Protocolo de Evaluación de Tests, Escalas y Cuestionarios (PETEYC)

Manual de uso

En el manual del Protocolo de Evaluación de Tests, Escalas y Cuestionarios (PETEYC) se describe cómo administrar la herramienta y cómo procesar la información recogida con ella.

Los pasos que deben seguirse para utilizar PETEYC son los siguientes:

1. Recopilar toda la información disponible sobre el instrumento de evaluación objetivo, incluyendo el propio instrumento, el manual de uso (en caso de que exista) y cualquier dato empírico que haya sido recogido con el fin de evaluar las propiedades del instrumento (datos cualitativos proporcionados por expertos, datos cuantitativos procedentes de un estudio piloto, entrevistas con usuarios/as, etc…).
2. Administrar PETEYC al instrumento de evaluación objetivo, completando cada uno de los apartados con la información disponible.
3. Descargar la "Herramienta para procesar las evaluaciones" (PROC_PETEYC) de la página web de IAPA y seguir las instrucciones del presente documento para completar la información que en ella se demanda. Edite únicamente las celdas en blanco.

A continuación, se indica cómo trasladar la información recogida durante la administración de PETEYC a PROC_PETEYC. Para ello, se utilizarán los números entre corchetes situados en PETEYC- Herramienta para la evaluación.

[1] El primer paso consiste en identificar el uso previsto del test. El uso señalado en PETEYC determinará la pestaña de PROC_PETEYC en la que se realizará la evaluación. Por ejemplo, si el uso previsto del test es “Diagnóstico” se seleccionará la pestaña correspondiente en la parte inferior de PROC_PETEYC, tal y como se indica en la imagen.

[2] Los apartados señalados con códigos cuya primera cifra es el 2 servirán para evaluar la idoneidad de las **características y el tamaño de la muestra** utilizada en el estudio piloto. La información sobre la muestra incluida en el apartado **[2a]** de PETEYC servirá para establecer la puntuación de cada una de las celdas de la fila dedicada a analizar “Tamaño e idoneidad de la muestra”, de la siguiente manera:

- **Tamaño:** El tamaño muestral se evaluará considerando el número total de observaciones y el número de personas e ítems respondidos. Se considerará que el tamaño muestral es adecuado cuando el número de observaciones sea igual o superior a 200 (Ferrando y Anguiano-Carrasco, 2010) o cuando, siendo menor a 200 se recogen cinco o diez respuestas por ítem (Muñiz y Fonseca-Pedrero, 2019). En ambos casos, se asignará una puntuación de 1, en caso contrario se asignará un 0.

- **Representatividad:** La representatividad de la muestra se evaluará comparando las características de la muestra con las características de la población objetivo especificadas en el apartado **[2b]**. Se asignará una puntuación de 1 si las

características de la muestra coinciden con las de la población objetivo, y de 0 en caso contrario.

- Selección de la muestra: Se asignará una puntuación de 1 si la muestra ha sido seleccionada aleatoriamente o mediante un muestreo no aleatorio que haya permitido representar adecuadamente las características de la población, y de 0 en cualquier otra situación.

[3] Los apartados señalados con códigos cuya primera cifra es el 3 servirán para evaluar la **fiabilidad** del instrumento.

- Inclusión: Se asignará una puntuación de 1 si se aportan datos sobre la fiabilidad del instrumento, y de 0 en caso contrario. Si la evaluación no se ha incluido, por favor, rellene únicamente la siguiente celda sobre "adecuación de la decisión" y continúe con el siguiente bloque.

- **Adecuación de la decisión:** Se asignará una puntuación de 1 si se considera relevante la evaluación de la fiabilidad, y de 0 en caso contrario. La evaluación de la fiabilidad se considera relevante cuando el test evalúa un único constructo y aporta una puntuación total o, en caso de que evalúe varias dimensiones, si presenta un análisis de fiabilidad para cada una de ellas. La evaluación de la fiabilidad no será relevante cuando el instrumento objetivo persiga evaluar más de un constructo a través de ítems independientes que no formen parte de una única escala.

- **Adecuación del procedimiento:** Se asignará una puntuación de 1 si se considera adecuado el procedimiento utilizado para evaluar la fiabilidad, y de 0 en caso contrario. El procedimiento será adecuado si se ajusta a las características del instrumento de evaluación. Por ejemplo, será inadecuado si se evalúa la

fiabilidad mediante test-retest en instrumentos que evalúan constructos que pueden cambiar con el paso del tiempo.

- **Apoyo de los resultados al uso previsto:** Se asignará una puntuación de 1 si los resultados reflejan que el instrumento es fiable, y 0 en caso contrario. Por ejemplo, el test será fiable si se obtienen valores superiores a .7 en indicadores como Alpha de Cronbach u Omega de McDonald. Valores superiores a .9 podrían indicar la presencia de contenido redundante (Panayides, 2013). En el caso de procedimientos como test-retest o dos mitades, se considerarán adecuados valores de correlación superiores a .3. Se asignará una puntuación de 0,5 cuando se hayan obtenidos resultados adecuados en algunas subescalas e inadecuados en otras, o cuando los valores de Alpha u Omega sean cercanos al criterio de .7.

[4] Los apartados señalados con códigos cuya primera cifra es el 4 servirán para evaluar en qué medida los resultados reflejan que las propiedades psicométricas de los ítems son apropiadas.

- **Inclusión:** Se asignará una puntuación de 1 si se aportan datos sobre las propiedades psicométricas de los ítems, y de 0 en caso contrario. Si la evaluación no se ha incluido, por favor, rellene únicamente la siguiente celda sobre "adecuación de la decisión" y continúe con el siguiente bloque.

- **Adecuación de la decisión:** Se asignará una puntuación de 1 si se consideran relevantes los datos psicométricos aportados, y de 0 en caso contrario. Los datos serán relevantes cuando incluyan la distribución de las respuestas en las distintas alternativas, índices de discriminación, índices de dificultad (sólo para tests de rendimiento óptimo) y fiabilidad del instrumento al eliminar cada ítem. Se asignará una puntuación parcial (0,5) cuando se aporten algunos de los datos.

- **Adecuación del procedimiento:** Se asignará una puntuación de 1 si se considera adecuado el procedimiento utilizado para evaluar cada una de las propiedades psicométricas de acuerdo a la aproximación seguida (Teoría Clásica de los Tests o Teoría de Respuesta al Ítem), y de 0 en caso contrario. El procedimiento será adecuado si se ajusta a las características del instrumento de evaluación.

- **Apoyo de los resultados al uso previsto:** La puntuación asignada en este apartado reflejará la proporción de ítems que muestran propiedades adecuadas. De manera que se asignará una puntuación de 1 cuando todos los ítems funcionen adecuadamente. En el caso de que el 80% de los ítems reflejen propiedades adecuadas la puntuación será 0,8. Serán adecuados valores superiores a .3 al calcular el índice de discriminación mediante la correlación ítem-total (Bichi, 2016).

[5] Los apartados señalados con códigos cuya primera cifra es el 5 servirán para evaluar en qué medida las evidencias basadas en el contenido del test apoyan el uso y la interpretación de las puntuaciones prevista.

- **Inclusión:** Se asignará una puntuación de 1 si se aportan datos derivados de analizar el solapamiento entre el modelo teórico y el contenido del test, y de 0 en caso contrario. Si la evaluación no se ha incluido, por favor, rellene únicamente la siguiente celda sobre “adecuación de la decisión” y continúe con el siguiente bloque.

- **Adecuación de la decisión:** Se asignará una puntuación de 1 si, dadas las características del instrumento, las evidencias de validez basadas en el contenido aportan o podrían aportar información relevante. Esta fuente de evidencias será

relevante cuando sea necesario mostrar que el instrumento ha recogido adecuadamente los indicadores del constructo objetivo.

- **Adecuación del procedimiento:** Se asignará una puntuación de 1 si se considera adecuado el procedimiento utilizado para obtener evidencias de validez basadas en el contenido del test, y de 0 en caso contrario. Por ejemplo, el procedimiento será adecuado si arroja resultados sobre la representatividad y la relevancia de los ítems creados para evaluar el constructo, y se presentan índices de validez de contenido. Se asignará una puntuación de 1 si se presentan datos cualitativos que permitan identificar ítems potencialmente problemáticos y se incorporan sugerencias o cambios dirigidos a mejorar la calidad de los ítems.

- **Apoyo de los resultados al uso previsto:** Se asignará una puntuación de 1 si los resultados apoyan el uso previsto del test, es decir, si se aportan evidencias que confirmen que el instrumento evalúa el constructo perseguido. Por ejemplo, cuando los valores de CVR según el Índice V de Aiken sean iguales o superiores a .8 (Penfield y Giacobbi, 2004). Pueden consultarse detalles adicionales en la siguiente fuente: Sireci y Faulkner-Bond, 2014. Se asignará una puntuación de 0 en caso contrario. Se asignará también una puntuación de 1 cuando se aporten datos cualitativos que apoyen el uso previsto del instrumento.

[6] Los apartados señalados con códigos cuya primera cifra es el 6 servirán para evaluar en qué medida las evidencias basadas en la estructura interna del test apoyan el uso y la interpretación de las puntuaciones previstas.

- **Inclusión:** Se asignará una puntuación de 1 si se aportan datos derivados de analizar la dimensionalidad del instrumento, y/o la invarianza de la medida, y de 0 en caso contrario. Si la evaluación no se ha incluido, por favor, rellene

únicamente la siguiente celda sobre "adecuación de la decisión" y continúe con el siguiente bloque.

- **Adecuación de la decisión:** Se asignará una puntuación de 1 si, dadas las características del instrumento, las evidencias de validez basadas en la estructura interna aportan o podrían aportar información relevante. Esta fuente de evidencias será relevante cuando sea necesario mostrar que la estructura del instrumento refleja adecuadamente las dimensiones teóricas del constructo y cuando sea relevante mostrar la invarianza de la medida entre distintos grupos.

- **Adecuación del procedimiento:** Se asignará una puntuación de 1 si se considera adecuado el procedimiento utilizado para obtener evidencias de validez basadas en la estructura interna del test, y de 0 en caso contrario. Por ejemplo, el procedimiento será adecuado si refleja el solapamiento entre la configuración del constructo y la del instrumento.

- **Apoyo de los resultados al uso previsto:** Se asignará una puntuación de 1 si los resultados apoyan el uso previsto del test, es decir, si se aportan evidencias que confirmen que el instrumento recoge las dimensiones teóricas del constructo. Por ejemplo, cuando se aporten valores adecuados sobre el ajuste del modelo obtenido mediante análisis exploratorios o confirmatorios (CFI >.95; TLI>.95, SRSM<.08; RMSEA<.06). Pueden consultarse detalles sobre los criterios habitualmente utilizados en las siguientes fuentes: Bentler, 1990; Hu y Bentler, 1999; Van Laar y Braeken, 2021. Se asignará una puntuación de 0 en caso contrario.

[7] Los apartados señalados con códigos cuya primera cifra es el 7 servirán para evaluar en qué medida las evidencias basadas en las relaciones con otras variables apoyan el uso y la interpretación de las puntuaciones prevista.

- **Inclusión:** Se asignará una puntuación de 1 si se aportan datos derivados de analizar relaciones de las puntuaciones del instrumento objetivo con puntuaciones de otros instrumentos que miden variables teóricamente relacionadas, y de 0 en caso contrario. Si la evaluación no se ha incluido, por favor, rellene únicamente la siguiente celda sobre "adecuación de la decisión" y continúe con el siguiente apartado.

- **Adecuación de la decisión:** Se asignará una puntuación de 1 si, dadas las características del instrumento, las evidencias de validez basadas en las relaciones con otras variables aportan o podrían aportar información relevante. Esta fuente de evidencias será relevante cuando sea necesario mostrar que las puntuaciones del instrumento son coherentes con las puntuaciones en otros instrumentos que evalúan variables teóricamente relacionadas.

- **Adecuación del procedimiento:** Se asignará una puntuación de 1 si se considera adecuado el procedimiento utilizado para obtener evidencias de validez basadas en las relaciones con otras variables, y de 0 en caso contrario. Por ejemplo, el procedimiento será adecuado si refleja relaciones entre las puntuaciones del instrumento y puntuaciones de otros instrumentos que evalúan variables teóricamente relacionadas.

- **Apoyo de los resultados al uso previsto:** Se asignará una puntuación de 1 si los resultados apoyan el uso previsto del test, es decir, si se aportan evidencias que confirmen que existe la relación prevista, y ~~Se asignará~~ una puntuación de 0 en caso contrario. Por ejemplo, cuando se aporten valores de correlaciones superiores a .3 (positivas o negativas, según corresponda) o valores que reflejen un área bajo la curva (AUC) cercanos a 1 (en análisis de curvas ROC; Muñiz, 2018). Se asignarán puntuaciones parciales (0,5) cuando se encuentren las relaciones

esperadas con algunas variables, pero no sea así con todas ellas; o cuando los valores obtenidos sean cercanos a los esperados.

[8] Los apartados señalados con códigos cuya primera cifra es el 8 servirán para evaluar en qué medida las evidencias basadas en los procesos de respuesta apoyan el uso y la interpretación de las puntuaciones prevista.

- Inclusión: Se asignará una puntuación de 1 si se aportan datos derivados de analizar relaciones de los procesos de respuesta desarrollados al responder al instrumento, y de 0 en caso contrario. Si la evaluación no se ha incluido, por favor, rellene únicamente la siguiente celda sobre "adecuación de la decisión" y continúe con el siguiente bloque.

- **Adecuación de la decisión:** Se asignará una puntuación de 1 si, dadas las características del instrumento, las evidencias de validez basadas en los procesos de respuesta aportan o podrían aportar información relevante. Esta fuente de evidencias será relevante cuando sea necesario recoger información que refleje que las personas desarrollan procesos de respuesta alineados con el constructo objetivo.

- **Adecuación del procedimiento:** Se asignará una puntuación de 1 si se considera adecuado el procedimiento utilizado para obtener evidencias de validez basadas en los procesos de respuesta, y de 0 en caso contrario. Por ejemplo, el procedimiento será adecuado si aporta información sobre cómo han respondido las personas al instrumento.

- **Apoyo de los resultados al uso previsto:** Se asignará una puntuación de 1 si los resultados apoyan el uso previsto del test, es decir, si se aportan evidencias que confirmen que las personas piensan en los indicadores previstos. Por ejemplo,

cuando aporten argumentos que reflejen que las respuestas hacen referencia al constructo previsto. Se asignará una puntuación de 0 en caso contrario. Se asignarán puntuaciones parciales cuando se aporten evidencias sólo en algunos casos.

[9] Los apartados señalados con códigos cuya primera cifra es el 9 servirán para evaluar en qué medida las evidencias basadas en las consecuencias de la evaluación apoyan el uso y la interpretación de las puntuaciones prevista.

- **Inclusión:** Se asignará una puntuación de 1 si se aportan datos derivados de analizar las consecuencias de la evaluación, y de 0 en caso contrario. Si la evaluación no se ha incluido, por favor, rellene únicamente la siguiente celda sobre "adecuación de la decisión" y continúe con el siguiente apartado.

- **Adecuación de la decisión:** Se asignará una puntuación de 1 si, dadas las características del instrumento, las evidencias de validez basadas en las consecuencias de la evaluación aportan o podrían aportar información relevante. Esta fuente de evidencias será relevante cuando sea necesario recoger información que refleje que la evaluación no ha tenido consecuencias imprevistas para las personas evaluadas.

- **Adecuación del procedimiento:** Se asignará una puntuación de 1 si se considera adecuado el procedimiento utilizado para obtener evidencias de validez basadas en las consecuencias de la evaluación, y de 0 en caso contrario. Por ejemplo, el procedimiento será adecuado si aporta información sobre las consecuencias que la evaluación ha tenido para las personas evaluadas.

- **Apoyo de los resultados al uso previsto:** Se asignará una puntuación de 1 si los resultados apoyan el uso previsto del test, es decir, si se aportan evidencias que

confirmen que las personas no han sufrido consecuencias imprevistas. Por ejemplo, cuando aporten argumentos que reflejen que la evaluación no ha generado una situación de injusticia. Se asignará una puntuación de 0 en caso contrario.

Una vez completado PROC_PETEYC, se obtendrá una puntuación total que estará comprendida entre 0 y 100. Puntuaciones cercanas a 100 indicarán que el instrumento dispone de evidencia suficiente para ser utilizado en el contexto previsto. Puntuaciones cercanas a 0 indican la necesidad de revisar profundamente el instrumento. PROC_PETEYC permite además identificar las áreas más débiles. Aquellas filas cuya puntuación total (columna G) sea más baja que el valor incluido en el peso (Columna F) serán las áreas con las que podría trabajarse para mejorar la calidad del instrumento. Para trabajarlas, pueden seguirse las indicaciones en rojo incorporadas en PETEYC.

Financiado por el Programa 39 del Plan Propio de Investigación y Transferencia de la Universidad de Granada, 2021

Referencias

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological bulletin*, *107*(2), 238-246.

Bichi, A. A. (2016). Classical Test Theory: An introduction to linear modeling approach to test and item analysis. *International Journal for Social Studies*, *2*(9), 27-33.

Ferrando, P. J., & Anguiano-Carrasco, C. (2010). El análisis factorial como técnica de investigación en psicología. *Papeles del psicólogo*, *31*(1), 18-33.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, *6*(1), 1-55.

Muñiz, J. (2018). *Introducción a la Psicometría: Teoría clásica y TRI*. Pirámide.

Muñiz, J., & Fonseca-Pedrero, E. (2019). Diez pasos para la construcción de un test. *Psicothema*, *31*(1).

Panayides, P. (2013). Coefficient Alpha: Interpret With Caution. *Europe's Journal of Psychology*, *9*(4), 687-696. https://doi.org/10.5964/ejop.v9i4.653

Penfield, R. D., & Giacobbi, J. (2004). Applying a score confidence interval to Aiken's item content-relevance index. *Measurement in physical education and exercise science*, *8*(4), 213-225.

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema*, *26.1*, 100-107.

Van Laar, S., & Braeken, J. (2021). Understanding the Comparative Fit Index: It's all about the base! *Practical Assessment, Research & Evaluation*, *26*(1).