Ciencias Psicológicas
ISSN: 1688-4094
ISSN: 1688-4221
Facultad de Psicología. Universidad Católica del Uruguay.

Correa-Rojas, Jossué
Coeficiente de Correlación Intraclase: Aplicaciones para estimar la estabilidad temporal de un instrumento de medida
Ciencias Psicológicas, vol. 15, núm. 2, e2318, 2021
Facultad de Psicología. Universidad Católica del Uruguay.

DOI: https://doi.org/10.22235/cp.v15i2.2318

Disponible en: https://www.redalyc.org/articulo.oa?id=459569568011

ISSN en línea 1688-4221 *Ciencias Psicológicas* julio-diciembre 2021; 15(2): e-2318
doi: https://doi.org/10.22235/cp.v15i2.2318

Coeficiente de Correlación Intraclase: Aplicaciones para estimar la estabilidad temporal de un instrumento de medida

Intraclass Correlation Coefficient: Applications to estimate the temporal stability of a measuring instrument

Coeficiente de correlação intraclasse: Aplicações para estimar a estabilidade temporal de um instrumento de medição

Jossué Correa-Rojas [1], *ORCID 0000-0002-4166-7210*
[1] *Universidad Peruana de Ciencias Aplicadas, Perú*

***Resumen*:** El presente artículo expone aspectos teóricos y prácticos acerca del uso del Coeficiente de Correlación Intraclase (C_{CI}), se describen sus ventajas respecto al coeficiente producto momento de Pearson para determinar la estabilidad temporal de las puntuaciones de un instrumento de medida. Este trabajo de investigación corresponde a un artículo metodológico. Para la aplicación del método se seleccionaron intencionalmente 42 estudiantes universitarios, en su mayoría mujeres (53.4 %), con edades entre los 17 y 26 años. Se les administró el Índice de Reactividad Interpersonal (IRI), luego de tres semanas se realizó el retest. Los resultados muestran la versatilidad del C_{CI} para proporcionar información respecto al *r* de Pearson. Asimismo, se encontró que en todos los casos el coeficiente *r* Pearson sobreestima ligeramente la estabilidad de las puntuaciones del IRI. Se concluye que el C_{CI} reporta valores estables y menos sesgados para determinar las evidencias de estabilidad temporal de un instrumento de medida.

Palabras clave: fiabilidad; estabilidad temporal; correlación; medidas repetidas; ANOVA.

***Abstract*:** This article presents theoretical and practical aspects about the use of the Intraclass Correlation Coefficient (ICC); it describes its advantages with respect to the Pearson's product-moment coefficient to determine the temporal stability of the scores of a measurement instrument. This research work corresponds to a methodological article. For the application of the method, 42 university students were intentionally selected, mostly women (53.4 %), aged between 17 and 26 years. The Interpersonal Reactivity Index (IRI) was administered; after three weeks the retest was performed. The results show the versatility of the ICC to provide information regarding Pearson's *r*. Likewise, it was found that in all cases the Pearson *r* coefficient slightly overestimates the stability of the IRI scores. It is concluded that the ICC reports stable and less-biased values to determine the evidence of temporal stability of a measurement instrument.

***Keywords**:* reliability; temporal stability; correlation; repeated measurements; ANOVA.

***Resumo*:** Este artigo apresenta aspectos teóricos e práticos sobre o uso do Coeficiente de Correlação Intraclasse (CCI), são descritas suas vantagens em relação ao coeficiente produto momento de Pearson para determinar a estabilidade temporal das pontuações de um instrumento de medição. Este trabalho de pesquisa corresponde a um artigo metodológico. Para a aplicação do método, foram selecionados intencionalmente 42 estudantes universitários, em sua maioria mulheres (53,4 %), com idades entre 17 e 26 anos. Foi administrado o Índice de Reatividade Interpessoal (IRI), após três semanas foi realizado o reteste. Os resultados demostram a versatilidade do CCI para proporcionar informações a respeito do r de Pearson. Da mesma forma, verificou-se que em todos os casos o coeficiente r de Pearson superestima ligeiramente a estabilidade das pontuações do IRI. Conclui-se que o CCI relata valores estáveis e menos enviesados para determinar as evidências de estabilidade temporal de um instrumento de medição.

***Palavras-chave**:* confiabilidade; estabilidade temporária; correlação; medidas repetidas; ANOVA.

Recibido: 10/10/2020| Aceptado: 19/10/2021

Cómo citar:
Correa-Rojas, J. (2021). Coeficiente de Correlación Intraclase: Aplicaciones para estimar la estabilidad temporal de un instrumento de medida. *Ciencias Psicológicas*, *15*(2), e-2318. doi: https://doi.org/10.22235/cp.v15i2.2318

Correspondencia: Jossué Correa-Rojas, Universidad Peruana de Ciencias Aplicadas, Perú. E-mail: jossue.correa@upc.pe

En los últimos años las medidas de acuerdo han cobrado popularidad en la investigación psicológica, específicamente en el campo de la psicometría; ellas se utilizan sobre todo para estimar las evidencias de validez y fiabilidad (Muñiz, 2018). Así, entre los coeficientes más utilizados, se encuentran el índice de acuerdo de Guilford (1954), el coeficiente de Kappa (Cohen, 1960), el coeficiente de Lawshe (1975), el índice de congruencia (Rovinelli & Hambleton, 1977), la prueba binomial (Siegel, 1980), el coeficiente de validez (Aiken, 1980; 1985), el índice de congruencia (Hambleton, 1984), el índice de escalamiento multidimensional (Sireci & Geisinger, 1992) y el coeficiente de validez de contenido (Hernández-Nieto, 2011).

Estos coeficientes son efectivos para analizar la concordancia entre observadores, cuando el nivel de medida es categórico, situación que es bastante usual cuando se utiliza el procedimiento de juicio de expertos (Martínez, 2005; Muñiz, 2018). Dicho de otra forma, estos coeficientes permiten cuantificar una evaluación cualitativa de *n* evaluadores que expresan su punto de vista acerca de la calidad de los ítems que componen una prueba, dichas valoraciones son cuantificadas en un formato de respuestas que aborda aspectos como el dominio, relevancia y representatividad de estos reactivos respecto a un constructo

subyacente (American Educational Research Association [AERA], American Psychological Association [APA] & National Council on Measurement In Education [NCME], 2018).

Cabe mencionar que la razón principal por la que estos coeficientes se han popularizado radica en la sencillez de su cálculo y en la fácil interpretación de sus valores (Bartko, 1994; Benavente, 2009). No obstante, a pesar de la cantidad de coeficientes con los que se cuenta, existe cierta preferencia de parte de los investigadores por utilizar el coeficiente V_{Aiken} (Aiken, 1980; 1985; Merino & Livia, 2009; Pedrosa, Suárez-Álvarez & García-Cueto, 2014), aunque muchas veces en su uso no se recogen los aspectos de calidad mencionados y solo se recogen valoraciones superficiales como *De acuerdo* o *En Desacuerdo*.

A diferencia de estos coeficientes, existe otro conjunto de ellos que permiten el análisis de variables cuantitativas (escala de intervalo); es decir, con puntajes directos (Livia & Ortiz, 2014). Con estas puntuaciones también es posible analizar las evidencias de validez y fiabilidad por medio de distintos procedimientos. Así, por ejemplo, al reportar las evidencias de validez basada en relación con otras variables, esta se suele reportar por medio de la aplicación de diferentes coeficientes de correlación (Martínez, 2005; Muñiz, 2018), dentro de los más conocidos destaca el uso del coeficiente de correlación producto momento de Pearson y la matriz multirasgo-multimétodo (Rodríguez-Miñón, Moreno & Sanjuán, 2000).

Asimismo, para estimar las evidencias de fiabilidad de una medida se pueden emplear diferentes métodos, entre ellos la consistencia interna, formas paralelas y la estabilidad temporal, este último también se denomina *test-retest*, y con él se obtiene la concordancia de las puntuaciones de una medida. Para estos casos, se suele recurrir al uso del coeficiente de correlación producto momento de Pearson (Martínez, 2005), a pesar de los inconvenientes que puede traer su uso (Shrout & Fleiss, 1979).

Coeficiente de Correlación Intraclase (C_{CI})

En lo referente a las evidencias de fiabilidad, uno de los métodos más utilizados es la consistencia interna (Cascaes da Silva et al., 2015; Ledesma, Molina & Valero Mora, 2002). Dentro de los coeficientes con los que se trabaja en este método, se destaca el uso del coeficiente Alfa (Livia & Ortiz, 2014; Muñiz, 2010), que ha recibido críticas debido al incumplimiento de los supuestos requeridos para su aplicación (Domínguez & Merino, 2015; Ventura-León, 2018), como por ejemplo el supuesto *tau-equivalencia*, requerido para estimar coeficientes alfa por dimensiones (Raykov, 1997). Por ello, la literatura especializada sugiere el uso de otros coeficientes, como el Omega (Ventura-León, 2017; Viladrich, Angulo-Brunet & Doval, 2017) o el coeficiente de fiabilidad compuesta (Hair, Anderson, Tatham & Black, 2010), que arrojan estimaciones menos sesgadas.

No obstante, existen otros procedimientos para demostrar la fiabilidad de un instrumento. Por ejemplo, la estabilidad temporal, menos popular que la consistencia interna, pero no menos importante. Este método hace alusión a la concordancia de la puntuación en dos momentos diferentes en el tiempo (Muñiz, 2010; 2018). Este procedimiento también es conocido como *test-retest*. Las aplicaciones del procedimiento suelen recurrir al cálculo del coeficiente de correlación producto momento de Pearson (*r*), con el cual es posible verificar la relación entre las dos mediciones, aunque generalmente este valor sea sobreestimado

(Spence-Laschinger, 1992) debido a la naturaleza lineal del coeficiente (Shrout & Fleiss, 1979).

El uso de este coeficiente implica una limitación importante, ya que si un instrumento mide sistemáticamente momentos diferentes uno del otro la correlación puede ser perfecta, a pesar de que la concordancia sea nula (Pita & Pértegas, 2004). Por este motivo, el uso del coeficiente de Pearson puede constituir una fuente de error en la medición, ya que se omite en el cálculo la variabilidad intra e inter sujeto (Shrout & Fleiss, 1979), exponiendo al investigador a errores sistemáticos en sus interpretaciones (Bartko, 1994; Ledesma et al., 2002).

Para resolver esto, desde la teoría de la generalizabilidad (TG) se ofrece un desarrollo teórico profundo acerca de la fiabilidad, definiéndola como la proporción de la varianza de un puntaje observado, que no es atribuible a errores en la medición (Spence-Laschinger, 1992), con lo cual se alienta a especificar y estimar los componentes de varianza de puntaje verdadero, varianza de puntaje de error y varianza de puntaje observado, y a calcular coeficientes basados en estas estimaciones (Mandeville, 2005; Pita & Pértegas, 2004). Desde este enfoque, se sugiere considerar el uso del C_{CI} para determinar la concordancia entre dos mediciones realizadas en un intervalo de tiempo (Esquivel et al., 2006; Koo & Li, 2016; Mandeville, 2005; Shrout & Fleiss, 1979; Weir, 2005). A diferencia de otros coeficientes, el C_{CI} permite detectar el sesgo sistemático de la medición (Esquivel et al., 2006), además de verificar la estabilidad temporal de las puntuaciones (Martínez, 2005; Muñiz, 2018).

En este punto se hace necesario revisar la complejidad de la definición de fiabilidad, pues ella contempla la relación de la varianza entre el puntaje verdadero respecto de la varianza de puntaje total (AERA, APA & NCME, 2018), esta definición resulta importante cuando el objetivo del estudio tiene que ver con determinar la consistencia interna (Vargha, 1997). Sin embargo, cuando se pretende medir la concordancia de las puntuaciones de un instrumento de medida en dos momentos en el tiempo sobre una muestra sin alterar, la literatura científica no sugiere un procedimiento específico (Muñiz, 2018) y la razón principal tiene que ver con la escala de medida, tratándose para la estabilidad temporal de medidas continuas (Benavente, 2009; Mandeville, 2005).

En este marco, el cálculo de la fiabilidad a través de la estabilidad temporal (test-retest) no es el procedimiento al cual se recurra comúnmente (Camacho-Sandoval, 2008; Pita & Pértegas, 2004; Prieto, Lamarca & Casado, 1998), ello no significa que su estimación sea irrelevante. Ello responde más bien a aspectos de conveniencia. Pues en el método test-retest se busca constatar que la variabilidad de las puntuaciones no difieren significativamente entre sí (Weir, 2005). Sin embargo, cuando los puntajes asignados difieren consistentemente entre cada observación es necesario recurrir a métodos de cálculo más sofisticados que permitan reducir el error de la medición. Uno de los procedimientos sugeridos es el cálculo de coeficientes de correlación producto de los residuos resultantes de un ANOVA de medidas repetidas (Cerda & Villarroel, 2008; Koo & Li, 2016; Shieh, 2016).

Originalmente el C_{CI} fue desarrollado por Fisher (1954) como una modificación del coeficiente de correlación de Pearson. Así, el C_{CI} actual se calcula a partir de la media de cuadrados producto de un análisis de varianza de medidas repetidas y es ampliamente utilizado en otras disciplinas (Cortés, Rubio & Gaitán, 2010; Koo & Li, 2016) para evaluar la validez y fiabilidad de los instrumentos de medición.

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y}$$

Donde:
σ_{XY} es la covarianza de (X, Y)
σ_X es la desviación estándar de la variable X
σ_Y es la desviación estándar de la variable Y

Los aspectos más importantes para aconsejar el uso del C_{CI} en la investigación psicológica son el hecho de que considera el error de medición necesario para poder controlar el sesgo (Camacho-Sandoval, 2008) y la variabilidad intra e inter sujeto (Hazra & Gogtay, 2016). Lo cual muestra sus beneficios en comparación con coeficientes como Pearson o Spearman (Esquivel et al., 2006). Al respecto, Abad, Olea, Ponsoda y García (2011) señalan que al descomponer la variabilidad de los datos en función de las fuentes de error se estiman los correspondientes componentes de la varianza, estos elementos refieren a una estimación de la variabilidad atribuida a los sujetos, ítems y la residual. Por lo tanto, el cálculo del C_{CI} constituye una estimación más precisa y menos sesgada. Asimismo, en términos de componentes de varianza el C_{CI} se obtiene de la siguiente manera:

σ_s^2: Variabilidad intersujeto (atribuible a las diferencias entre los sujetos, *s*)
σ_j^2: Variabilidad intrasujeto (se refiere a las diferencias de las mediciones de un mismo sujeto, *j*)
σ_e^2: Variabilidad residual (variabilidad aleatoria asociada a los errores de medición, *e*)

$$C_{CI} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_j^2 + \sigma_e^2}$$

De acuerdo con Shrout y Fleiss (1979), el C_{CI} expresa cantidades únicas de la magnitud relativa de los dos componentes de varianza de una puntuación. A medida que disminuye la proporción de la varianza del error de la varianza total en un conjunto de puntajes, los valores posibles del C_{CI} oscilan entre 0 y 1 (Manterola et al., 2018; Müller & Büttner, 1994). Donde una gran proporción de varianza de error en un conjunto de puntajes produce un coeficiente C_{CI} bajo e indica poca confiabilidad (Turner & Carlson, 2003). Asimismo, señalan que el valor mínimo aceptable para el ICC es .75 (Haggard, 1958; Shrout & Fleiss, 1979). Al respecto, Prieto et al. (1998) modificaron el cálculo del C_{CI} a partir de la variabilidad de los puntajes observados: cuanto más homogénea sea la muestra de estudio este tiende a ser más bajo.

De acuerdo con la TG una aproximación a la medida de la varianza del error se puede obtener descomponiendo la variabilidad de los datos a partir de cada fuente de variación; se estiman así los elementos de la varianza (variabilidad atribuida al sujeto, a los ítems y al error de medición). La aplicación del ANOVA permite realizar estas estimaciones. Para esto es necesario definir la cantidad de niveles de la variable intrasujeto (cantidad de medidas realizadas en un periodo de tiempo). Entre los resultados se seleccionan las sumas de cuadrados (*SC*), grados de libertad (*gl*) y medias cuadráticas (*MC*), con los cuales es posible realizar el cálculo del C_{CI}.

$$Cci = \frac{k + SC\ entre - SS\ total}{(k-1)\ SS\ total}$$

Donde:
k: Numero de medidas
SC entre: Media cuadrática de los puntajes
SS total: Suma de error cuadrático medio de one-way ANOVA

Se ha expuesto acerca de la conveniencia y ventajas del C_{CI} en relación con otros coeficientes de correlación (concordancia). A continuación, se presenta una aplicación del $C_{CI,}$ la misma tiene como objetivo determinar la estabilidad temporal de las puntuaciones del Índice de Reactividad Interpersonal (IRI) en una muestra de universitarios limeños, se compara el procedimiento tradicional que mide la concordancia de las medidas mediante el coeficiente de correlación de Pearson y el procedimiento sugerido a través del C_{CI} que proviene de un ANOVA de medidas repetidas.

Método

Participantes

Para realizar una demostración acerca de la aplicabilidad del C_{CI} se seleccionaron intencionalmente 41 estudiantes de universidades públicas (67.6 %) y privadas (32.4 %), en su mayoría mujeres (53.4 %), con edades que oscilaron entre los 17 y 26 años. El criterio de selección de los estudiantes responde a su accesibilidad, asistencia regular a clases y aprobación del consentimiento informado. Todos los estudiantes presentaron un nivel cultural y socioeconómico medio.

Instrumento

Se utilizó el Índice de Reactividad Interpersonal de Davis (1983). Este instrumento de autoaplicación, de lápiz y papel, evalúa la empatía cognitiva y afectiva (Esteban-Guitart, Rivas & Pérez, 2012), mediante un formato de respuesta tipo Likert con cinco opciones: *No me describe bien* (1), *Me describe un poco* (2), *Me describe bien* (3), *Me describe bastante bien* (4) y *Me describe muy bien* (5). Consta de 28 ítems que permiten medir las diferencias individuales del constructo empatía mediante las siguientes cuatro subescalas (7 ítems cada una): toma de perspectiva y fantasía (componente cognitivo) y preocupación empática y malestar personal (componente emocional). Para la presente investigación se ha empleado la adaptación española de Mestre, Frías y Samper (2004), que mantiene la estructura de los ítems en cada una de las categorías de la versión original.

Procedimiento

La administración de los instrumentos fue realizada en los meses de abril y mayo del 2020, las mediciones se llevaron de forma individual, como se trata de una medición longitudinal (dos medidas), se trató de que las mediciones se realicen en condiciones similares (día y hora) y dejando un lapso de tres semanas. Se consideraron las recomendaciones y normativas para la aplicación de pruebas propuestas por la International Test Commission (2000), con el objetivo de minimizar la varianza irrelevante al constructo proclive a ocurrir durante la administración de pruebas psicológicas. Previo a la administración de las pruebas, los participantes firmaron el consentimiento informado, en el

que se dio a conocer el carácter voluntario del estudio, la liberad de su participación, la ausencia de daño físico y psicológico, el anonimato y la confidencialidad de la información recabada. De esta manera, se respetaron los lineamientos éticos según los derechos de Helsinki acoplándose además al Código de Ética del Perú (Colegio del Psicólogo del Perú, 2017).

Análisis de datos

El análisis estadístico se realizó mediante una sintaxis desarrollada para el software IBM SPSS versión 25. El análisis de datos se realizó por etapas, inicialmente se exploraron los estadísticos descriptivos y distribucionales de los ítems. Así, el supuesto de normalidad univariada se evaluó mediante los coeficientes de asimetría y curtosis, considerando como criterio los valores dentro del rango de ± 1.5 (Pérez & Medrano, 2010). Posteriormente, se aplicó el procedimiento test-retest, la concordancia de las puntuaciones se analizó por medio del coeficiente de correlación producto momento de Pearson (*r*), los criterios para su interpretación se basaron en las sugerencias de Cohen quien señala que este es en sí mismo un tamaño de efecto (Cohen, 1992). La segunda estimación test-retest se realizó a través de un ANOVA de medidas repetidas en donde se definieron dos niveles. Este procedimiento también permitió verificar las variaciones intra e inter sujeto, se asumieron diferencias estadísticamente significativas $\alpha \leq .05$. Los resultados hacen referencia a la variabilidad de la medición en el mismo sujeto y en el segundo caso a la variabilidad entre la respuesta de un participante en relación con las otras personas. Se ha incluido una sintaxis mediante la cual se puede reproducir, debido a que en esta oportunidad, lo que se busca es identificar el acuerdo absoluto. Las variaciones de sujeto a sujeto se evalúan mediante un estadístico *F* con su respectiva significancia estadística y además el tamaño de efecto (eta parcial al cuadrado [η_p^2]), asumiendo los criterios de Cohen para su interpretación (Cohen, 1992). Además, se añade la variabilidad inter sujetos (las variaciones del sujeto con otro sujeto) con un estadístico *F* con su respectiva significancia estadística y además el tamaño de efecto (eta parcial al cuadrado [η_p^2]), asumiendo los criterios de Cohen para su interpretación (Cohen, 1992).

Resultados

En la tabla 1 se presentan las medidas descriptivas para TP, F, CE y M para dos medidas reportadas con un margen de tres semanas. Los resultados muestran que los promedios de TP evidencian poca variación (M_1 = 20.710 y M_2 = 20.120), las medias de F muestran un comportamiento similar (M_1 = 18.900 y M_2 = 17.760), en cuanto a CE las medidas resultan bastante parecidas (M_1 = 25.370 y M_2 = 23.220). Asimismo, los promedios *M* muestran el mismo estado (M_1 = 15.020 y M_2 =16.170). Finalmente, los coeficientes de asimetría y curtosis se encuentran por debajo de 1.5, lo que sugiere que las variables presentan normalidad univariada.

Tabla 1.
Estadísticos descriptivos

Variables	Medida 1			Medida 2		
	$M(DE)_1$	*g1*	*g2*	$M(DE)_2$	*g1*	*g2*
Toma de perspectiva	20.71(4.18)	-0.06	0.29	20.12(4.64)	-0.21	-0.49
Fantasía	18.90(4.19)	0.09	-0.17	17.76(4.65)	0.06	-0.61
Comprensión empática	25.37(3.81)	0.88	0.36	23.22(5.44)	0.11	-0.41
Malestar	15.02(4.05)	0.29	0.02	16.17(4.24)	0.52	-0.27

Notas: *M*: Media; *DE*: Desviación estándar; g_1: Coeficiente de asimetría; g_2: Coeficiente de curtosis.

Análisis de varianzas

En la tabla 2 se muestran los resultados del ANOVA de medidas repetidas para dos factores. Los resultados de la dimensión *F* muestran que a nivel intrasujeto no se encontraron diferencias estatistamente significativas y el tamaño de efecto es inexistente ($F = .531$; $p > .05$; η_p^2=0.013). Sin embargo, en la prueba de efecto intersujeto las variaciones son estadísticamente significativas y la magnitud de las diferencias es grande ($F = 1327.275$; $p < .001$; $\eta_p^2 = 0.971$). En lo referente a la dimensión F, la prueba de efecto intrasujeto arroja que no existen diferencias estadísticamente significativas y el tamaño de efecto no resulta importante ($F = 2.832$; $p > .05$; $\eta_p^2 = 0.066$). Mientras que la prueba de efecto intersujeto indica que las variaciones individuo-grupo son estadísticamente significativas y la magnitud de estas es grande ($F = 928.659$; $p < .001$; $\eta_p^2 = 0.959$). Los resultados en CE indican que no existen diferencias estadísticamente significativas intrasujeto, alcanzando un tamaño de efecto muy pequeño ($F = 9.156$; $p > .05$; $\eta_p^2 = 0.186$). Sin embargo, sí se encontraron diferencias estadísticamente significativas a nivel intersujeto, siendo la magnitud de estas grande ($F = 1327.275$; $p < .001$; $\eta_p^2 = 0.973$). Por último, la dimensión *M*, los resultados a nivel intrasujeto reflejan que no existen diferencias estadísticamente significativas y el tamaño de efecto no resulta importante ($F = 3.800$; $p > .05$; $\eta_p^2 = 0.087$). Mientras que la prueba de efecto intersujeto indica que las variaciones son estadísticamente significativas y la magnitud de estas es grande ($F = 729.928$; $p < .001$; $\eta_p^2 = 0.948$).

Tabla 2.
Prueba de efecto intra e inter sujeto

Variables	*MC (gl)*	*F*	*p*	η_p^2
Toma de Perspectiva (TP)				
Prueba de efecto intrasujeto				
factor1	14.049(1)	.531	.470	.013
Error (factor1)	26.449(40)			
Prueba de efecto intersujeto				
Intersección	17087.049(1)	1327.275	.000	.971
Error	12.874(40)			
Fantasía (F)				
Prueba de efecto intrasujeto				
factor1	14.049(1)	2.832	.100	.066
Error (factor1)	26.449(40)			
Prueba de efecto intersujeto				
Intersección	17087.049(1)	928.659	.000	.959
Error	12.874(40)			
Comprensión Empática (CE)				
Prueba de efecto intrasujeto				
factor1	94.439(1)	9.156	.004	.186
Error (factor1)	412.561(40)			
Prueba de efecto intersujeto				
Intersección	48391.024(1)	1327.275	.000	.973
Error	1351.976(40)			
Malestar (M)				
Prueba de efecto intrasujeto				
factor1	26.939(1)	3.800	.470	.087
Error (factor1)	283.561(40)			
Prueba de efecto intersujeto				
Intersección	19943.280(1)	729.928	.000	.948
Error	1093.220(40)			

Nota: *MC*: Media Cuadrática; *F*: Estadístico de Prueba ANOVA medidas repetidas; *p*: Significancia estadística; η_p^2: Eta Parcial al Cuadrado. Prueba de Efecto Intra-Sujeto: Evalúa la variabilidad de las mismas medidas en las personas. Prueba de Efecto Inter-Sujeto: Evalúa la variabilidad entre las mismas medidas entre las personas.

Estabilidad temporal de la medida

A partir del procedimiento de ANOVA de medidas repetidas, se obtuvieron la *MC*: media cuadrática de los puntajes y la *MSE*: la suma de errores cuadráticos medio de *one-way* elementos necesarios para el cálculo del $C_{CI,}$ con sus respectivos intervalos de confianza al 95 %. Asimismo, se presentan los coeficientes de correlación producto momento de Pearson (*r*) con la respectiva significancia estadística (tabla 3). Se comparan los coeficientes C_{CI} - *r*, de ellos se calculó el delta entre estos coeficientes obteniéndose cambios por encima de .001.

Tabla 3.
Comparativo entre los coeficientes producto momento de Pearson y C_{CI}

Test-retest	*n*	*r*	C_{ci}	$\Delta\ C_{ci}$ - *r*
Toma de perspectiva	41	.323*	.324 [.020-.572]	0.001
Comprensión empática	41	.567**	.487 [.200-.694]	-0.080
Fantasía	41	.517**	.503 [.242-.699]	-0.014
Malestar	41	.589**	.572 [.327-.746]	-0.017

Nota: $\Delta\ C_{ci}$ - *r*: Cambio entre los coeficientes; *$p < .05$, **$p < .001$.

Discusión

El C_{CI} es un índice de concordancia para datos continuos, evalúa el tamaño de los componentes de la varianza entre los grupos y dentro de éstos (Davis & Joseph, 2016; Shoukri, 2004). Asimismo, describe la proporción de la variación total, la cual es explicada por las diferencias entre las puntuaciones e instrumentos (Mandeville, 2005). Según Hazra y Gogtay (2016), el C_{CI} se desarrolla dentro del análisis de varianza y su cálculo se basa en la varianza verdadera (entre sujetos) y la varianza del error de medición, producida durante la medición repetida (Hazra & Gogtay, 2016; Manterola et al., 2018).

En tal sentido, la presente investigación tuvo como propósito realizar una revisión teórica acerca de la aplicabilidad del C_{CI} para estimar la estabilidad temporal de las puntuaciones de los instrumentos de medida. Para ello, se dirigió un estudio longitudinal de dos mediciones sobre las puntuaciones del IRI, las mismas que luego fueron analizadas desde una perspectiva tradicional mediante un análisis bivariado con el coeficiente de correlación de Pearson. Mientras en el segundo enfoque el análisis comprende un análisis de varianza de medidas repetidas (ANOVA).

Cabe mencionar que la evidencia de fiabilidad por el método de estabilidad temporal (test-retest) ya ha sido utilizada en el análisis psicométrico del IRI, encontrándose en estudios en población española (Carrasco, Delgado, Barbero, Holgado & Del Barrio, 2011), belga (De Corte et al., 2007) y chilena (Fernández, Dufey & Kramp, 2011), en cuyos casos se logró constatar una correlación test-retest entre moderada y alta.

Por otro lado, los coeficientes de correlación producto momento de Pearson indican que existe relación entre estas puntuaciones. Sin embargo, ello no indica que exista concordancia entre las medidas, lo cual ya ha sido bastante discutido en la literatura (Davis & Joseph, 2016; Koo & Li, 2016; Shoukri, 2004); además, al tratarse de un procedimiento de cálculo lineal las interpretaciones son parciales y existe el riesgo de sobre estimación (Hazra & Gogtay, 2016; Manterola et al., 2018). Por su parte, el ANOVA de medidas repetidas provee los insumos para el cálculo del C_{CI}, el cual por su naturaleza no lineal constituye una medida ajustada de la concordancia entre las mediciones. Con ello, se identificó que las cuatro dimensiones del IRI (TP, CE, F y M) no presentan mayor diferencia en las puntuaciones dentro del grupo (intrasujeto) apreciándose diferencias no significativas con magnitudes de efecto inexistentes. Sin embargo, al analizar las variaciones entre grupos se pudo apreciar que sí existían diferencias estadísticamente significativas, con tamaños de efecto grandes.

Con ello, se pudo corroborar la utilidad práctica del cálculo del $C_{CI,}$ pues no solo brinda información acerca de la relación entre las dos medidas, sino que también brinda información sobre el cumplimiento de los supuestos de no variaciones intra e intergrupos. Los cuales permiten la estimación del error de medición (Pita & Pértegas, 2004).

Asimismo, al comparar los coeficientes de Pearson y C_{CI} se pudo apreciar que los primeros son ligeramente superiores; asimismo, se interpretan como correlaciones significativas y muy significativas, pero esto no implica que se han analizado las varianzas y por ende no se está evaluando la concordancia en sí. Lo que este coeficiente expresa es la relación producto momento entre dos mediciones. Desconociendo la variación a nivel inter e intra sujeto (Shoukri, 2004; Shrout & Fleiss, 1979).

Adicionalmente, para evaluar si los cambios entre los coeficientes de correlación eran significativos, se calcularon los diferenciales (Δ) y se consideró el criterio de Byrne (2008) para determinar la invarianza de la medición. Se puede apreciar que con excepción de la dimensión TP, en las restantes estas diferencias son significativas, lo cual evidencia la sobre estimación que suele ocurrir al emplear el coeficiente de correlación de Pearson como estadístico de concordancia.

En cuanto al método de estimación utilizado, es importante recalcar que el procedimiento test-retest ha sido previamente utilizado en otros estudios. Como es el caso de la investigación de Carrasco et al. (2011), en donde se analizó la estabilidad temporal del IRI en una muestra de adolescentes españoles, reportándose correlaciones producto momento de Pearson que oscilan entre .44 y .65 después del intervalo de un año. Lo mismo que lo reportado por Fernández et al. (2011), quien encontró correlaciones producto momento de Pearson superiores a .70 luego de un intervalo de 60 días, en universitarios chilenos. Estos estudios denotan que el constructo examinado no está sujeto a fluctuaciones aleatorias (Reidl-Martínez, 2013); por el contrario, parece ser bastante estable en el tiempo. Por otro lado, a pesar de que los intervalos de tiempo utilizados en estos antecedentes son distintos a los de la presente investigación, es necesario recalcar que estos se han establecido en concordancia con los criterios sugeridos por la literatura (Martínez, 2005). Lo expuesto se indica como referencia para resaltar que los hallazgos de la investigación no responden a un comportamiento anómalo del constructo, ni a algún otro aspecto resultante propio de la tarea realizada (Medrano & Pérez, 2019).

Un aspecto importante tiene que ver con la aplicabilidad del procedimiento para el cálculo del $C_{CI,}$ ya que este este no solo se limita a la estimación de la estabilidad temporal de las puntuaciones de un instrumento, siendo posible utilizarlo en estudios cuasiexperimentales (más de una medición). En dichos diseños se emplea comúnmente la *t* relacionada o la suma de rangos de wilcoxon, estimaciones que solo expresan la diferencia puntual entre antes-después y no la variación intra e inter sujeto como producto del efecto de un factor (programa de intervención) (Abad et al., 2011).

Una limitación importante tiene que ver con el tamaño de muestra y el tipo de muestreo, lo cual restringe la capacidad de generalización de los resultados. No obstante, como en este caso lo que se busca es exponer la técnica de análisis, el tamaño de muestra no afecta ello. Asimismo, es necesario demostrar la aplicabilidad del C_{CI} en otros procedimientos como la validación por juicio de expertos, en cuyo caso se esperaría que se demuestre que arroje estimaciones más precisas que otros coeficientes como la $V_{Aiken.}$

Finalmente, es importante recalcar que estudios psicométricos recientes incluyen dentro de sus medidas de fiabilidad al procedimiento test-retest o la estabilidad temporal de la medida (Correa-Rojas, Grimaldo & Del Rosario-Gontaruk, 2020; Lascurain, Lavandera & Manzanares, 2017), esto como complemento a la consistencia interna; lo cual se hace necesario sobre todo si se pretende dar uso a estas medidas en estudios longitudinales (Abad et al., 2011; Muñiz, 2018) para garantizar que estas no constituyen una fuente de error sistemático.

Referencias

Abad, F., Olea, J., Ponsoda, V. & García, C. (2011). *Medición en ciencias sociales y de la salud*. Editorial Sintesis.

Aiken, L. (1980). Content validity and reliability of single items or questionnaries. *Educational and Psychological Measurement*, *40*(1), 955-959.

Aiken, L. (1985). Three coefficients for analyzing the reliability and validity or ratings. *Educational and Psychological Measurement*, *45*(1), 131-142.

American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement In Education (NCME). (2018). *Estandares para pruebas educativas y psicologicas*. Washington: American Educational Research Association.

Bartko, J. J. (1994). Measures of agreement: A single procedure. *Statistics in Medicine*, *13*(5-7), 737-745. doi: https://doi.org/10.1002/sim.4780130534

Benavente, A. P. (2009). *Medidas de acuerdo y desacuerdo entre jueces*. Universidad de Murcia.

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema*, *20*(4), 872-882.

Camacho-Sandoval, J. (2008). Coeficiente de concordancia para variables continuas. *Acta Médica Costarricense*, *50*(4), 211-212.

Cascaes da Silva, F., Gonçalves, E., Valdivia Arancibia, B. A., Graziele Bento, S., Da Silva Castro, T. L., Soleman Hernandez, S. S. & Da Silva, R. (2015). Estimadores de consistencia interna en las investigaciones en salud: el uso del coeficiente alfa. *Revista Peruana de Medicina Experimental y Salud Pública*, *32*(1), 129. doi: https://doi.org/10.17843/rpmesp.2015.321.1585

Carrasco, M., A., Delgado, B., Barbero, M., Holgado, F. & Del Barrio, M. (2011). Propiedades psicométricas del Interpersonal Reactivity Index (IRI) en población infantil y adolecente española. *Psicothema*, *23*(4), 824-831.

Cerda, J. & Villarroel, L. (2008). Evaluación de la concordancia inter-observador en investigación pediátrica: Coeficiente de Kappa. *Revista Chilena de Pediatria*, *79*(1), 54-58.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37-46. doi: https://doi.org/10.1177/001316446002000104

Cohen, J. (1992). Statistical Power Analysis. *Current Directions in Psychological Science*, *1*(3), 98-101. doi: https://10.1111/1467-8721.ep10768783

Cortés, É., Rubio, J. & Gaitán, H. (2010). Métodos estadísticos de evaluación de la concordancia y la reproducibilidad de pruebas diagnósticas. *Revista Colombiana de Obstetricia y Ginecología*, *61*, 247-255.

Correa-Rojas, J., Grimaldo, M. & Del Rosario-Gontaruk, S. (2020). Propiedades psicometricas d de la Fear of Missing Out Scale en universitarios peruanos. *Revista Aloma*, *28*(2), 113-120. doi: https://doi.org/10.51698/aloma.2020.38.2.113-120

Colegio del Psicólogo del Perú. (2017). *Código de Ética y Deontología*. Lima: Autor.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology, 44*(1), 113-126. doi: https://doi.org/10.1037/0022-3514.44.1.113

Davis, M. D. & Joseph, J. (2016). Determining agreement using rater characteristics. *Journal of Biopharmaceutical Statistics*, *26*(4), 619-630. doi: https://doi.org/10.1080/10543406.2015.1052490

De Corte, K., Buysse, A., Verhofstadt, L. L., Roeyers, H., Ponnet, K. & Davis, M. H. (2007). Measuring Empathic Tendencies: Reliability And Validity of the Dutch Version of the Interpersonal Reactivity Index. *Psychologica Belgica*, *47*(4), 235-260. doi: http://doi.org/10.5334/pb-47-4-235

Domínguez, S. & Merino, C. (2015). Sobre el reporte de confiabilidad del CLARP-TDAH, de Salamanca (2010). *Revista Latinoamericana de Ciencias Sociales, Niñez y Juventud*, *13*(2).

Esquivel, C. G., Velasco, V. M., Martinez, E., Barbachano, E., González, G. & Castillo, C. E. (2006). Coeficiente de correlación intraclase vs correlación de Pearson de la glucemia capilar por reflectometría y glucemia plasmática. *Medicina Interna de Mexico*, *22*(3), 165-171.

Esteban-Guitart, M., Rivas, M. J. & Pérez, M. (2012). Empatía y tolerancia a la diversidad en un contexto educativo intercultural. *Universitas Psychologica, 11*(2), 415-426.

Fernández, A. M., Dufey, M. & Kramp, U. (2011). Testing the psychometric properties of the Interpersonal Reactivity Index (IRI) in Chile: Empathy in a different cultural context. *European Journal of Psychological Assessment, 27*(3), 179-185. doi: https://doi.org/10.1027/1015-5759/a000065

Fisher, R. A. (1954). *Statistical Methods for Research Workers*. Oliver and Boyd.

Guilford, J. P. (1954). *Psychometric methods* (2ª ed). Bombay -New Deli: Tata McGraw-Hill

Hambleton, R. K. (1984). Validating the test scores. En R. A. Berk (Ed.), *A guide to critterion referenced test construction* (pp. 199-230). Baltimore: Johns Hopkins University Press.

Hair, J., Anderson, R., Tatham, R. & Black, W. (2010). *Analisis multivariante* (2ª ed.). Madrid: Pearson Prentice Hall.

Haggard, E. A. (1958). *Intraclass correlation and the analysis of variance.* Holt.

Hazra, A. & Gogtay, N. (2016). Biostatistics series module 6: Correlation and linear regression. *Indian Journal of Dermatology*, *66*(1). 593-601. doi: https://doi.org/10.4103/0019-5154.193662

Hernández-Nieto, R. (2011). *Instrumentos de recolección de datos en ciencias sociales y ciencias biomédicas.* Venezuela: Universidad de Los Andes.

International Test Commission. (2000). *Guidelines on Test Use: Spanish Version*. Recuperado de https://www.intestcom.org/files/guideline_test_use.pdf

Koo, T. K. & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155-163. doi: https://doi.org/10.1016/j.jcm.2016.02.012

Lascurain, P. C., Lavandera, M. C. & Manzanares, E. L. (2017). Propiedades psicométricas de la escala de actitudes sobre el amor (LAS) en universitarios peruanos. *Acta Colombiana de Psicología*, *20*(2), 270-281. doi: https://doi.org/10.14718/ACP.2017.20.2.13

Lawshe, C. H. (1975). *A Quantitative Approach To Content Validity*. *Personnel Psychology*, *28*(4), 563-575. doi: https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Ledesma, R., Molina Ibañez, G. & Valero Mora, P. (2002). Análisis de consistencia interna mediante Alfa de Cronbach: un programa basado en gráficos dinámicos. *Psico-USF*, *3*(7600), 143-152. doi: https://doi.org/10.1590/S1413-82712002000200003

Livia, J. & Ortiz, M. (2014). *Construcción de pruebas. Aplicaciones en ciencias sociales y de la salud*. Lima: UNFV.

Martínez, R. (2005). *Psicometría: Teoria de los test psicologicos y educativos*. Editorial Sintesis Psicología.

Mandeville, P. (2005). El Coeficiente de Correlación Intraclase. *Ciencia UANL*, *8*(3), 414-416.

Manterola, C., Grande, L., Otzen, T., García, N., Salazar, P. & Quiroz, G. (2018). Confiabilidad, precisión o reproducibilidad de las mediciones. Métodos de valoración, utilidad y aplicaciones en la práctica clínica. *Revista Chilena de Infectología*, *35*(6), 680-688. doi: https://doi.org/10.4067/s0716-10182018000600680

Medrano, L. & Pérez, E. (2019). *Manual de psicometría y evaluación psicológica*. Editorial Brujas.

Merino, C. & Livia, J. (2009). Intervalos de confianza asimétricos para el índice la validez de contenido: Un programa Visual Basic para la V de Aiken. *Anales de Psicología*, *25*(1), 169-171. doi: http://revistas.um.es/analesps

Mestre, V., Frías, M. D. & Samper, P. (2004). La medida de la empatía: análisis del Interpersonal Reactivity Index. *Psicothema*, *16*(2), 255-260.

Muñiz, J. (2010). *Teoria Clasica de los Test*. Madrid: Piramide.

Muñiz, J. (2018). *Introducción a la psicometria*. Madrid: Piramide.

Müller, R. & Büttner, P. (1994). A critical discussion of intraclass correlation coefficients. *Statistics in Medicine*, *13*(23-24), 2465-2476. doi: https://doi.org/10.1002/sim.4780132310

Pedrosa, I., Suárez-Álvarez, J. & García-Cueto, E. (2014). Content validity evidences: Theoretical advances and estimation methods. *Acción Psicológica*, *10*(2), 3-18. doi: https://doi.org/10.5944/ap.10.2.11820

Pérez, E. R. & Medrano, L. A. (2010). Análisis factorial exploratorio: bases conceptuales y metodológicas. *Revista Argentina de Ciencias del Comportamiento (RACC)*, *2*(1), 58-66.

Pita, S., & Pértegas, S. (2004). La fiabilidad de las mediciones clínicas: el análsiis de concordancia para variables numéricas. *Atencion Primaria En La Red*, (1995), 1-11.

Prieto, L., Lamarca, R. & Casado, A. (1998). El coeficiente de Correlación Intraclase. *Medicina Clínica*, (October), 142-145.

Raykov, T. (1997). Scale Reliability, Cronbach's Coefficient Alpha, and Violations of Essential Tau-Equivalence with Fixed Congeneric Components. *Multivariate Behavioral Research, 32*(4), 329-353. doi: https://doi.org/10.1207/s15327906mbr3204_2

Rodríguez-Miñón, P., Moreno, E. & Sanjuán, P. (2000). La matriz multimétodo-multirrasgo aplicada al estudio de la sensibilidad. *Psicothema, 12*(2), 492-495.

Rovinelli, R. J. & Hambleton, R. K. (1977). On the Use of Content Specialists in the Assessment of Criterion-Referenced Test Item Validity. *Dutch Journal of Educational Research, 2*, 49-60.

Reidl-Martínez, L. (2013). Confiabilidad en la medición. *Investigación en Educación Médica, 2*(6), 107-111.

Siegel, S. (1980). *Estadísticas no Paramétricas Aplicadas a las Ciencias de la Conducta.* México: Trillas.

Sireci, S. G. & Geisinger, K. F. (1992). Analyzing test content using cluster analysis and multidimen-siamal scaling. *Applied Psychological Measurement, 16*, 17-31.

Shieh, G. (2016). Choosing the best index for the average score intraclass correlation coefficient. *Behavior Research Methods, 48*(3), 994-1003. doi: https://doi.org/10.3758/s13428-015-0623-y

Shoukri, M. (2004). *Measures of interobserver agreement.* Estados Unidos: Chapman & Hall.

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

Spence-Laschinger, H. (1992). Intraclass Correlations as Estimates of Interrater Reliability in Nursing Research. *Western Journal of Nursing Research, 14*(2), 246-251.

Turner, R. & Carlson, L. (2003). Indexes of Item-Objective Congruence for Multidimensional Items. *International Journal of Testing, 3*(2), 163-171. doi: https://doi.org/10.1207/s15327574ijt0302_5

Vargha, P. (1997). Letter to the editor a critical discussion of intraclass correlation coefficients by R. Müller and P. Büttner. *Statistics in Medicine, 16*(7), 821-822. doi: https://doi.org/10.1002/(sici)1097-0258(19970415)16:7<821::aid-sim558>3.0.co;2-b

Ventura-León, L. (2017). El coeficiente Omega: un método alternativo para la estimación de la confiabilidad. *Revista Latinoamericana En Ciencias Sociales, Niñez y Juventud, 15*(1), 625-627.

Ventura-León, J. L. (2018). ¿Es el final del alfa de Cronbach? *Adicciones, 31*(1), 2016-2017. doi: https://doi.org/10.20882/adicciones.1037

Viladrich, C., Angulo-Brunet, A. & Doval, E. (2017). A journey around alpha and omega to estimate internal consistency reliability. *Anales de Psicologia, 33*(3), 755-782. doi: https://doi.org/10.6018/analesps.33.3.268401

Weir, J. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research, 19*(1), 231-240. doi: https://doi.org/10.1007/978-3-642-27872-3_5

Contribución de los autores: a) Concepción y diseño del trabajo; b) Adquisición de datos; c) Análisis e interpretación de datos; d) Redacción del manuscrito; e) revisión crítica del manuscrito.
J. C-R. ha contribuido en a, b, c, d, e.

Editora científica responsable: Dra. Cecilia Cracco.

Anexo

Sintaxis para IBM SPSS

```
GLM Medida1 Medida2
  /WSFACTOR=factor1 2 Polynomial
  /METHOD=SSTYPE(3)
  /EMMEANS=TABLES(factor1) COMPARE ADJ(LSD)
  /EMMEANS=TABLES(OVERALL)
  /PRINT=ETASQ
  /CRITERIA=ALPHA(.05)
  /WSDESIGN= factor1.
```

Reliability

```
  /VARIABLES=Medida1 Medida2
  /SCALE('ALL VARIABLES') ALL
  /MODEL=ALPHA
  /ICC=MODEL(MIXED) TYPE(ABSOLUTE) CIN=95 TESTVAL=0.
```