UNIVERSIDAD NACIONAL DE LA PLATA
FACULTAD DE PSICOLOGIA

CATEDRA: *FUNDAMENTOS, TECNICAS E INSTRUMENTOS DE EXPLORACION PSICOLOGICA I*

FICHA Nº 5 (Unidad 3)
LA VALIDEZ Y CONFIABILIDAD DE LOS INSTRUMENTOS DE EVALUACIÓN PSICOLÓGICA
Año 2003

Autor: Prof. Telma Piacente

INTRODUCCIÓN

El problema de la exploración psicológica se encuentra estrechamente relacionado, entre otras cuestiones, al problema de la medición en Psicología. En un sentido amplio medir significa aplicar las propiedades del número a las propiedades de los objetos o eventos materia de medición. Ahora bien, como los objetos o eventos de carácter psicológico se caracterizan por su naturaleza abstracta, corresponde restringir la noción de medición como "el proceso de vincular conceptos abstractos con indicadores empíricos" (Hernández Sampieri y ot., 1998). En el caso de los instrumentos de exploración psicológica, tales indicadores son los puntajes con los que se expresan los resultados. Esos puntajes han sido cuidadosamente seleccionados en el momento de construcción de un instrumento. Los conceptos subyacentes no observables que se intenta examinar se operacionalizan a través de reactivos (problemas, cuestiones, etc) que promueven respuestas observables, a las que se adjudican valores que permiten cuantificar o clasificar. Es decir que se realiza una operación que permite observar aquello que no es directamente observable, y sobre la que es posible estimar valores.

No obstante hablar de observables, recordemos sucintamente el rol de las teorías en la delimitación u operacionalización de los observables. Al contrario de lo que se supone ordinariamente, los hechos "no hablan por si mismos" (Cole, 1998). Son las teorías las que proporcionan el marco a partir del cual observar. Einstein describe el rol central de las teorías en los siguientes términos "es equivocado tratar de fundar las teoría en los observables. En realidad ocurre lo contrario. Es la teoría la que decide lo que debe ser observado" (citado en Cole, 1998). Esa teoría puede ser más o menos explícita, pero es conveniente que esté especificada. Por otra parte en muchos casos, especialmente en la disciplina psicológica, se trata de grandes teorías (por ejemplo la teoría factorial o la teoría piagetiana de la inteligencia), y en otros casos de teorías de alcance medio.

Volviendo a la problemática de los instrumentos de exploración, los requisitos indispensables que informan sobre su valor refieren a que "registren datos

observables que representen verdaderamente los conceptos o variables que el investigador tiene in mente" (Hernández Sampieri, Fernández Collado, & Baptista Lucio, 1998).

Sea cual fuere el propósito de la exploración psicológica (hacer un psicodiagnóstico, hacer una evaluación psicológica de aspectos parciales del sujeto de examen o de otra unidad de análisis –contextos-interacciones-intervenciones) usualmente se aplican diferentes instrumentos para medir las variables contenidas en las hipótesis que se formulan y, en el caso en el que no hubieran hipótesis, las variables a ser analizadas. La medición es adecuada cuando el instrumento de recolección de datos *realmente examina las variables que debe examinar y las examina adecuadamente*, aunque obviamente en el campo de las ciencias fácticas la medición no es perfecta. Todo instrumento de recolección de datos debe reunir en consecuencia dos requisitos esenciales de calidad: *validez y confiabilidad.*

VALIDEZ

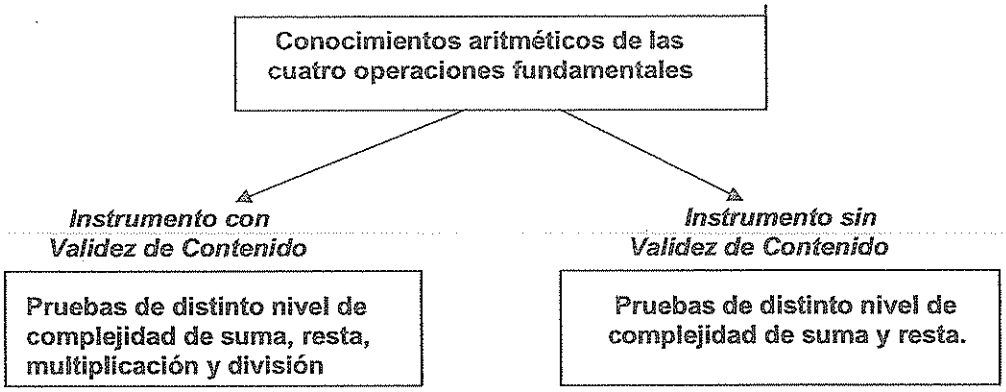
"La validez, en términos generales, se refiere al grado en que un instrumento realmente mide la variable que pretende medir. Por ejemplo, un instrumento válido para medir inteligencia debe medir la inteligencia y no la memoria" (Hernández Sampieri, Fernández Collado, & Baptista Lucio, 1998). La validez es un concepto del cual pueden tenerse diferentes tipos de evidencia:

1. Evidencia relacionada con el contenido.
2. Evidencia relacionada con el criterio.
3. Evidencia relacionada con el constructo.

1. Evidencia relacionada con el contenido

Se habla en este caso de *validez de contenido.* Se refiere al grado en el que un instrumento muestrea un dominio específico del contenido de aquello que se mide. Por ello es necesario especificar el contenido de lo que se mide. Por ejemplo la exploración de *conocimientos matemáticos*, cuyos contenidos no estuvieran suficientemente especificados, podría estar muestreada de diferentes maneras, en una prueba que abarcara desde conocimientos de nociones de numeración hasta conocimientos sobre resolución de ecuaciones de dos incógnitas. Si en cambio la variable "conocimientos matemáticos" está mejor especificada, entendiendo por ello, para determinadas edades, el dominio de las cuatro operaciones fundamentales (suma, resta, multiplicación y división), en las que se utilicen números enteros y fraccionarios, se debería armar una prueba con operaciones de ese tipo, de diferente nivel de complejidad, que involucren el tipo de números mencionados. Hernández Sampieri (1998) afirma que un instrumento, para poseer este tipo de validez, debe contener representados todos los ítems del dominio de contenido de las variables a medir. Esto aparece ilustrado en la Figura 1.

Figura 1
Dominio de la variable

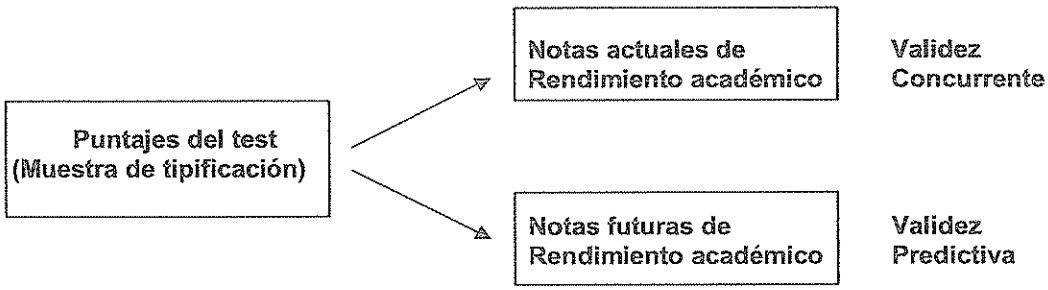


2. Evidencia relacionada con el criterio

Se habla en este caso de *validez empírica o de criterio*. Recibe este nombre porque se comparan lo resultados del test (de todos o algunos de los sujetos incluidos en la muestra de tipificación) con los resultados en una medición externa al test, denominada criterio[1]. Por ejemplo un test de rendimiento académico, puede tomar como criterio externo las notas de rendimiento académico de los sujetos de la muestra.

Si la comparación con el criterio externo se hace en el presente, se habla de *validez concurrente*, es decir que los resultados del tests correlacionan positivamente (correlaciones moderadamente altas) con los resultados de las notas académicas obtenidas en el mismo tiempo que los resultados de la prueba.

Si en cambio la comparación con el criterio se hace en un tiempo futuro, se habla de *validez predictiva,* es decir *que* los resultados del tests correlacionan positivamente (correlaciones moderadamente altas) con los resultados de las notas académicas obtenidas en un tiempo posterior a los resultados de la prueba.



Este tipo de validez se obtiene constatando la correlación que existe entre los dos sistemas de puntuaciones. Un instrumento resultará válido si el coeficiente de correlación es moderadamente alto.

---

[1] **criterio** (del gr. «kritérion»; «Aplicar, Servir de») m. *Norma para *juzgar una cosa (Diccionario M. Moliner, 1996).

3. Evidencia relacionada con el constructo

Se habla aquí de *validez de constructo*. Constituye la validez más importante, en la medida que se refiere al "grado en que una medición se relaciona consistentemente con otras mediciones de acuerdo con hipótesis derivadas teóricamente y que conciernen a los conceptos (o constructos) que están siendo medidos. Un constructo es una variable medida y que tiene lugar dentro de una teoría o esquema teórico" (Hernández Sampieri, Fernández Collado, & Baptista Lucio, 1998). Para afirmar la validez de constructo suelen correlacionarse un número significativo de mediciones de variables que teóricamente y de acuerdo con estudios precedentes están relacionadas.

La validez de constructo requiere del establecimiento de tres etapas:

1. Establecimiento de la relación teórica entre los conceptos, por ejemplo, el rendimiento en lectura se relaciona con el nivel de conciencia fonológica, partir del marco teórico seleccionado.

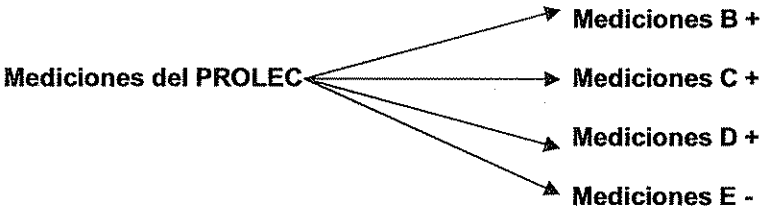
2. Constatación de la correlación entre ambos conceptos.

3. Interpretación de la evidencia empírica (alcance y significación de las correlaciones encontradas).

En la Figura 3 se ejemplifica este tipo de validez.

Fig. 3
Instrumento que mide el constructo rendimiento en lectura

Teoría: La Psicología Cognitiva afirma que el rendimiento en lectura depende en gran medida del nivel de conciencia fonológica (conciencia fonémica). Investigaciones realizadas con el PROLEC[2] (resultados A) encontraron que se relaciona positivamente con otras mediciones del rendimiento en lectura.
Si el instrumento mide realmente el rendimiento en lectura, sus resultados deben relacionarse positivamente con los resultados en otras mediciones del rendimiento en lectura ya probadas y negativamente con mediciones de otras variables.



La validez de una prueba se constata sobre la base de los tres tipos de evidencia y en cada instrumento debe especificarse de que tipo de validez se trata. No obstante cuanto mayor validez tenga en un instrumento en los tres tipos descriptos, mayor será su ajuste a aquello que se propone medir (validez total = validez de contenido + validez de criterio + validez de constructo).

[2] Test de Procesos Lectores

Procedimientos para encontrar la validez

El procedimiento para obtener la *validez de contenido* es complejo. En primer lugar deben determinarse los alcances de la variable a medir (definir que se va a medir y con que intensidad). Luego se debe revisar cómo ha sido examinada la variable en la literatura especializada, a partir de la cual poder elaborar un universo exhaustivo de ítems posibles. Posteriormente se debe realizar una consulta con expertos en el tema para corroborar si ese universo es exhaustivo. Finalmente se seleccionan los ítems, luego de una cuidadosa evaluación. En el caso de que la variable a examinar incluya distintas facetas o dimensiones[3], se selecciona una muestra probabilística de los ítems correspondientes a cada una de ella, para que todas la facetas se encuentren adecuadamente representadas. Luego se correlacionan las puntuaciones de los ítems entre sí. Tales correlaciones deben se altas, especialmente entre los ítems correspondientes a una misma faceta o dimensión.

La validez empírica o de criterio resulta más fácil de estimar. El procedimiento que se sigue es correlacionar las puntuaciones del test con las puntuaciones del criterio (por ejemplo notas académicas o notas proporcionadas por expertos (ver figura 4).

Puntajes del test de los Sujetos de la muestra	correlación	Puntajes del Criterio externo

La estimación de la validez de constructo es muy compleja. Se relaciona con una verificación experimental de las hipótesis que corresponden a los rasgos psicológicos o construcciones teóricas que justifican la ejecución del test. Usualmente se utilizan para ello estudios factoriales, que permiten establecer en qué medida el test mide una capacidad común en un grupo de pruebas similares. Otro ejemplo de validación de constructo lo constituye la técnica de diferenciación por edades, utilizada frecuentemente para validar tests infantiles de inteligencia. Parte del supuesto que es propio de la inteligencia que progrese con la edad. En tal caso se examinan las puntuaciones del test para ver si incrementan a medida que se avanza hacia edades cronológicas mayores.

UN POCO DE HISTORIA SOBRE EL CONCEPTO DE VALIDEZ

El concepto de validez hasta la primera mitad del Siglo XX estuvo referido básicamente al uso predictivo de los tests. Las dos circunstancias histórico científicas que impulsaron la creación de instrumentos refieren a la instauración de la enseñanza primaria obligatoria y a la Primera Guerra Mundial (1914 - 1918) por un lado y al auge del estudio de las diferencia individuales a la luz de concepciones psicológicas que progresivamente se alejaban del asociacionismo. La primera dio lugar a la creación de los primeros *tests individuales*, producto de la labor de A. Binet, quien introduce dos innovaciones de gran importancia. Una alude a la ruptura con la *tradición empirista*, según la cual el examen de los procesos psicológicos debería recaer en el examen de los procesos sensoriales, a través de *pruebas simples*. Esta perspectiva es

[3] Aspectos o partes de la variable

reemplazada por una concepción que tomó en consideración la evaluación de procesos complejos a través de *pruebas complejas*. La otra refiere a la *metodología de la medida*, que le permite a Binet construir una escala graduada por niveles de edad, de modo tal que el rendimiento de un sujeto puede ser expresado en términos del nivel al que corresponde, dando como resultado las diferencias que pueden encontrarse no solo entre niños y adultos sino además entre niños de diferente edad, permitiendo asimilar el rendimiento de los retrasados al de los sujetos de menor edad. Los resultados de sus pruebas permitían predecir la necesidad de enseñanza especial de un grupo de niños (Reuchlin, 1985).

La participación de EEUU en la Primer Guerra Mundial, impulsó a su vez el desarrollo de *tests colectivos*. En este caso, bajo la dirección de Robert M. Yerkes, se construyeron instrumentos que permitieran clasificar de manera rápida a millón y medio de reclutas respecto de su nivel intelectual general. Tal información "sirvió de ayuda en muchas decisiones administrativas, incluyendo la exención del servicio militar, la asignación de diferentes tipos de servicios, etc." (Anastasi, 1982).

Aunque el armisticio de 1918 no hizo posible arribar a conclusiones válidas respecto de las predicciones que se realizaron a través del rendimiento en los tests (Boring, 1961, citado por Muñiz, 1996), el prestigio de los instrumentos avanzó considerablemente y en ocasión de la Segunda Guerra Mundial "se creó un comité de psicólogos para la clasificación del personal militar" (Muñiz, 1996), que elaboró pruebas, cuya validez seguía siendo preferentemente predictiva.

Hacia la década del 50 el uso de los tests se extendió a la selección de personal, tarea que implicaba tomar decisiones respecto del desempeño actual de los sujetos en relación con su desempeño en los instrumentos de evaluación. Se realizaron entonces las primeras validaciones simultáneas de los resultados del test y los resultados del criterio externo (notas de desempeño en tareas reales). Dicho de otro modo la validez en este caso era *concurrente*.

Esta concepción de validez empírica o de criterio (predictiva o concurrente), no resulta sin embargo apta para los test de rendimiento académico o escolar, por cuanto los instrumentos que intentan explorar esta variable (rendimiento escolar o académico) generalmente se utilizan para examinar el desempeño de un grupo de sujetos luego de un periodo de instrucción. Es decir que se utilizan los resultados del test como indicadores de la variable que se pretende evaluar y no como predictores de conductas ajenas al test. Se habla entonces de *validez de contenido*, es decir de la congruencia del contenido con la variable que se pretende medir, especificada por el juicio de expertos sobre la pertinencia de los ítems incluidos (Muñiz, 1996).

Paralelamente a ese recorrido histórico, comienzan a surgir diferentes teorías sobre "la estructura de las variables psicológicas y se comienzan a elaborar instrumentos para probar dichas teorías" (Muñiz, 1996). Es decir que se pasa de la construcción de tests empíricos a la construcción de tests teóricos. En la década del 30, el auge de las teorías factorialistas de la inteligencia conduce a la elaboración y uso de técnicas estadísticas multivariadas para la construcción

y validación de los tests. Se produce así un cambio de perspectiva en la consideración de la validez. La denominada *validez de constructo* pretende examinar el grado en que cada test refleja el constructo que pretende medir.

Esta concepción triple de la validez, que apunta a diferentes aspectos se mantiene hasta la década de los ochenta. En resumen, a partir de cada una de ellas es posible realizar distintos tipos de inferencias:

1. Validez de criterio = sobre las relaciones de los resultados del test con el desempeño de los sujetos de la muestra de tipificación en un área independiente del test.
2. Validez de contenido = sobre como se muestrean en el test los contenidos que se desean examinar
3. Validez de constructo = sobre como mide el test aquello que intenta medir, es decir el constructo psicológico que subyace al mismo.

Estos tipos de validez fueron especificados en los Estándares de la American Psychological Association (APA) en 1954, como validez predictiva, concurrente, de contenido y de constructo. Posteriormente en el año 1966 se unifica la validez predictiva y concurrente como validez empírica o de criterio.

En los años 80 se produce dos innovaciones de importancia. En primer lugar se considera a la validez como única (Cronbach, 1980, citado por Muñiz, 1998). Efectivamente, en los Estándares de la APA, a partir del año 1985 ya no se habla de tipos de validez sino de categorías de validez, queriendo significar con esto que "una validación ideal incluye varios tipos de evidencia implicados en los tres tipos tradicionales". En segundo lugar se considera a la validez no ya como un concepto inherente al test sino que lo que se valida corresponde a las inferencias realizadas a partir del mismo. Este cambio de enfoque tiene dos importantes consecuencias: por un lado el responsable de establecer la validez de un test no es sólo el constructor, sino también el usuario. Por la otra la validez no se establece de una vez y para siempre, se trata de un proceso continuo, que finalizaría en el momento en el que se encontrara evidencia en contrario, a la manera de la contrastación popperiana.

Esta concepción unitaria de validez no puede dejar de lado la consideración de que la única validez admisible es la de constructo: un test debe medir aquello que se propone y no otra cosa. La validez de criterio y la de contenido, quedarían subsumidas en aquella. No obstante se sigue manteniendo esta clasificación, aunque no ya como tipos de validez, tal como se señalara, sino como estrategias de validación, de acuerdo al tipo de inferencias que se quieran realizan.

CONFIABILIDAD

El propósito fundamental de los estudios de fiabilidad o confiabilidad es estimar la magnitud de los errores cometidos al medir las variables psicológicas, dado que en la práctica es casi imposible que una medición sea perfecta.

Existen diversos procedimientos para estimar la confiabilidad de un instrumento de medición. Todos ellos utilizan fórmulas que utilizan el coeficiente de

confiabilidad, cuya magnitud, como en todo coeficiente de correlación puede oscilar entre -1, 0 y 1. Es decir que un coeficiente de 0 indicaría una confiabilidad nula y uno de 1 una confiabilidad perfecta. En la práctica se obtienen valores más próximos a uno u otro extremo. Cuando los valores que se obtengan estén más próximos a 1, más confiable serán los resultados del test.

Los diseños o procedimientos más usuales para obtener dos sistemas de puntuaciones cuya correlación pueda dar cuenta de la confiabilidad de las mediciones, en tanto se están midiendo cosas iguales son los siguientes:

- Formas paralelas, alternativas o equivalentes.
- Test-retest
- Dos mitades
- Homogeneidad y equivalencia de todos los elementos

Fiabilidad de las formas alternativas, paralelas o equivalentes

Este procedimiento para determinar la fiabilidad de un test consiste, tal como su nombre lo indica en correlacionar dos series de puntuaciones, cada una de las cuales corresponde a una de las formas paralelas, alternativas o equivalentes del test. Es decir que se administran dos versiones del mismo test a los sujetos de la muestra de tipificación (o a parte de ellos) y se halla la correlación entre ambos. Para que dos formas de un mismo test se consideren equivalentes deben ser similares en contenido, instrucciones, forma de registro, y en todas las otras características del test. Se considera que el instrumento es confiable si esa correlación es significativamente positiva. El coeficiente de confiabilidad obtenido se denomina *coeficiente de equivalencia*, puesto que indica en que medida ambas puntuaciones son equivalentes, miden de la misma manera.

Si bien esta forma de encontrar la confiabilidad del test es teóricamente correcta, en la práctica tiene el inconveniente de obligar a construir dos formas paralelas. Esta construcción es costosa y difícil.

Fiabilidad del test-retest

El procedimiento aquí consiste en correlacionar dos series de puntuaciones, cada una de las cuales corresponde a la administración del mismo test en dos ocasiones diferentes. Es decir que se administra el mismo test a los sujetos de la muestra de tipificación (o a parte de ellos) en dos oportunidades distintas y se halla la correlación entre ambas serie de puntuaciones. Se considera que el instrumento es confiable si esa correlación es significativamente positiva. El coeficiente de confiabilidad obtenido se denomina *coeficiente de estabilidad*, puesto que indica en que medida ambas puntuaciones mantienen su estabilidad en las dos mediciones realizadas.

Si bien este procedimiento es más económico que el anterior, tiene la desventaja de tener que determinar el tiempo óptimo que debe transcurrir entre una administración y la otra. Si el período es demasiado largo puede potenciarse excesivamente el cambio en la variable examinada. Esto es particularmente sensible en el caso de pruebas destinadas a niños, en los que

se observan generalmente grandes cambios con la edad. Por el contrario si el período es demasiado corto los sujetos examinados pueden recordar las respuestas proporcionadas en la primera ocasión. En el primer caso aparecerían falta de estabilidad debido al cambio en la variable y no a la falta de fiabilidad. En el segundo, por el contrario, una estabilidad mayor que la que verdaderamente corresponde, atribuible al efecto de memoria.

Fiabilidad de la división en dos mitades

En este tipo de procedimiento se divide a la prueba en dos mitades, obviamente equivalentes en el grado de dificultad. Para lograr esta equivalencia, dado que los ítems están graduados en orden de dificultad creciente, la prueba se divide en *ítems pares e impares.* La administración se realiza en una sola ocasión y se correlacionan los resultados obtenidos por los sujetos de la muestra de tipificación en ambas mitades. Se considera que el instrumento es confiable si esa correlación es significativamente positiva. El coeficiente de confiabilidad obtenido se denomina *coeficiente de consistencia interna*, puesto que indica en que medida ambas puntuaciones son consistentes en las dos mediciones realizadas.

Equivalencia y homogeneidad entre los elementos del test

Un cuarto método para estimar la confiabilidad de un test, se basa en la consistencia de las respuestas de los sujetos a todos los elementos del test (Anastasi, pg. 118). La administración se realiza en una sola ocasión y se examinan los resultados obtenidos por los sujetos de la muestra de tipificación en todos los ítems del test. El procedimiento más utilizado es el desarrollado por Kuder-Richardson. También se utiliza con propósitos similares el coeficiente alfa de Cronbach. Constituye el promedio de todos los coeficientes de división en dos mitades que resulte de las diferentes divisiones de un test (para ver su modo de cálculo ver Anastasi, 1985, cap. 5). Se considera que el instrumento es confiable si esa correlación es significativamente positiva. Este tipo de confiabilidad proporciona una medida de *equivalencia y homogeneidad de los elementos.*

Finalmente, debe considerarse que, en términos generales, puede afirmarse que la confiabilidad se relaciona con el número de ítems incluidos en la prueba: a mayor cantidad de ítems, mayor confiabilidad. Es por ello que el psicólogo debe estar prevenido frente a pruebas muy sencillas que a través de pocos ítems pretender examinar variables complejas.

Fiabilidad del puntuador o examinador

Los tests proporcionan instrucciones muy precisa acerca de cómo administrar, puntuar y evaluar un test, de modo tal que dejan lugar a pocas variaciones en la forma de puntuar del examinador. No obstante ese grado de precisión en las instrucciones, a través de las cuales se pretende controlar el juicio subjetivo del examinador, varía en función de cada tipo de instrumento. En el caso de tests colectivos las respuestas proporcionadas por el sujeto o bien son positivas o negativas (por ejemplo el test de Matrices Progresivas de Raven) o bien de elección múltiple (por ejemplo cuando se trata de una escala Likert de

puntuación): no dejan lugar a dudas respecto de su ponderación. Pero en el caso de otros instrumentos, especialmente individuales y de aplicación clínica, el examinador debe usar su juicio para ponderar las respuestas del sujeto. En estos casos para examinar la fiabilidad, es decir el grado de confianza que merecen los examinadores, se realizan estudios de confiabilidad de los puntuadores. Estos estudios consisten en comparar las puntuaciones que otorgan dos examinadores diferentes en los resultados de los mismos sujetos, y observar si existen diferencias significativas entre ellas. En algunos estudios realizados se han encontrado diferencias de hasta 13 puntos en el mismo grupo de sujetos. Estas constataciones conducen a volver a plantear dos cuestiones:

- Las instrucciones del test deben estar suficientemente especificadas, para estrechar el margen de las decisiones que debe tomar el examinador.
- El entrenamiento de los examinadores debe ser tan exhaustiva como lo requiera la complejidad del test en cuestión.

Actividades guías para el estudio de los instrumentos de exploración

1. Identificar si el manual del instrumento que está usando informa sobre datos de validez y confiabilidad.
2. Identificar cual ha sido la evidencia de validez utilizada en el test.
3. Identificar el procedimiento para establecer la confiabilidad y sobre que segmento de la muestra de tipificación (sujetos incluidos).
4. Hacer un breve comentario sobre la validez y confiabilidad del test

REFERENCIAS BIBLIOGRÁFICAS

Anastasi, A. (1974). *Tests Psicológicos.* Madrid: Ed. Aguilar.

Anastasi, A. (1978). El examen psicológico los tests. En L. Kanner, H.I. Kaplan et al., *La psiquiatría infantil.* Buenos Aires: Paidos.

Cole, M. (1999). *The Development of children*. USA: Freeman Worth

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (1998). *Metodología de la investigación*. Buenos Aires: McGraw-Hill

Muñiz, J. (1996). *Psicometría.* Madrid: Editorial Universitas

Reuchlin, M. (1980). *Psicología*. Madrid: Ediciones Morata.