase y la intraclas es que la primera (r de Pearson) permite comparar mediciones de distintas "clases" o tipos de datos —por ejemplo, peso o número de kilómetros que una persona puede correr—, mientras que la segunda solo permite comparar elementos pertenecientes a la misma clase, como en el caso de puntuaciones asignadas por distintos evaluadores utilizando el mismo instrumento (McGraw & Wong, 1996). En consecuencia, el ICC constituye la medida más adecuada para evaluar la fiabilidad test–retest, ya que proporciona un único valor que oscila entre 0 (ausencia total de fiabilidad) y 1 (fiabilidad excelente) e integra tanto la correlación como el acuerdo entre mediciones.

Tareas incluidas en el presente estudio

Las cuatro tareas seleccionadas para este estudio se emplean con frecuencia en investigaciones sobre los procesos cognitivos de inhibición, actualización y atención sostenida en contextos experimentales y clínicos. Estos procesos son fundamentales para el funcionamiento cognitivo y la conducta. La inhibición, por ejemplo, regula el comportamiento al controlar impulsos considerados inapropiados en una situación determinada (Diamond, 2013). La capacidad de actualizar el contenido de la memoria de trabajo (en adelante, actualización) es esencial para el procesamiento de la información, dado que la capacidad de la memoria de trabajo es muy limitada y la información debe ser sustituida continuamente (Kessler & Oberauer, 2014). Por último, la capacidad de mantener la atención en una tarea es indispensable para muchas actividades cotidianas, como conducir o hacer compras, y suele verse afectada en diversas psicopatologías (Brands et al., 2005; Ebert & Kohnert, 2011).

Otra razón relevante para seleccionar estas tareas es que sus propiedades psicométricas en muestras WEIRD se han documentado en numerosos artículos (una descripción detallada de cada tarea se ofrece en la sección de Métodos). La primera de estas tareas, la Tarea de Señal de Parada, se desarrolló para evaluar el control inhibitorio de un impulso (Logan et al., 1997). El valor más importante derivado de esta tarea es el tiempo de reacción a la señal de parada (SSRT), una medida de la velocidad con la que los participantes pueden inhibir una respuesta prepotente una vez presentada la señal. Diversos laboratorios en Europa (Bekker et al., 2005; De Zeeuw et al., 2008), Estados Unidos (Blaskey, Harris, & Nigg, 2008) y Canadá (Toplak et al., 2009) han mostrado diferencias en el control inhibitorio entre pacientes con TDAH y controles mediante esta tarea. La fiabilidad test–retest se ha evaluado en muestras clínicas con TDAH y otros trastornos en Canadá y Estados Unidos (Soreni et al., 2009; Weafer et al., 2013), hallándose una fiabilidad moderada (véase la Tabla 1). Laboratorios europeos también han evaluado su fiabilidad test–retest, aunque algunos han obtenido resultados heterogéneos. En un estudio del Reino Unido y los Países Bajos sobre la fiabilidad de varias tareas, se informó una fiabilidad extremadamente baja para el SSRT (Kuntsi et al., 2001), y otro estudio del Reino Unido y Alemania (Wöstmann et al., 2013) no encontró fiabilidad alguna, es decir, una correlación interclase no significativa. Una búsqueda en revistas latinoamericanas de las bases de datos Scopus y Redalyc utilizando el término "Stop Signal" mostró que, aunque la tarea aparece mencionada en más de 50 estudios, solo cuatro la han empleado con una población latinoamericana y ninguno informó propiedades psicométricas.

La Tarea Stroop de palabras y colores es también una medida habitual del control inhibitorio, aunque no fue diseñada originalmente con ese propósito (Stroop, 1935). Se pide a los participantes que respondan al color en el que está impresa una palabra mientras ignoran su significado; por ejemplo, si la palabra "rojo" aparece impresa en tinta azul, deben responder "azul", inhibiendo así la respuesta automática derivada del contenido semántico. La tarea también se ha utilizado como medida de atención selectiva o sesgo atencional (Atkinson et al., 2009; Epp et al., 2012) y de velocidad de procesamiento (Van Den Heuvel et al., 2006). Ha sido validada como una tarea de función ejecutiva de distintas maneras; por ejemplo, en varios estudios de neuroimagen se ha observado que la corteza cingulada anterior, una región considerada fundamental para la atención selectiva, desempeña un papel destacado durante la ejecución del Stroop (Botvinick et al., 2004; Pardo et al., 1990). La versión original de la tarea se administraba mediante tarjetas impresas, aunque la mayoría de los estudios actuales emplean una versión informatizada capaz de registrar la latencia de las respuestas con un alto nivel de precisión (Dalgleish, 1995).

En un estudio con una muestra estadounidense, se han hallado puntuaciones de fiabilidad test–retest adecuadas con el formato de tarjetas (Franzen et al., 1987). La fiabilidad test–retest también se ha evaluado en poblaciones suizas y estadounidenses utilizando la versión

informatizada de la tarea, encontrándose igualmente una buena fiabilidad (Siegrist, 1995, 1997; Strauss et al., 2005). Hedge y otros (2018) observaron fiabilidades moderadas en una muestra del Reino Unido en dos estudios distintos utilizando la versión informatizada. Varios estudios han evaluado la fiabilidad test–retest del formato de tarjetas en muestras latinoamericanas (Rodríguez Barreto et al., 2016), encontrando en general puntuaciones altas; sin embargo, hasta donde sabemos no se han realizado análisis de fiabilidad de la versión informatizada en esta población. Las versiones en tarjeta e informatizada registran las respuestas de manera muy distinta: en la versión en tarjeta se registra el tiempo total necesario para leer todos los estímulos de un tipo, mientras que en la versión informatizada se registra el tiempo de respuesta (TR) ante cada estímulo presentado. Estas diferencias podrían afectar la fiabilidad de la tarea, por lo que resulta importante evaluar la fiabilidad test–retest en esta población utilizando la versión informatizada, hoy ampliamente difundida.

Tabla 1

Puntuaciones de fiabilidad de estudios previos realizados con muestras WEIRD

Estudio	Origen de la muestra[a]	ICC	*r* de Pearson
Stroop de palabras y colores (puntuación de inhibición)			
Hedge et al. (2018)[b]	Reino Unido	0.60	
Hedge et al. (2018)[c]	Reino Unido	0.66	
Siegrist (1995) [b]	Suiza		0.73
Siegrist (1997) [c]	Suiza		0.68
Strauss et al. (2005)	Estados Unidos		0.46
SART (errores de comisión)			
Robertson et al. (1997)	Reino Unido		0.76
Tarea de Señal de Parada (SSRT)			
Hedge et al. (2018) [b]	Reino Unido	0.47	
Hedge et al. (2018) [c]	Reino Unido	0.43	
Kuntsi et al. (2001)	Reino Unido	0.11	
Soreni et al. (2009)	Canadá	0.72	
Weafer et al. (2013)	Estados Unidos		0.65
Wöstmann et al. (2013)	Alemania	0.03	0.03

Nota. ICC = Coeficiente de correlación intraclase.

[a] Esta columna indica el país en el que se obtuvo la muestra. Desafortunadamente, los estudios incluidos no informaron la nacionalidad de los miembros de la muestra, por lo que no podemos excluir la posibilidad de que se incluyeran participantes de países no-WEIRD.

[b] Estudio 1

[c] Estudio 2

La SART se desarrolló como una medida de atención sostenida y requiere que los participantes respondan a un flujo continuo de estímulos mientras inhiben la respuesta ante un objetivo poco frecuente. Los autores originales

(Robertson et al., 1997) aportaron evidencia de su validez y fiabilidad en una muestra británica. En ese estudio se observó una correlación significativa entre la SART y otras pruebas de atención sostenida, además de una correlación test–retest adecuada. Se realizó una búsqueda en las bases de datos Redalyc y SciELO para localizar estudios que evaluaran las propiedades psicométricas de esta tarea en poblaciones latinoamericanas, pero no se encontraron estudios que reportaran fiabilidad[2]. Cheyne y otros (2009) han propuesto que problemas atencionales más específicos —a saber, inatención focal, inatención global y desconexión conductual— pueden evaluarse utilizando puntuaciones adicionales derivadas de la SART como indicadores de dificultades de implicación con la tarea. Estos problemas atencionales aparecen en forma de etapas progresivas: en la primera etapa los participantes se desconectan ligeramente de la tarea, lo que provoca fluctuaciones en su rendimiento (por ejemplo, si divagan brevemente, su latencia de respuesta aumenta); en la segunda, la atención se vuelve inestable y, finalmente, en la tercera, la atención del participante está completamente desvinculada de la tarea y no responde en absoluto. Una limitación de estas medidas es que, hasta donde sabemos, no se han publicado estimaciones de fiabilidad test–retest.

La Tarea de Actualización de Memoria de Letras (Updating Letter Memory task, ULMT) (Miyake et al., 2000) se adaptó de una tarea de reconocimiento (Morris & Jones, 1990) y se modificó para permitir la evaluación del proceso de actualización y supervisión de la información mantenida en memoria de trabajo. Aunque se han utilizado otros métodos para evaluar esta función (por ejemplo, Garavan, 1998; Kessler & Oberauer, 2014), la ULMT ha mostrado una sólida evidencia de validez, evaluada mediante análisis factorial confirmatorio junto con otras tareas diseñadas para medir la actualización (Friedman et al., 2008; Miyake & Friedman, 2012). El proceso de actualización ha sido investigado con menor profundidad que los procesos evaluados por otras tareas (Ecker et al., 2010; Kessler & Oberauer, 2014), y existe poca información sobre la fiabilidad de los instrumentos empleados para medirlo, incluida la ULMT. No obstante, los autores originales han informado en varias ocasiones de una consistencia interna adecuada (Friedman et al., 2006; Miyake et al., 2000). En un estudio con una muestra brasileña de personas diagnosticadas con esquizofrenia, Berberian y otros (2015) investigaron las propiedades psicométricas de la ULMT. La tarea correlacionó adecuadamente con otra prueba que requería la actualización de información procedente de distintas categorías y mostró una buena consistencia interna. Lamentablemente, no se evaluó la fiabilidad test–retest.

Metodología

Diseño

Este estudio empleó un diseño test–retest para evaluar la fiabilidad de cuatro tareas cognitivas ampliamente utilizadas —la Tarea Stroop de palabras y colores, la SART, la Tarea de Señal de Parada y la ULMT— en una muestra de estudiantes de psicología de la Universidad de Guayaquil, Ecuador. Los participantes fueron evaluados en grupos mediante versiones informatizadas de estas tareas, las cuales miden procesos cognitivos como la inhibición, la atención sostenida y la actualización de la memoria de trabajo. Los datos se recogieron en dos sesiones separadas por una semana y la fiabilidad se evaluó mediante coeficientes de correlación intraclase (ICC) para estimar la estabilidad de las puntuaciones. El diseño buscó examinar si las diferencias culturales y demográficas de esta población no WEIRD influían en las propiedades psicométricas de las tareas.

Participantes

El estudio fue anunciado mediante folletos en la Facultad de Psicología de la Universidad de Guayaquil como una investigación de psicología cognitiva centrada en procesos mentales evaluados a través de tareas experimentales. Un total de 185 estudiantes de grado (111 mujeres; edad media = 19.54, DE = 3.53) respondió al anuncio y participó en la primera sesión de evaluación. Se esperaba que todos los participantes completaran todas las tareas; sin embargo, algunos no aceptaron realizar todas las pruebas y se produjeron fallos técnicos en los ordenadores utilizados para la recogida de datos, lo que provocó la pérdida de parte de la información. En consecuencia, 96 participantes completaron la segunda administración de la Tarea Stroop, 88 la SART, 86 la Tarea de Actualización de Memoria de Letras y 115 la Tarea de Señal de Parada. Algunos participantes señalaron dificultades para comprender las instrucciones y sus datos fueron excluidos del análisis. Los tamaños finales de muestra fueron 94 para la Stroop, 87 para la SART, 86 para la ULMT y 112 para la Tarea de Señal de Parada. No se aplicaron criterios específicos de inclusión; todos los estudiantes que se ofrecieron voluntariamente y otorgaron su consentimiento informado pudieron participar.

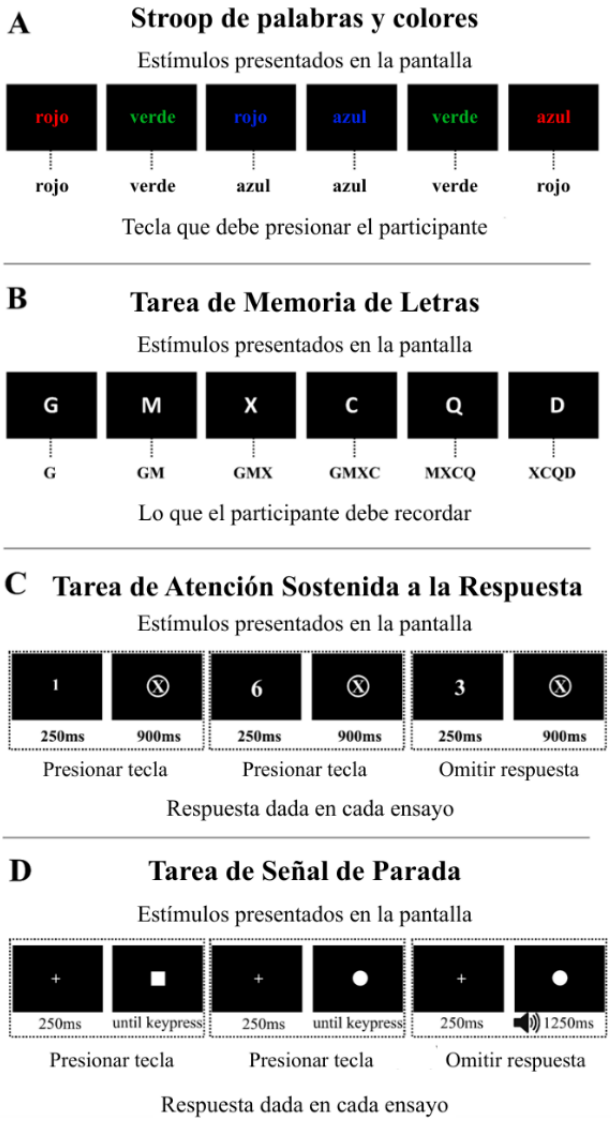
[2] Los términos "SART", "sustained attention" y "atención sostenida" se utilizaron como cadenas de búsqueda.

Instrumentos

Tarea Stroop de palabras y colores (Stroop, 1935)

En esta tarea se presentaban las palabras "red", "green" o "blue" una a la vez en el centro de la pantalla, y el color de la fuente era congruente con la palabra (por ejemplo, la palabra "red" en color rojo) o incongruente (por ejemplo, la palabra "blue" en color verde). Se pedía a los participantes que pulsaran una tecla específica en respuesta al color de la fuente, ignorando el significado de la palabra (véase la Figura 1A). La tarea incluía 48 ensayos congruentes y 24 incongruentes, además de 18 ensayos congruentes y 9 incongruentes de práctica. La proporción de ensayos congruentes e incongruentes fue similar a la empleada por Kane y Engle (2003), donde se aumentó la proporción de ensayos congruentes para intensificar el efecto Stroop. Solo se utilizaron las condiciones congruente e incongruente, presentadas en un único bloque en un orden seudorrandomizado. La medida principal utilizada es la diferencia en el TR entre ensayos incongruentes y congruentes (MacLeod, 1991), que representa la puntuación de inhibición.

Figura 1



Nota. Presentación de estímulos en las cuatro tareas. Cada cuadrado negro representa lo que se muestra en la pantalla en un momento determinado. En el panel A, las palabras "rojo", "verde" y "azul" se presentan solo como ejemplo, ya que en la tarea real utilizada con la muestra ecuatoriana se emplearon las traducciones al español de estas palabras ("rojo", "verde" y "azul", respectivamente). Los grupos de cuadrados negros encerrados en líneas punteadas en los paneles C y D representan un ensayo.

Tarea de Atención Sostenida a la Respuesta (Robertson et al., 1997)

Se presentaban dígitos blancos de distintos tamaños de fuente sobre un fondo negro durante 250 ms, seguidos de una máscara compuesta por una X blanca dentro de un círculo durante 900 ms. La presentación de cada dígito y su máscara correspondiente constituía un ensayo. Se instruía a los participantes para que pulsaran la barra espaciadora cada vez que aparecía un dígito distinto de 3 (ensayos go) y que inhibieran la respuesta cuando aparecía el dígito 3 (ensayos no-go; véase la Figura 1B). Los dígitos utilizados eran los números del 1 al 9 y se presentaban en orden seudoaleatorizado. Los tamaños de fuente (48, 72, 94, 100 y 120 píxeles de altura) se distribuían equitativamente entre los nueve dígitos. Los participantes completaron primero 27 ensayos de práctica consistentes en tres presentaciones de cada dígito. Tras la práctica, completaron 225 ensayos experimentales con 25 presentaciones de cada dígito. La medida más empleada es el número de errores por comisión, producidos al pulsar la tecla en un ensayo no-go. En relación con las tres fases progresivas de desimplicación de la tarea descritas por Cheyne et al. (2009): (1) la inatención focal se midió mediante el coeficiente de variabilidad del tiempo de reacción (RTCV), calculado dividiendo la desviación estándar del TR en ensayos go por la media del TR en estos ensayos; (2) la inatención global se midió mediante el número de respuestas anticipadas, es decir, respuestas emitidas dentro de los primeros 200 ms tras la aparición del estímulo; y (3) la desconexión de la respuesta se midió mediante el número de errores por omisión (no pulsar en ensayos go).

Tarea de Actualización de Memoria de Letras

Esta tarea se adaptó de Miyake et al. (2000) y se utiliza habitualmente para evaluar el rendimiento en actualización de la memoria de trabajo. Se presentaban letras mayúsculas una a la vez en el centro de la pantalla en una serie continua. Los participantes debían mantener un registro de las últimas cuatro letras presentadas. El número de letras por lista podía ser 5, 7, 9 u 11, y era desconocido para el participante. Por ello, era necesario actualizar continuamente la información almacenada, recordando la nueva letra presentada y olvidando las anteriores más allá de las cuatro últimas (véase la Figura 1C). Tras la presentación completa de una lista, se pedía al participante que escribiera las últimas cuatro letras en el orden correcto. Las listas se generaban aleatoriamente al inicio de la tarea e incluían únicamente consonantes. Las letras “B” y “W” no se utilizaron, ya que la “B” es fonéticamente similar a otra letra y la “W” es fonéticamente más larga que el resto. Se presentaron cuatro listas de práctica con longitudes seleccionadas al azar. Tras la práctica, se presentaron cuatro listas de cada longitud (16 en total) en orden aleatorio como ensayos experimentales. Cada letra permanecía en pantalla 2 segundos, con un intervalo entre estímulos de 500 ms. La medida de resultado fue el número de letras recordadas correctamente.

Tarea de Señal de Parada (Logan, 1994)

Para este estudio se empleó el programa STOP-IT para la recogida de datos y el programa ANALYZE-IT para su procesamiento y obtención de las medidas de resultado, ambos desarrollados por Verbruggen y otros (2008). Los autores han puesto ambos programas a disposición del público para su descarga en Open Science Framework (https://osf.io/wuhpv/). Cada ensayo comenzaba con la presentación de un signo de fijación (“+”) en el centro de la pantalla durante 250 ms, seguido del estímulo de la tarea primaria, que consistía en un cuadrado o un círculo. Este estímulo permanecía en pantalla 1250 ms o hasta que el participante pulsaba la tecla correspondiente (“z” para el cuadrado y “/” para el círculo). Se instruía a los participantes para responder lo más rápido y precisamente posible cuando se presentaba el estímulo. En el 25% de los ensayos, el estímulo iba seguido de una señal de parada: un sonido de 750 Hz presentado durante 75 ms que indicaba al participante que debía inhibir su respuesta y evitar pulsar cualquier tecla. El sonido se presentaba siempre con un retraso respecto al estímulo; este retraso, denominado Stop Signal Delay (SSD), comenzaba en 250 ms y aumentaba en 50 ms tras cada inhibición exitosa, disminuyendo en 50 ms tras cada fallo en inhibir la respuesta. El intervalo interestímulos era de 2 segundos y se presentaron 224 ensayos en total. Véase la Figura 1D para ejemplos de ensayos. Los ensayos se agruparon en cuatro bloques: el primero consistió en 32 ensayos de práctica y los otros tres, que eran experimentales, incluyeron 64 ensayos cada uno. La medida más utilizada es el tiempo de reacción a la señal de parada (SSRT), que se calcula restando la SSD media del tiempo de reacción medio en los ensayos sin señal de parada; esta medida representa el proceso interno de detención de la respuesta (para más detalles sobre el modelo véase Logan, 1994).

Todas las tareas se administraron mediante ordenador y se presentaron con una resolución de 1920 × 1080 píxeles. Todos los ordenadores emplearon teclados USB. La Tarea Stroop, la SART y la ULMT se programaron en PsychoPy2 (Peirce, 2008).

Procedimiento

Los participantes fueron invitados a inscribirse en una de ocho sesiones grupales de evaluación. Estas sesiones tuvieron lugar en una sala silenciosa de la Facultad de Psicología equipada con 30 ordenadores. El número de participantes por grupo osciló entre 25 y 30, y todas las sesiones se realizaron en el transcurso de una semana. Al inicio de cada sesión, los participantes recibieron una explicación sobre el experimento, en la que se mencionaba que una nueva evaluación tendría lugar la semana siguiente, y se les pidió firmar el consentimiento informado. Tras otorgar su consentimiento, comenzó la evaluación. El orden de presentación de las tareas se contrabalanceó mediante un cuadrado latino. Cada tarea iniciaba con instrucciones presentadas en pantalla, que los participantes debían leer en silencio mientras el investigador las leía en voz alta para todo el grupo. En el caso de la Tarea de Señal de Parada, las instrucciones se entregaron impresas en español a cada participante al inicio de la sesión, dado que el programa presenta las instrucciones en inglés. Durante la administración de las tareas, todos los participantes utilizaron auriculares debido a que la Tarea de Señal de Parada incluye estímulos auditivos. Además de las cuatro tareas descritas en este artículo, los participantes completaron otra tarea perteneciente a un estudio distinto y que no se describe aquí.

Por limitaciones de espacio y recursos, se realizaron sesiones de retest separadas para cada tarea, llevadas a cabo una semana después de la sesión inicial. Estas sesiones se realizaron en grupos de 30 participantes en la misma sala utilizada para la evaluación inicial.

Ética y consentimiento para participar

El estudio fue aprobado por el comité de ética de la Facultad de Psicología de la Universidad de Guayaquil, y los participantes no recibieron compensación alguna por su participación. Antes de comenzar, se explicaron los objetivos del estudio y se firmó un consentimiento informado. Todos los participantes fueron informados de que los datos serían anonimizados y utilizados con fines de investigación.

Análisis de datos

El análisis se realizó para evaluar la fiabilidad test–retest de cuatro tareas cognitivas utilizando el ICC, siguiendo el modelo de efectos mixtos bidireccional para acuerdo absoluto (Shrout & Fleiss, 1979). Antes de los análisis de fiabilidad se aplicó la prueba de Shapiro–Wilk y se generaron gráficos de distribución para analizar si los datos seguían una distribución normal. Los valores de ICC se interpretaron según las directrices de Koo y Li (2016), clasificando la fiabilidad como baja (<0.5), moderada (0.5–0.75), buena (0.75–0.9) o excelente (>0.9). Además, se calcularon coeficientes de correlación de Pearson (r) para facilitar la comparación con estudios previos. El análisis incluyó estimaciones puntuales e intervalos de confianza del 95% para los ICC, pero no se reportaron valores p, de acuerdo con las recomendaciones para estudios de fiabilidad (McGraw & Wong, 1996). Se informaron estadísticas descriptivas del rendimiento en las tareas en ambos momentos de evaluación y se calcularon ICC para las medidas principales de cada tarea, incluidas latencias de respuesta, tasas de error y puntuaciones derivadas, como el índice de inhibición en la tarea Stroop. Los resultados se resumieron para facilitar comparaciones transculturales con estimaciones de fiabilidad en poblaciones WEIRD.

Resultados

Los datos y Materiales Suplementarios de este estudio pueden consultarse en línea en https://osf.io/da39c/. Los resultados de las pruebas de Shapiro–Wilk indicaron que la mayoría de las variables se desviaban significativamente de la normalidad según sus valores p. Sin embargo, la mayoría de los valores W superaron .9 y los gráficos de distribución no mostraron desviaciones sustanciales respecto a la normalidad. Esto sugiere que, aunque los datos no se ajusten estrictamente a una distribución normal, estas desviaciones probablemente no introduzcan sesgos en los resultados al utilizar ICC y coeficientes de correlación de Pearson. Los detalles de los valores W y los gráficos de distribución se encuentran disponibles en los Materiales Suplementarios.

Los análisis de la Tarea Stroop se realizaron con datos de 94 participantes (60 mujeres; edad media = 19.65, DE = 3.83). Como se observa en la Tabla 2, los ICC de los ensayos congruentes e incongruentes son moderados, con intervalos de confianza del 95% que abarcan de moderado a bueno. Los análisis de la ULMT se realizaron con datos de 86 participantes (56 mujeres; edad media = 19.33, DE = 2.21). Como muestra la Tabla 2, el ICC del número de letras recordadas correctamente es bueno, con un intervalo de confianza del 95% que va de moderado a bueno. Finalmente, los análisis de la Tarea de Señal de Parada se realizaron con datos de 112 participantes (69 mujeres; edad media = 19.54, DE = 3.56). Como se observa en la Tabla 2, los ICC de la probabilidad de responder en ensayos con señal de parada, la demora de la señal de parada y el tiempo medio de respuesta en ensayos sin

señal son buenos, y los intervalos de confianza del 95% van de moderado a bueno. En el caso del tiempo de reacción a la señal de parada, el ICC es moderado, con un intervalo de confianza del 95% que oscila entre moderado y bueno.

Tabla 2

Resumen de puntajes de fiabilidad test-retst en la muestra ecuatoriana

	N	Momento 1		Momento 2		*R de Pearson*	ICC	Intervalo de confianza 95%	
		M	*DE*	*M*	*DE*			Limite inferior	Limite superior
Stroop									
Congruente	94	0.82	0.23	0.74	0.18	0.59**	0.69	0.50	0.81
Incongruente	94	0.95	0.31	0.82	0.21	0.69**	0.73	0.50	0.84
Inhibicion	94	0.12	0.15	0.08	0.12	0.25*	0.37	0.07	0.58
SART									
Correcto	87	205.91	17.63	198.06	21.65	0.65**	0.74	0.55	0.84
Comisiones	87	9.85	5.93	11.26	5.60	0.58**	0.72	0.57	0.82
TR Medio	87	0.43	0.09	0.45	0.10	0.62**	0.76	0.63	0.84
RTCV	87	0.29	0.10	0.34	0.13	0.46**	0.58	0.33	0.73
Anticipaciones	87	6.25	11.59	10.95	13.88	0.64**	0.75	0.57	0.84
Omisiones	87	2.99	4.12	4.72	6.69	0.43**	0.54	0.30	0.70
ULMT									
Letras recordadas	86	35.31	10.48	38.95	12.61	0.69**	0.79	0.65	0.87
Stop-Signal Task									
p(r|s)	112	52.02	15.21	52.52	14.22	0.69**	0.82	0.73	0.87
SSD	112	323.38	167.84	347.64	174.76	0.68**	0.81	0.72	0.87
SSRT	112	303.44	81.71	285.80	74.81	0.57**	0.71	0.58	0.80
No signal RT	112	627.55	133.31	634.58	141.21	0.64**	0.78	0.68	0.85

Nota. RTCV = Coeficiente de variabilidad de tiempo de respuesta; p(r|s) = probabilidad de responder a los ensayos de señal de parada; SSD = retraso de la señal de parada; SSRT = tiempo de respuesta de la señal de parada

*Correlación significativa al 0.05

** Correlación significativa al 0.001

Los análisis de la ULMT se realizaron con datos de 86 participantes (56 mujeres; edad media = 19.33, DE = 2.21). Como muestra la Tabla 2, el ICC del número de letras recordadas correctamente es bueno, con un intervalo

de confianza del 95% que va de moderado a bueno. Finalmente, los análisis de la Tarea de Señal de Parada se realizaron con datos de 112 participantes (69 mujeres; edad media = 19.54, DE = 3.56). Como se observa en la Tabla 2, los ICC de la probabilidad de responder en ensayos con señal de parada, la demora de la señal de parada y el tiempo medio de respuesta en ensayos sin señal son buenos, y los intervalos de confianza del 95% van de moderado a bueno. En el caso del tiempo de reacción a la señal de parada, el ICC es moderado, con un intervalo de confianza del 95% que oscila entre moderado y bueno.

Discusión

Se ha observado que pueden existir diferencias cognitivas entre culturas (Nisbett, 2003) y que estas diferencias pueden afectar de forma específica el rendimiento en tareas cognitivas según el contexto cultural (Masuda & Nisbett, 2001). Esto plantea la cuestión de cuán fiables son realmente estas tareas cognitivas, empleadas tanto en investigación como en entornos clínicos, cuando se aplican en poblaciones con antecedentes culturales distintos. La comparación de nuestros datos con los observados en diferentes estudios de fiabilidad test–retest sugiere que la fiabilidad parece verse menos afectada por variables culturales que por otros tipos de variables, como las características específicas del diseño de la tarea, las instrucciones dadas a los participantes o la presencia de variables no controladas en la muestra (Wöstmann et al., 2013). Este planteamiento se ve respaldado por la gran variabilidad observada en las puntuaciones de fiabilidad informadas en estudios realizados en poblaciones culturalmente muy similares (véase la Tabla 1). Por ejemplo, la estimación puntual del ICC para la puntuación de inhibición en la Tarea de Señal de Parada (SSRT) oscila entre 0.03 y 0.72, siendo la puntuación ecuatoriana la segunda más alta. Esto sugiere que los valores de SSRT observados en la muestra ecuatoriana son tan fiables como los registrados en otras poblaciones. La otra tarea que permite este tipo de comparación, debido al número de estudios que han informado fiabilidad test–retest, es la Tarea Stroop.

Las puntuaciones de fiabilidad de la medida de inhibición son más estables entre estudios que las observadas en la Tarea de Señal de Parada y tienden a interpretarse como moderadas en la mayoría de los casos, salvo en el informe de Strauss et al. (2005), donde se observa una correlación baja para un diseño test–retest. La fiabilidad encontrada en nuestro estudio para esta medida concreta es baja, lo que indica que constituye un indicador deficiente de inhibición.

Aunque los ICC de los ensayos congruentes e incongruentes de la Stroop en la muestra ecuatoriana son moderados a buenos, de acuerdo con los intervalos de confianza del 95%, el ICC de la puntuación de inhibición disminuye de forma notoria, situándose en gran parte dentro del rango considerado de fiabilidad baja (0 a 0.49). No es sorprendente hallar fiabilidades menores en puntuaciones derivadas, ya que suelen acumular el error de medición de las variables directas de las que proceden (Caruso, 2004; Thomas & Zumbo, 2012). Sin embargo, esto no basta para explicar una fiabilidad tan baja en la muestra ecuatoriana, especialmente cuando las fiabilidades halladas en otros estudios son sustancialmente más altas. Una posible explicación es que el elevado número de participantes evaluados simultáneamente afectara su rendimiento de diversas maneras. Por ejemplo, podrían haberse distraído por la presencia de otros participantes o incluso haber intentado mirar otras pantallas, lo que habría interferido en la concentración. Esta situación podría haber influido en el rendimiento y en la fiabilidad de la tarea.

La fiabilidad encontrada en la SART, medida mediante ICC, es similar a la informada en otros estudios de fiabilidad realizados con tareas distintas. Esto implica que el número de errores por comisión es una puntuación fiable en la muestra ecuatoriana, o al menos tan fiable como suelen serlo las puntuaciones derivadas de tareas cognitivas. En cuanto a las otras tres medidas propuestas por Cheyne et al. (2009) —RTCV, número de anticipaciones y número de omisiones— solo las anticipaciones mostraron una fiabilidad moderada a buena. Las otras dos medidas presentaron fiabilidades de pobres a moderadas. Es posible que el formato grupal de evaluación utilizado en este estudio también haya influido en el rendimiento en esta tarea, especialmente dada su monotonía, que podría desviar la atención hacia otros participantes o hacia sus pantallas. No obstante, si se pretende utilizar el RTCV y las omisiones, conviene realizar esfuerzos para mejorar la fiabilidad de la tarea. Una opción sería aumentar el número de ensayos. Una de las ventajas de la SART es que, siguiendo el diseño original (número de ensayos, duración del estímulo, intervalo entre estímulos, etc.), suele completarse en torno a cinco minutos, de modo que incrementar la longitud no supondría dificultades significativas.

La información disponible sobre la fiabilidad test–retest de la SART y la ULMT es más limitada que la existente para la Stroop y la Tarea de Señal de Parada; de hecho, solo hallamos un estudio previo que reportara esta información para la SART (Robertson et al., 1997) y ninguno para la ULMT, lo que imposibilita la

comparación entre distintas poblaciones. A pesar de esta limitación, el presente estudio constituye una contribución relevante, ya que se presentan los valores de fiabilidad ICC de ambas tareas. La ULMT, en el formato utilizado por Miyake y Friedman (2012) en sus estudios sobre funciones ejecutivas, ha sido empleada por diversos laboratorios (Dahlin et al., 2008; St Clair-Thompson & Gathercole, 2006) al ofrecer una evaluación directa del proceso de actualización de la memoria de trabajo; sin embargo, hasta donde sabemos, la estabilidad temporal de su puntuación no había sido evaluada. Según nuestros resultados, esta tarea puede utilizarse como una medida fiable de la capacidad de actualización. En el caso de la SART, el presente estudio aporta una estimación ICC, una métrica de fiabilidad más adecuada que la correlación interclase empleada por los autores originales.

Este estudio presenta varias limitaciones metodológicas. Las diferencias procedimentales entre las sesiones de pretest y retest —con todas las tareas completadas en una sola sesión en el pretest, pero administradas por separado en el retest— pueden haber introducido factores de confusión como fatiga, variaciones en la motivación y cambios en el orden de las tareas, lo que podría haber afectado el rendimiento y las estimaciones de fiabilidad. Estas diferencias también plantean dudas sobre la equivalencia de las condiciones de evaluación en el cálculo de los ICC, dado que estos coeficientes dependen de condiciones de medición consistentes. Además, problemas técnicos durante la recogida de datos provocaron la pérdida de parte de la información, reduciendo la solidez y representatividad de la muestra. Otra limitación es la ausencia de análisis de validez convergente, que permitiría determinar si las tareas miden de forma efectiva los procesos cognitivos que pretenden evaluar. Aunque este estudio se centró en la fiabilidad test–retest en una población no WEIRD, futuras investigaciones deberían examinar la validez de estas tareas para reforzar su aplicabilidad transcultural y su solidez psicométrica.

Respecto a la muestra, es importante señalar que, aunque los estudiantes universitarios de la Universidad de Guayaquil presentan diferencias importantes respecto a los de países más WEIRD, siguen siendo jóvenes que han podido completar los estudios secundarios y que buscan obtener un título universitario. Esto implica que, en cierta medida, han superado dificultades comunes en la población ecuatoriana general, como pobreza extrema, violencia en barrios de bajos ingresos y posibles problemas nutricionales. En este sentido, representan únicamente a la parte de la población ecuatoriana con acceso a educación superior y comparten algunas características con muestras típicas de estudios en países WEIRD. No obstante, esta población mantiene rasgos culturales distintos de los de las poblaciones WEIRD (Capella et al., 2019), lo que proporciona información valiosa sobre los efectos de variables culturales en el rendimiento cognitivo.

En conjunto, nuestros datos sugieren que las estimaciones de fiabilidad de las tareas en la población ecuatoriana son muy similares a las observadas en países WEIRD, lo que indica que las medidas son estables en poblaciones con diferencias culturales importantes. Esto respalda la idea de que los hallazgos de la investigación en cognición pueden generalizarse a distintas poblaciones y apunta a la existencia de elementos compartidos en los procesos cognitivos subyacentes. Además, el estudio ofrece una base para el desarrollo de futuras investigaciones en Ecuador y otros países de América Latina, una región con una producción científica muy limitada, especialmente en el ámbito de la psicología cognitiva.

Referencias Bibliográficas

Atkinson, L., Leung, E., Goldberg, S., Benoit, D., Poulton, L., Myhal, N., … Kerr, S. (2009). Attachment and selective attention: Disorganization and emotional Stroop reaction time. *Development and Psychopathology*, *21*(1), 99–126. https://doi.org/10.1017/S0954579409000078

Bekker, E. M., Overtoom, C. C., Kenemans, J. L., Kooij, J. J., De Noord, I., Buitelaar, J. K., & Verbaten, M. N. (2005). Stopping and changing in adults with ADHD. *Psychological Medicine*, *35*(6), 807–816. https://doi.org/10.1017/S0033291704003459

Berberian, A. A., Gadelha, A., Dias, N. M., Mecca, T. P., Bressan, R. A., & Lacerda, A. T. (2015). Investigation of cognition in schizophrenia: Psychometric properties of instruments for assessing working memory updating. *Jornal Brasileiro de Psiquiatria*, *64*(3), 238–246. https://doi.org/10.1590/0047-2085000000084

Berchtold, A. (2016). Test–retest: Agreement or reliability? *Methodological Innovations*, *9*, 205979911667287. https://doi.org/10.1177/2059799116672875

Blaskey, L. G., Harris, L. J., & Nigg, J. T. (2008). Are sensation seeking and emotion processing related to

or distinct from cognitive control in children with ADHD? *Child Neuropsychology*, *14*(4), 353–371. https://doi.org/10.1080/09297040701660291

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Sciences*, *8*(12), 539–546. https://doi.org/10.1016/j.tics.2004.10.003

Brands, A. M. A., Biessels, G. J., De Haan, E. H. F., Kappelle, L. J., & Kessels, R. P. C. (2005). Effects of type 1 diabetes on cognitive performance. *Diabetes Care*, *28*(3), 726–735.

Capella, M., & Andrade, F. (2017). Hacia una psicología ecuatoriana: Una argumentación intergeneracional sobre la importancia de la cultura y la glocalidad en la investigación. *Teoría y Crítica de La Psicología*, *9*, 173–195.

Capella, M., Jadhav, S., & Moncrieff, J. (2019). History, violence and collective memory: Implications for mental health in Ecuador. *Transcultural Psychiatry*, 1–20. https://doi.org/10.1177/1363461519834377

Caruso, J. C. (2004). A comparison of the reliabilities of four types of difference scores for five cognitive assessment batteries. *European Journal of Psychological Assessment*, *20*(3), 166–171. https://doi.org/10.1027/1015-5759.20.3.166

Cheyne, J. A., Solman, G. J. F., Carriere, J. S. A., & Smilek, D. (2009). Anatomy of an error: A bidirectional state model of task engagement/disengagement and attention-related errors. *Cognition*, *111*(1), 98–113. https://doi.org/10.1016/j.cognition.2008.12.009

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Dahlin, E., Neely, A. S., Larsson, A., Bäckman, L., & Nyberg, L. (2008). Transfer of learning after updating training mediated by the striatum. *Science*, *320*(5882), 1510–1512. https://doi.org/10.1126/science.1155466

Dalgleish, T. (1995). Performance on the emotional stroop task in groups of anxious, expert, and control subjects: A comparison of computer and card presentation formats. *Cognition and Emotion*, *9*(4), 341–362. https://doi.org/10.1080/02699939508408971

De Zeeuw, P., Aarnoudse-Moens, C., Bijlhout, J., König, C., Post Uiterweer, A., Papanikolau, A., … Oosterlaan, J. (2008). Inhibitory performance, response speed, intraindividual variability, and response accuracy in ADHD. *Journal of the American Academy of Child and Adolescent Psychiatry*, *47*(7), 808–816. https://doi.org/10.1097/CHI.0b013e318172eee9

Diamond, A. (2013). Executive functions. *The Annual Review of Psychology*, *64*, 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Dias, N. M., Gomes, C. M. A., Reppold, C. T., Bastos, A. C. M. F., Pires, E. U., Carreiro, L. R. R., & Seabra, A. G. (2015). Investigação da estrutura e composição das funções executivas: Análise de modelos teóricos. *Psicologia - Teoria e Prática*, *17*(2), 140–152. https://doi.org/10.15348/1980-6906/psicologia.v17n2p140-152

Duncan, G. J., & Brooks-Gunn, J. (2000). Family poverty, welfare reform, and child development. *Child Development*, *71*(1), 188–196.

Ebert, K. D., & Kohnert, K. (2011). Sustained attention in children with primary language impairment: A meta-analysis. *Journal of Speech, Language, and Hearing Research*, *54*(5), 1372–1384. https://doi.org/10.1044/1092-4388(2011/10-0231)

Ecker, U. K. H., Lewandowsky, S., Oberauer, K., & Chee, A. E. H. (2010). The components of working memory updating: An experimental decomposition and individual differences. *Journal of Experimental Psychology: Learning Memory and Cognition*, *36*(1), 170–189. https://doi.org/10.1037/a0017891

Epp, A. M., Dobson, K. S., Dozois, D. J. A., & Frewen, P. A. (2012). A systematic meta-analysis of the Stroop task in depression. *Clinical Psychology Review*, *32*(4), 316–328. https://doi.org/10.1016/j.cpr.2012.02.005

Fierro Bósquez, M. J., Olabarrieta-Landa, L., Christ, B. R., Arjol, D., Perrin, P. B., Arango-Lasprilla, J. C., & Rivera, D. (2024). Normative data for executive function tests in an Ecuadorian Waranka minority population. *The Clinical Neuropsychologist*, 1-21.

Franzen, M. D., Tishelman, A. C., Sharp, B. H., & Friedman, A. G. (1987). An investigation of the test-retest reliability of the Stroop Color-Word Test across two intervals. *Archives of Clinical Neuropsychology*, *2*, 265–272.

Friedman, N. P., Miyake, A., Corley, R. P., Young, S. E., Defries, J. C., & Hewitt, J. K. (2006). Not all executive functions are related to intelligence. *Psychological Science*, *17*(2), 172–179.

Friedman, N. P., Miyake, A., Young, S. E., DeFries, J. C., Corley, R. P., & Hewitt, J. K. (2008). Individual differences in executive functions are almost entirely genetic in origin. *Journal of Experimental*

Psychology. General, *137*(2), 201–225. https://doi.org/10.1037/0096-3445.137.2.201

Gaertner, L., Sedikides, C., Cai, H., & Brown, J. D. (2010). It's not WEIRD, it's WRONG: When Researchers Overlook uNderlying Genotypes, they will not detect universal processes. *Behavioral and Brain Sciences*, *33*(2–3), 93–94. https://doi.org/10.1017/S0140525X10000105

Garavan, H. (1998). Serial attention within working memory. *Memory and Cognition*, *26*(2), 263–276. https://doi.org/10.3758/BF03201138

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Instituto Nacional de Estadistica y Censos. (2018). *Encuesta nacional de empleo, desempleo y subempleo (ENEMDU)*. Retrieved from https://www.ecuadorencifras.gob.ec/documentos/web-inec/POBREZA/2018/Diciembre-2018/201812_Pobreza.pdf

Kane, M. J., & Engle, R. W. (2003). Working-memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology: General*, *132*(1), 47–70. https://doi.org/10.1037/0096-3445.132.1.47

Kessler, Y., & Oberauer, K. (2014). Working memory updating latency reflects the cost of switching between maintenance and updating modes of operation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(3), 738–754. https://doi.org/10.1037/a0035545

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*, 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kuntsi, J., Stevenson, J., Oosterlaan, J., & Sonuga-Barke, E. J. S. (2001). Test–retest reliability of a new delay aversion task and executive function measures. *British Journal of Developmental Psychology*, *19*(3), 339–348.

Leppink, J., & Pérez-Fuster, P. (2017). We need more replication research – A case for test-retest reliability. *Perspectives on Medical Education*, *6*(3), 158–164. https://doi.org/10.1007/s40037-017-0347-z

Logan, G. D. (1994). On the ability to inhibit thought and action: A users guide to the stop-signal paradigm. *Inhibitory Processes in Attention, Memory, and Language*. https://doi.org/10.1016/j.jsat.2006.09.008

Logan, G. D., Schachar, R. J., & Tannock, R. (1997). Impulsivity and inhibitory control. *Psychological Science*, *8*(1), 60–64.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*(2), 163–203.

Masuda, T., & Nisbett, R. E. (2001). Attending holistically versus analytically: Comparing the context sensitivity of Japanese and Americans. *Journal of Personality and Social Psychology*, *81*(5), 922–934. https://doi.org/10.1037/0022-3514.81.5.922

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review*, *15*(1), 28–50. https://doi.org/10.1177/1088868310366253

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46.

McLoyd, V. C. (1998). Socioeconomic disadvantage and child development. *American Psychologist*, *53*(2), 185–204.

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science*, *21*(1), 8–14. https://doi.org/10.1177/0963721411429458

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "Frontal Lobe" tasks: a latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100.

Morris, N., & Jones, D. M. (1990). Memory updating in working memory: The role of the central executive. *British Journal of Psychology*, *81*, 111–121.

Nisbett, R. E. (2003). *The geography of thought: Why we think the way we do*. New York: Free Press.

Oyserman, D., & Lee, S. W. S. (2008). Does culture influence what and how we think? Effects of priming individualism and collectivism. *Psychological Bulletin*, *134*(2), 311–342. https://doi.org/10.1037/0033-2909.134.2.311.supp

Pardo, J. V., Pardo, P. J., Janer, K. W., & Raichle, M. E. (1990). The anterior cingulate cortex mediates

processing selection in the Stroop attentional conflict paradigm. *Proceedings of the National Academy of Sciences*, *87*, 256–259. https://doi.org/10.1080/00107518308210682

Peirce, J. W. (2008). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, *2*, 10. https://doi.org/10.3389/neuro.11.010.2008

Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of Homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences*, *115*(45), 11401–11405. https://doi.org/10.1073/pnas.1721165115

Rivera, D., Perrin, P. B., Stevens, L. F., Garza, M. T., Weil, C., Saracho, C. P., ... & Arango-Lasprilla, J. C. (2015). Stroop color-word interference test: normative data for the Latin American Spanish speaking adult population. *NeuroRehabilitation*, 37(4), 591-624.

Robertson, I. H., Manly, T., Andrade, J., Baddeley, B. T., & Yiend, J. (1997). "Oops!": Performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. *Neuropsychologia*, *35*(6), 747–758. https://doi.org/10.1016/S0028-3932(97)00015-8

Rodríguez Barreto, L. C., Pulido, N. del C., & Pineda Roa, C. A. (2016). Propiedades psicométricas del Stroop, test de colores y palabras en población colombiana no patológica. *Universitas Psychologica*, *15*(2), 255. https://doi.org/10.11144/Javeriana.upsy15-2.ppst

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Siegrist, M. (1995). Reliability of the stroop test with single-stimulus presentation. *Perceptual and Motor Skills*, *81*(3 Pt 2), 1295–1298.

Siegrist, M. (1997). Test-retest reliability of different versions of the stroop test. *Journal of Psychology: Interdisciplinary and Applied*, *131*(3), 299–306. https://doi.org/10.1080/00223989709603516

Soreni, N., Crosbie, J., Ickowicz, A., & Schachar, R. (2009). Stop Signal and Conners' Continuous Performance Tasks: Test-retest reliability of two inhibition measures in ADHD children. *Journal of Attention Disorders*, *12*(9), 137–143. https://doi.org/10.1177/1087054708326110

St Clair-Thompson, H. L., & Gathercole, S. E. (2006). Executive functions and achievements in school: Shifting, updating, inhibition, and working memory. *Quarterly Journal of Experimental Psychology*, *59*(4), 745–759. https://doi.org/10.1080/17470210500162854

Strauss, G. P., Allen, D. N., Jorgensen, M. L., & Cramer, S. L. (2005). Test-retest reliability of standard and emotional Stroop tasks: An investigation of color-word and picture-word versions. *Assessment*, *12*(3), 330–337. https://doi.org/10.1177/1073191105276375

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Thomas, D. R., & Zumbo, B. D. (2012). Difference scores from the point of view of reliability and repeated-measures ANOVA: In defense of difference scores for data analysis. *Educational and Psychological Measurement*, *72*(1), 37–43. https://doi.org/10.1177/0013164411409929

Toplak, M. E., Bucciarelli, S. M., Jain, U., & Tannock, R. (2009). Executive functions: Performance-based measures and the behavior rating inventory of executive function (BRIEF) in adolescents with attention deficit/hyperactivity disorder (ADHD). *Child Neuropsychology*, *15*(1), 53–72. https://doi.org/10.1080/09297040802070929

United Nations Development Programme. (2018). *Human Development Indices and Indicators: 2018 Statistical Update*. New York. Retrieved from http://hdr.undp.org/sites/default/files/2018_human_development_statistical_update.pdf

Universidad de Guayaquil. (2016). Población Estudiantil – Universidad de Guayaquil. Retrieved September 3, 2019, from http://www.ug.edu.ec/poblacion-estudiantil/

Van Den Heuvel, D. M. J., Ten Dam, V. H., De Craen, A. J. M., Admiraal-Behloul, F., Olofsen, H., Bollen, E. L. E. M., … Van Buchem, M. A. (2006). Increase in periventricular white matter hyperintensities parallels decline in mental processing speed in a non-demented elderly population. *Journal of Neurology, Neurosurgery and Psychiatry*, *77*(2), 149–153. https://doi.org/10.1136/jnnp.2005.070193

Verbruggen, F., Logan, G. D., & Stevens, M. A. (2008). STOP-IT: Windows executable software for the stop-signal paradigm. *Behavior Research Methods*, *40*(2), 479–483. https://doi.org/10.3758/BRM.40.2.479

Weafer, J., Baggott, M. J., & De Wit, H. (2013). Test-retest reliability of behavioral measures of impulsive

choice, impulsive action, and inattention. *Experimental and Clinical Psychopharmacology*, *21*(6), 475–481. https://doi.org/10.1037/a0033659

Wöstmann, N. M., Aichert, D. S., Costa, A., Rubia, K., Möller, H. J., & Ettinger, U. (2013). Reliability and plasticity of response inhibition and interference control. *Brain and Cognition*, *81*(1), 82–94. https://doi.org/10.1016/j.bandc.2012.09.010.