Texto extraído de la cátedra Evaluación del deportista de Alto Rendimiento, del Comité Olímpico Español

FUNDAMENTOS DE LA MEDIDA Y DE LA EVALUACIÓN.

Los términos medición y evaluación son empleados con gran profusión, a menudo sin atender a su significado último. La medición es el proceso por el cuál se recoge información cuantitativa o cualitativa. La evaluación consiste en la utilización de mediciones para emitir un juicio de valor y adoptar decisiones. Ambos conceptos están interrelacionados. En este capítulo se analizan los requisitos que deben reunir los tests destinados a la evaluación de la aptitud física para que tengan el mínimo error posible.

1.1. Conceptos de medición y evaluación.

Los términos **medición**y **evaluación** son empleados con gran profusión, a menudo sin atender a su significado último. La medición es el proceso por el cuál se recoge información cuantitativa o cualitativa. La evaluación consiste en la utilización de mediciones para emitir un juicio de valor y adoptar decisiones. Ambos conceptos están interrelacionados. La evaluación es un proceso que utiliza mediciones, mientras que el propósito de las mediciones es reunir información. Las mediciones se obtienen mediante una serie de procedimientos denominados con carácter general **tests**.

El proceso de evaluación comporta la interpretación de la información obtenida mediante los tests, de acuerdo con unos criterios o valores de referencia, a veces denominados estándares. Los valores de referencia pueden ser de dos tipos: normativos o de criterio. Los valores de referencia normativos se obtienen a partir de la medición de la variable estudiada en una muestra representativa de la población a la cual pertenece el caso analizado. Por ejemplo, si medimos el VO_2max de un ciclista de 25 años de edad y obtenemos un valor de 70 $ml.kg^{-1}.min^{-1}$, nos puede interesar saber si el va lor registrado es normal o no, es decir, si entra dentro del intervalo de valores que presenta la mayoría de los sujetos sanos de 25 años de la población a la que pertenece ese ciclista. La referencia normativa nos indica que el 95 % de los varones sanos de 25 años tienen un VO_2max comprendido entre 38 y 45 $ml.kg^{-1}.min^{-1}$. Así pues, podríamos elaborar un primer juicio (evaluación) acerca del VO_2max de nuestro ciclista, afirmando que su VO_2max es muy superior al que presenta la mayoría de la población de su edad y sexo.

Probablemente su entrenador estará más interesado en saber si con este valor de VO_2max puede ganar la Vuelta Ciclista a España. Si sabemos que la Vuelta Ciclista a España nunca ha sido ganada por un ciclista con un VO_2max inferior a 75 $ml.kg^{-1}.min^{-1}$, podríamos adoptar este último valor como valor de referencia de criterio (el criterio sería el valor de VO_2max mínimo para poder ganar la Vuelta Ciclista a España). En consecuencia, podríamos emitir un juicio (evaluación) afirmando que el valor actual de VO_2max de nuestro ciclista no es suficientemente alto como para optar a ganar la Vuelta Ciclista a España. Por lo tanto, los tests aportan datos abstractos y la evaluación confiere un significado a los datos obtenidos.

La medición es la primera etapa en la evaluación, de tal manera que cuanto mayor sea la calidad de la medición mejor podrá ser la evaluación. Los tests de condición física constituyen procedimientos destinados a la medición de variables relacionadas con la capacidad de rendimiento deportivo.

1.2. Requisitos que deben reunir los tests de condición física.

El proceso de medición debe seguir unas pautas generales a las que nos referiremos en este capítulo, de tal manera que los tests practicados sean, en primer lugar, válidos, fiables y objetivos.

La **validez** es una estimación del grado de veracidad del test, es decir, hasta qué punto es adecuado un determinado test para medir la cualidad objeto de estudio.

La **fiabilidad** es una medida de la reproductibilidad o repetibilidad del test.

Finalmente, la **objetividad** hace referencia a la fiabilidad de medidas repetidas, pero efectuadas por diferentes evaluadores. Además, los tests han de ser específicos, sensibles, fáciles de administrar y de fácil interpretación. Además los tests deben ser respetuosos con los derechos del deportista.

1.2.1. Los tests deben ser válidos.

Se dice que un test es válido cuando realmente mide lo que pretende medir. Por ejemplo, supongamos que deseamos determinar la velocidad máxima de carrera de un corredor. Un procedimiento sencillo podría consistir en medir el tiempo invertido en correr una distancia de 60 m, puesto que la mayoría de los corredores alcanzan la velocidad máxima de carrera antes de los 60 m (Groser 1992). Además, existe una relación muy intensa entre la velocidad máxima de carrera y la marca obtenida en 60 m.

Sin embargo, la marca en 60 m también depende de otros factores como la velocidad de reacción ante la señal de salida. Por eso será más válido utilizar otro procedimiento para determinar la velocidad máxima de carrera que permitiera excluir el tiempo de reacción. Esto es fácil de conseguir si se mide la marca obtenida en un test de 60 m con salida lanzada. Una vez alcanzada la velocidad máxima, ésta se mantiene durante 10 ó 20 m, por lo tanto, deberíamos disponer las células fotoeléctricas, al menos cada 5 o 10 m para poder medir la velocidad máxima. Si sólo colocamos una célula a la salida y otra a la llegada, tan sólo podremos calcular la velocidad media en 60 m. Por lo tanto, tendrá *más validez* para determinar la velocidad máxima de carrera un test en el que se determine la velocidad cada 10 m en el curso de una carrera de 60 ó 70 m, que utilizar únicamente la marca obtenida en un test de 60 m.

La marca obtenida en un test de 60 m no es un indicador válido para estimar la capacidad de resistencia cardiorrespiratoria, que guarda más relación con la máxima distancia recorrida en 12 minutos o test de Cooper (Noakes 1988; Sharkey 1990). Así pues, la marca alcanzada en un test de 60 m estima de forma válida la velocidad máxima de carrera, mientras que la máxima distancia recorrida en un test de Cooper no es un test válido para medir la velocidad máxima de carrera.
Para que un test sea válido es imprescindible que sea fiable y relevante. Si un test no es fiable no puede ser válido. No obstante, un test muy fiable puede ser poco válido.

La relevancia hace referencia a la importancia que tiene la variable medida mediante el test para la cualidad física analizada. El grado de validez de un test indica la fidelidad con el que el test es capaz de alcanzar ciertos objetivos de medición.

Por ejemplo, la masa corporal se puede determinar con gran fiabilidad, es decir, si pesamos varias veces en un mismo día al mismo grupo de sujetos obtendremos valores muy próximos. Sin embargo, la determinación de la masa corporal es un test poco válido del porcentaje de grasa corporal, ya que la masa corporal no sólo depende del grado de obesidad, sino que también depende de la talla, el sexo, la constitución corporal y la edad. Por otro lado, el grosor del pliegue cutáneo abdominal se puede determinar con buena fiabilidad y guarda muy buena relación con el porcentaje de grasa corporal. Así pues, para la determinación del porcentaje de grasa corporal carece de validez la determinación de la masa corporal aislada dada su escasa relevancia, mientras que sí es válida la determinación del grosor del pliegue cutáneo abdominal dada su relevancia en relación con el porcentaje de grasa corporal.

1.2.1.1. ¿Cómo se determina la validez de un test?

La validez puede ser establecida por procedimientos lógicos y estadísticos, distinguiéndose dos tipos de validez: validez de contenido y validez relacionada con un criterio.

A la validez relacionada con un criterio también se le llama validez estadística o validez correlacional, ya que se establece por procedimientos estadísticos. Existen dos clases de validez estadística: la validez concurrente y la validez predictiva, ambos tipos de validez se determinan mediante el test de correlación de Pearson La principal diferencia entre la validez concurrente y la validez predictiva reside en el momento en que se efectúan las mediciones con el test objeto de estudio y con el método patrón. Las medidas se obtienen simultáneamente, o casi simultáneamente para efectuar una validación concurrente. En el caso de la validez predictiva, las medidas correspondientes al test objeto de análisis se obtienen antes que las medidas predichas.

Se dice que un instrumento de medición, o test, tiene **validez de contenido** o **validez lógica**, cuando mide variables a partir de las cuales será posible elaborar las conclusiones en las que se fundamentará la evaluación. La validez de contenido se establece por criterios lógicos, simplemente examinando las capacidades que van a medirse y comprobando hasta qué punto la medida obtenida está realmente relacionada con la variable deseada. Generalmente, el establecimiento de la validez de contenido comporta la realización de pruebas complementarias para verificar que se cumplen ciertos requisitos. Por ejemplo, en el caso de la velocidad máxima de carrera, tanto el test de 60 metros como el test de velocidad lanzada (por ejemplo, velocidad lanzada en 30 metros) podrían tener validez de contenido para determinar la velocidad máxima de carrera en seres humanos. Para ello se tendría que comprobar que la distancia y el tiempo de aceleración resultan suficientes para permitir alcanzar la velocidad máxima. Además, se debería determinar durante cuánto tiempo se puede mantener la velocidad máxima mediante tests, en los que se utilizarían distancias superiores a la del test en cuestión y en las que se colocarían múltiples células fotoeléctricas a lo largo del recorrido. De este modo también se podrían detectar los cambios de velocidad durante la prueba y así descartar aquellas distancias a partir de las cuales la fatiga pudiera provocar una pérdida importante de velocidad. Al final de este proceso tanto la lógica, como los datos recogidos mediante pruebas complementarias, nos permitirán propugnar que un test de 60 metros (o 30 metros lanzados) tiene validez para medir la velocidad máxima de carrera en seres humanos. Otro ejemplo: un test destinado a la selección de futuros jugadores de voleibol debería incluir entre las variables analizadas aquellas que resultan determinantes del rendimiento en voleibol como la talla y la capacidad de salto vertical. Por lo tanto, la validez de contenido se establece fundamentalmente a través del análisis lógico de los elementos o pruebas incluidas en el test utilizando ya sea, la propia experiencia o el conocimiento aportado por expertos en libros, artículos, etc.

Para determinar la **validez estadística** se comparan y analizan las relaciones entre los resultados obtenidos mediante el test objeto de estudio y otro test, el que proporciona los datos de referencia o criterio (test patrón). El test patrón será preferiblemente un test de gran validez y fiabilidad, a ser posible el más válido y fiable. Es importante señalar que la

relación entre los resultados obtenidos en dos tests puede ser casi perfecta (r=0.999), pero pueden existir diferencias entre ellos. Esto último ocurre cuando el test sometido a estudio subestima o sobrestima sistemáticamente los resultados obtenidos con el método patrón. Por ejemplo, los valores de porcentaje de grasa corporal obtenidos mediante la ecuación de Yuhasz correlacionan intensamente (r>0.90) con el porcentaje de grasa corporal obtenido mediante un método patrón, no obstante los valores son subestimados. Para descartar la existencia de diferencias significativas entre el método patrón y el test sometido a estudio es necesario efectuar un test de comparación de medias como, por ejemplo, la prueba de la "t de Student para datos apareados".

La **validez concurrente** es una medida de la correlación de un test con cierto criterio patrón. Para determinar la validez concurrente primero hay que escoger un criterio patrón o "gold standard" (en inglés), posteriormente se efectúan mediciones de la variable mediante el procedimiento patrón y el test que estemos analizando.
Posteriormente, se comprueba que no existan diferencias significativas entre los valores medios obtenidos mediante cada test y se calcula el coeficiente de correlación de Pearson. El coeficiente de correlación de Pearson permite determinar si la relación entre dos variables es de tipo lineal, pudiendo adoptar valores comprendidos entre -1 y 1.
Cuanto más cercano a 1 (ó -1) se encuentre el valor del coeficiente de correlación de Pearson, mayor será la validez del test analizado. Cuando el coeficiente de correlación es inferior a 0.6 las diferencias entre el test evaluado y el método patrón resultan excesivas como para poder equiparar ambos tests. Por ejemplo, supongamos que se mide la velocidad de carrera en 100 sujetos y posteriormente la capacidad de salto vertical. Supongamos, nuevamente, que el coeficiente de correlación de Pearson (representado por la letra "r") entre la marca en 60 metros y la altura alcanzada en el salto vertical es de r=0.88. Esto indicaría que se puede estimar la velocidad de carrera a partir de la marca en el salto vertical, pero que el grado de similitud no es muy alto, por lo que nos encontraríamos algunos sujetos que saltando lo mismo que otros, o incluso un poco menos, serían capaces de correr a mayores velocidades. Si el coeficiente de correlación de Pearson fuera 1, esto indicaría que ambos tests responden exactamente en la misma dirección a los cambios experimentados por la variable analizada. Es decir, si un sujeto salta más que otro también será con toda seguridad más veloz. No obstante, tal y como veremos más adelante, el coeficiente de correlación nunca puede ser igual a 1, ya que siempre existe una pequeña variabilidad debida a múltiples factores como la imprecisión de los instrumentos de medida, la variabilidad biológica y las faltas o errores cometidos al efectuar las mediciones.

Obviamente, resulta crucial disponer de un criterio patrón de gran validez para poder evaluar la validez concurrente de tests alternativos. Los criterios patrón pueden obtenerse midiendo magnitudes físicas relacionadas con la cualidad de la condición física en cuestión. Por ejemplo, potencia muscular y capacidad de salto; consumo máximo de oxígeno y capacidad de resistencia, etc.

Muchas veces interesa conocer cuál será la capacidad de rendimiento de un atleta en una competición futura. Esta circunstancia es común al proceso de selección tanto para deportes individuales como para deportes de equipo. Con esta finalidad los entrenadores pasan tests para elegir a aquellos deportistas que parecen hallarse en mejores condiciones de rendir al máximo nivel el día de la competición. Los tests aplicados con esta intención han de tener **validez predictiva**. La validez predictiva también se determina mediante el cálculo del coeficiente de correlación de Pearson entre el rendimiento predicho y el rendimiento realmente alcanzado el día de la competición, por lo tanto este tipo de validación se efectúa a posteriori.

Por lo que respecta a las variables relacionadas con la condición física la validez predictiva está muy condicionada por el tiempo transcurrido entre la realización del test y el día de la competición. Evidentemente, cuanto más lejana sea la predicción mayores son las posibilidades de error, es decir, menor es la validez predictiva. Sin embargo, muchos tests no deben administrarse justo antes de las competiciones puesto que podrían disminuir la capacidad de rendimiento en las horas o días siguientes. Por ello es conveniente que el test tenga validez predictiva aún cuando se administre mucho antes de que tenga lugar la competición. En cualquier caso, es necesario conocer el error asociado a cada predicción.

Así mismo, hay que tener en cuenta que el error de predicción fuera del intervalo de medición puede ser muy elevado. Por ejemplo, supongamos que hemos desarrollado un test de marcha que nos permite predecir el VO_2max y que el VO_2max de la población empleada para desarrollar el test estaba comprendido entre 35 y 70 $ml.kg^{-1}.min^{-1}$. Si administramos este test de marcha a un corredor y obtenemos un valor predicho de 80 $ml.kg^{-1}.min^{-1}$, el error asociado a esta predicción será mucho mayor que el error asociado a una predicción que cayera dentro del intervalo comprendido entre 35 y 70 $ml.kg^{-1}.min^{-1}$. Algo parecido ocurre con los tests desarrollados para predecir marcas.
Otro aspecto que debemos destacar es que, en ocasiones, determinado test sólo es válido cuando se administra a cierto grupo poblacional. Por ejemplo, se han desarrollado procedimientos antropométricos para estimar el porcentaje de grasa corporal, basados en la medición del grosor de pliegues cutáneos. Estas técnicas aplican fórmulas que se han obtenido midiendo los pliegues en un grupo de sujetos y determinando el porcentaje de grasa corporal mediante un procedimiento criterio, generalmente por hidrodensitometría. A partir de los resultados de la hidrodensitometría se pueden obtener ecuaciones que relacionan el grosor de los pliegues con el porcentaje de grasa corporal. Pues bien, la mayoría de estas ecuaciones son específicas de población, es decir, sólo son válidas para ser aplicadas a sujetos pertenecientes a la misma población que se empleó para confeccionar la fórmula antropométrica. Así, una ecuación antropométrica que ha sido obtenida a partir de una población de mujeres tiene escasa validez para medir el porcentaje de grasa corporal en un grupo de hombres.

1.2.2. Los tests deben ser fiables.

La fiabilidad hace referencia a la reproductibilidad o repetibilidad del test. Un test es fiable cuando al efectuar varias mediciones de una determinada variable, cuyo valor no ha cambiado, los resultados obtenidos son consistentes, es decir similares. En ocasiones puede ocurrir que un test tenga una gran reproductibilidad, pero que los valores obtenidos se alejen sistemáticamente del valor real. En este último caso el test será reproducible, pero poco exacto. La **exactitud** representa la fidelidad con la que es posible representar cada valor real, por lo que depende de la precisión de los instrumentos de medida. La precisión hace referencia a la magnitud absoluta de la diferencia mínima entre dos medidas sucesivas que puede ser detectada. Lo ideal sería que los tests fueran la vez reproducibles, exactos y precisos, no obstante, si son muy reproducibles es posible corregir los errores sistemáticos (o desviaciones del valor real), ya sean constantes o proporcionales.

Por ejemplo, una báscula que es capaz de detectar cambios mínimos en la masa corporal de las personas de hasta 50 g será más precisa que otra, que sólo pueda detectar cambios mínimos superiores a 200 g. Así mismo, una báscula tendrá una gran reproductibilidad si al pesar siempre da los mismos resultados. Supongamos que la lectura de la báscula es 52.0 y 102.0 kg cuando colocamos sobre ella masas reales de 50.00 y 100.00 kg. Esta báscula tendrá una gran fiabilidad pero será poco exacta pues comete un error constante de 2.0 kg que, una vez identificado, puede ser corregido. También podría ocurrir que la lectura obtenida en otra báscula fuera siempre de 52.0, 104.0 y 156.0 kg al colocar sobre ella masas reales de 50.00, 100.00 y 150.00 kg. En este último caso la báscula también sería tan reproducible como la primera, pero menos exacta (la diferencia entre el valor observado y el valor real es mayor). Además, la segunda báscula cometería un error proporcional al sobrestimar en un 2 % el valor real de la masa. Los errores sistemáticos son fácilmente detectables y se pueden corregir, generalmente, por procedimientos informáticos, incluidos en las rutinas de calibración de los instrumentos modernos. Un test perfectamente fiable, o fiable al 100 %, daría resultados siempre idénticos. Esta circunstancia sólo es posible cuando la variable es constante (no varía) y cuando no existe error de medida. Sin embargo, ningún test es fiable al 100 % o perfecto, ya que todo proceso de medición lleva asociado un error. Así, se puede considerar que los valores observados (valores obtenidos en un proceso de medición) resultan de la suma del valor verdadero más el error. Es importante señalar con respecto al error que:

- El error puede adoptar valores positivos y negativos, incrementando o disminuyendo el valor verdadero y, en consecuencia, el valor observado.

- El valor medio de los errores es igual a cero.

- Tanto los valores observados, como los valores verdaderos y los errores presentan variabilidad.

- La varianza de los valores observados (S_o) es igual a la suma de la varianza de los valores verdaderos (S_v) más la varianza de los errores (S_e).

$$S_o = S_v + S_e$$

La varianza de los errores depende de múltiples factores, que contribuyen en mayor o menor medida al error total. Entre ellos destacan:

- El error de medida de los instrumentos.

- La imprecisión de las mediciones debidas a diferencias en el proceso de administración del test.

Un test es fiable si cuando es administrado por diferentes testadores y en diferentes entornos ofrece los mismos resultados. Es decir, las diferencias achacables al proceso de administración del test han de ser mínimas. Para ello es necesario que los testadores dominen perfectamente las técnicas de medición y que los sujetos estén familiarizados con el procedimiento.

- La variabilidad biológica.

1.2.2.1. Las mediciones múltiples.

Para reducir la variabilidad en los resultados obtenidos se suelen emplear mediciones múltiples. La cuestión que surge es qué valor tomar cómo representativo de la medición efectuada cuando se practican múltiples mediciones. Se han propuesto dos métodos alternativos: calcular el valor medio o tomar el valor máximo. Los defensores de la utilización del valor medio aducen que de este modo se reducen los errores debidos a la falta de reproductibilidad por errores de medición y por variabilidad biológica. Por ejemplo, el grosor de los pliegues cutáneos se suele determinar como el valor medio de, al menos, tres mediciones consecutivas. Sin embargo, cuando se desea conocer cuál es el nivel máximo de rendimiento que es capaz de alcanzar un sujeto en determinado test, parece más apropiado tomar el valor máximo obtenido en las múltiples mediciones efectuadas, de modo similar a cómo se realiza en el deporte. Por ejemplo, la marca de un saltador en competición se obtiene computando el mejor salto, no como el valor medio de los saltos válidos.

En general, cuanto mayor es el número de mediciones más reproducibles serán los resultados, con la salvedad de que cada medición debe efectuarse en las mismas condiciones, evitando que la fatiga pueda alterar el rendimiento en cada repetición. El número final de mediciones dependerá en cada caso de la cualidad a medir, de las características del test y de la disponibilidad de tiempo. No obstante, la realización de

múltiples mediciones puede provocar, por sí misma, una mejora en el rendimiento por un efecto aprendizaje y/o por efecto de entrenamiento. El efecto aprendizaje se manifiesta especialmente en tests relativamente complejos como suelen ser los tests que pretenden medir las capacidades coordinativas o habilidades deportivas específicas. El efecto entrenamiento tiene lugar cuando la repetición del test mejora la capacidad de rendimiento del sujeto evaluado. Esta última situación puede darse con cierta facilidad cuando la cualidad valorada puede cambiar rápidamente o es muy entrenable, cómo puede suceder con la movilidad articular o con la fuerza muscular en sujeto que nunca ha entrenado fuerza. Tanto el efecto aprendizaje como el efecto entrenamiento se dan con más facilidad en los sujetos no entrenados. Obviamente, los sujetos no entrenados son más entrenables, por lo que cuando se evalúa la condición física en esta población hay que ser menos ambiciosos en lo que respecta al número de mediciones a efectuar.

2.2.2.2. ¿Cómo se determina el grado de fiabilidad de un test?

La fiabilidad (r_{xx}) se puede definir como la proporción de la varianza de los valores observados achacable a la varianza de los valores verdaderos:

$$r_{xx} = S_v/S_o = (S_o - S_e)/S_o$$

Si fuera posible efectuar mediciones sin error, S_e sería igual a cero, por lo que el coeficiente de fiabilidad r_{xx} sería igual a 1. A medida que aumenta la magnitud del error el valor de r_{xx} disminuye aproximándose a 0 en los casos extremos. En general, lo deseable es trabajar con tests cuya fiabilidad sea superior a 0.8 (Morrow y col. 1995).

Aunque no es posible determinar de forma directa el valor verdadero de cada observación, sí se pueden obtener directamente los valores observados, así como su varianza (S_o). Por otro lado, existen varios procedimientos que permiten estimar la varianza del error (S_e) repitiendo, en las mismas condiciones, dos o más veces las mediciones. Los coeficientes de fiabilidad se han clasificado en dos grandes categorías:

coeficientes **interclase** (basados en el cálculo del coeficiente de correlación de Pearson)

y coeficientes **intraclase** que se obtienen por análisis de la varianza (ANOVA).

Además, también se puede utilizar como índice de la fiabilidad de un test el coeficiente de variación.

El coeficiente de correlación interclase se puede determinar por el método de test-retest, que consiste en el cálculo del coeficiente de correlación de Pearson entre dos mediciones

efectuadas en la misma población, en idénticas condiciones. El valor mostrado por el coeficiente de correlación de Pearson se toma como coeficiente de fiabilidad y es denominado, en ocasiones, coeficiente de correlación interclase. Cuanto más fiable sea el test, más próximo a 1 resultará el valor del coeficiente de fiabilidad. Si el tiempo transcurrido entre las dos determinaciones en las que se basa el cálculo del coeficiente de correlación test-retest es prolongado (días o semanas) el test de correlación test-retest puede ser, entonces, denominado coeficiente de estabilidad.
La mayoría de las cualidades físicas pueden ser medidas con coeficientes de fiabilidad de 0.80 a 0.95. No obstante, el valor de fiabilidad obtenido es sólo una aproximación, una estimación de la fiabilidad real del test. No hay que olvidar que el coeficiente de fiabilidad obtenido es específico de la población empleada en su cálculo y que puede variar en función de múltiples aspectos relacionados con la administración del test. Un coeficiente de fiabilidad cercano a 1 indica que el error de medida ha sido pequeño, que el instrumento de medida es fiable (aunque no necesariamente preciso) y que la variable analizada ha permanecido más o menos estable durante el tiempo transcurrido entre las dos mediciones.

Más recientemente, se ha propugnado el cálculo del coeficiente de correlación intraclase como coeficiente de fiabilidad, debido a que el establecimiento de la fiabilidad de un test por el procedimiento de test-retest tiene ciertas limitaciones. En primer lugar, sólo pueden emplearse dos mediciones por persona para calcular el coeficiente de correlación de Pearson. Si se han obtenido más de dos mediciones por persona, hay que reducirlas a dos, ya sea descartando las restantes o dividiendo el total de mediciones efectuadas en dos grupos (por ejemplo mediciones de orden par e impar). Posteriormente se calcula la media de cada grupo de mediciones. De este modo a cada sujeto se le asignan dos medidas que pueden emplearse para obtener el coeficiente de correlación de Pearson. Esta técnica es aceptable para variables que permanecen relativamente constantes.Otra limitación de la técnica test-retest es que no aporta ninguna información acerca de las diferencias absolutas entre las dos mediciones efectuadas, lo cual obliga a descartar la existencia de diferencias significativas entre ambas mediciones aplicando, por ejemplo, un test de comparación de medias como la prueba de la "**t de Student** ". Un índice de correlación de Pearson elevado junto con diferencias significativas entre ambos tests (prueba "t de Student" significativa) sugiere la existencia de un error sistemático. Por estas razones, numerosos autores defienden la determinación del coeficiente de fiabilidad intraclase como el procedimiento más adecuado para obtener el coeficiente de fiabilidad de un test (Kroll 1962; Safrit 1981; Baumgartner y Jackson 1987).

El coeficiente de correlación intraclase permite emplear en el cómputo del coeficiente de fiabilidad más de dos mediciones por persona y toma en consideración el valor medio de la medidas efectuadas así como su dispersión (desviación estándar), proporcionando una estimación más apropiada de la fiabilidad del test en cuestión. Los modelos más utilizados para el cálculo de los coeficientes de fiabilidad intraclase son el coeficiente alfa de Cronbach, la Fórmula 20 de Kuder-Richardson (KR_{20}) y el coeficiente de correlación

intraclase obtenido mediante análisis de varianza (Baumgartner y Jackson 1987; Morrow y col. 1995).

Otro procedimiento muy utilizado para estimar la fiabilidad de un test es ladeterminación del coeficiente de variación individual y conjunto. El **coeficiente de variación individual** (CVi) y el **coeficiente de variación conjunto** (CVc) se obtienen efectuando múltiples mediciones en varios sujetos para, posteriormente, determinar el valor medio y calcular la desviación estándar de los resultados obtenidos, aplicando las fórmulas:

$$CVi = \frac{SD \times 100}{\overline{x_i}} \qquad CVc = \frac{\sqrt{\frac{\sum (n-1) \times SD^2{}_i}{\sum (n-1)}}}{\overline{x}} \times 100$$

Donde, "CV_i" es el coeficiente de variación intrasujeto, "SD_i" es la desviación estándar de los valores adoptados por la variable estudiada en las diferentes pruebas que realizó un mismo sujeto, "x_i" es la media aritmética de dichos valores, "n_i" es el número de pruebas realizadas por cada sujeto, "x" es la media conjunta de todos los valores adoptados por la variable en todos los sujetos y "CV_c" es el coeficiente de variación conjunto. En la Tabla 2.2 se muestra un ejemplo de cálculo del coeficiente de variación individual y conjunto de los valores de potencia máxima alcanzada al final de un test de esfuerzo de intensidad progresivamente creciente hasta el agotamiento. En la tabla 2.3, también se pueden observar los coeficientes de variación individuales, así como el coeficiente de variación conjunto de la determinación de la lactatemia máxima, tras un esfuerzo supramáximo repetido en tres ocasiones.

En un estudio reciente 6 estudiantes de Educación Física efectuaron en días separados 4 tests de esfuerzo en cicloergómetro. Se calcularon los coeficientes de variación individuales y el coeficiente de variación conjunto aplicando la fórmula expuesta en el apartado anterior. Obsérvese que para calcular el coeficiente de variación conjunto se calcula primero la media ponderada de las desviaciones estándar individuales, tal y como hemos señalado en el apartado anterior.

Tabla 2.2. Valores de potencia máxima (Wmax) medidos en 4 ocasiones, a lo largo de 8 semanas en 6 estudiantes de Educación Física. (x_e: valor medio entre sujetos; SD_e: desviación estándar entre sujetos; x_i: valor medio intrasujetos; SD_i: desviación estándar intrasujetos; CVi: Coeficiente de variación individual).

Sujetos	Wmax1	Wmax2	Wmax3	Wmax4	x_i	SD_i	CV_i
1	357	363	331		350.3	171	**49**
2	338	337	354	354	345.8	95	**27**
3	254	267	281	270	268.0	111	**41**
4	335	339	336	335	336.3	19	**6**
5	341	361	357	352	352.8	87	**25**
6	340	354	347	360	350.3	87	**25**
x_e	327.5	336.8	334.3	337.3	334.0		
SD_e	36.8	35.9	28.0	34.0	31.8		

La determinación de los coeficientes de variación ha sido muy utilizada, especialmente para comparar la variabilidad que presentan distintos tests entre sí. En general, podemos considerar como aceptables aquellos tests de condición física que presenten coeficientes de variación inferiores al 10%. Cuanto mayor sea el coeficiente de variación de un test, mayor tendrá que ser la magnitud de los cambios producidos por el entrenamiento (o cualquier otra intervención), para que puedan ser detectados por el test en cuestión. Aunque el coeficiente de variación es más fácil de calcular y parece de más fácil interpretación, para determinar la fiabilidad de un test lo más adecuado es calcular el coeficiente de fiabilidad intraclase.

1.2.2.3. Factores a considerar en la determinación de la fiabilidad de un test.

En la determinación del grado de fiabilidad de un test debemos considerar algunos factores que pueden influir en el resultado del test de fiabilidad, de tal manera que podemos esperar un grado aceptable de fiabilidad en aquellos tests en los que se den las siguientes condiciones:

1) *Que la población en la que se va a administrar el test sea heterogénea en el comportamiento de la variable analizada.* Decimos que un grupo es **homogéneo** cuando las diferencias entre sujetos son pequeñas. Por ejemplo, si determinamos el VO_2max en corredores de maratón de elite (nivel internacional), las diferencias entre sujetos son muy pequeñas, por lo general, inferiores a 10 $ml.kg^{-1}.min^{-1}$, para un valor medio que se sitúa próximo a los 80 $ml.kg^{-1}.min^{-1}$. Es decir, podemos considerar que la población de corredores de maratón de elite es homogénea para el VO_2max. Sin embargo, el grupo sería **heterogéneo** para el VO_2max si incluyera una muestra representativa del universo de todos los seres humanos normales, que va desde ancianos con valores de VO_2max inferiores a 15 $ml.kg^{-1}.min^{-1}$, hasta corredores de maratón de elite, que pueden alcanzar valores entre 80 y 90 $ml.kg^{-1}.min^{-1}$. Si los sujetos son muy homogéneos en el comportamiento de la variable analizada, la fiabilidad del test será baja. Esto es debido a que el test no tendrá suficiente sensibilidad para detectar diferencias entre sujetos, quedando éstas, en ocasiones, enmascaradas por la variabilidad biológica y el error de medida. En estas

circunstancias hay que modificar el test para que tenga una mayor sensibilidad, o bien hay que cambiar a otro tipo de test que tenga mayor poder discriminante entre sujetos.

2) *Que los sujetos estén muy motivados para efectuar el test con el objetivo de obtener la máxima puntuación.* Para lograr estas condiciones es conveniente crear cierto ambiente de competición entre los individuos que van a ser medidos. En ocasiones, contribuye a mantener el nivel de motivación la actitud del evaluador, animando continuamente a los sujetos a esforzarse al máximo.

3) *Que los sujetos estén preparados para ser sometidos al test,* es decir, que conozcan bien el tipo de test que van a efectuar, que hayan calentado adecuadamente, etc...

4) *Que se tomen suficientes mediciones* en cada sujeto como para poder poner de manifiesto su má xima capacidad de rendimiento, sin olvidar las limitaciones comentadas anteriormente con respecto a la repetición del test.

5) *Que el ambiente y la organización de todo el proceso relacionado con el test, sean propicios para facilitar el máximo rendimiento.* Hay que conseguir similares condiciones ambientales (temperatura, humedad, viento, público, etc.) y una técnica similar al administrar el test en repetidas ocasiones.

6) *Que el responsable de la administración del test sea competente,* o sea que conozca perfectamente los procedimientos a emplear, que esté motivado, que efectúe su trabajo con un buen nivel de concentración y sea capaz de recoger los datos de forma segura (es recomendable utilizar sistemas de registro que eviten errores de notación e introducción de datos).

1.2.3. Los tests deben ser precisos.

Uno de los aspectos cruciales en cualquier proceso que entrañe la medición de variables es conocer la **incertidumbre** o **imprecisión** con la que se realiza la medición. La imprecisión estima el error asociado al proceso de medición. Los valores de imprecisión se obtienen asumiendo que la magnitud medida tiene un valor constante durante la medición y que el proceso de medición no modifica de forma impredecible el valor de la magnitud medida. La imprecisión determina la diferencia mínima detectable entre dos valores de la variable analizada. Debemos señalar, que mientras no nos indiquen lo contrario los valores de imprecisión se han de considerar como indicativos de la imprecisión máxima en la escala completa. Puesto que la imprecisión suele ser mayor en los extremos de la escala de medida, cuando se emplea el instrumento de medición en la zona media de la escala de medición suelen obtenerse valores de imprecisión más bajos.

Otros aspectos muy importantes por su influencia en la precisión de las mediciones están relacionados con la sensibilidad de los instrumentos y con el rango de medición.

El término **sensibilidad** (sensitivity, en inglés; no confundir con el término inglésensibility) se utiliza para referirse al valor mínimo detectable. El **rango** de medición delimita el valor mínimo y máximo, o límites inferior y superior, respectivamente, de la escala de medida. Por lo tanto, la elección de un instrumento de medida debe fundamentarse en el conocimiento de los valores que puede tomar la variable analizada, de tal manera que los valores medidos estén incluidos en el rango de medida de los instrumentos a utilizar. También se debe conocer cuál es la imprecisión o tolerancia admitida.

Supongamos que deseamos medir el tiempo de vuelo en un salto vertical mediante filmación. Si utilizamos una cámara de vídeo capaz de filmar a una velocidad máxima de 50 imágenes por segundo y una cámara de cine capaz de alcanzar las 1000 imágenes por segundo, la imprecisión de las mediciones será de 20 veces superior al filmar con la cámara de vídeo (20 ms) que al utilizar la cámara de cine (1 ms). Sin embargo, en este tipo de salto hay dos momentos críticos el momento del despegue (pérdida del contacto de los pies con el suelo) y el aterrizaje (nuevo contacto de los pies con el suelo). Por lo tanto, la imprecisión con la que podremos determinar el tiempo de vuelo será posiblemente aún mayor para ambos instrumentos y ésta es la imprecisión que deberíamos atribuir a nuestro test. Otro ejemplo, lo constituyen las celdas de carga cuya imprecisión y rango de medición viene reflejados en las especificaciones técnicas.

Una celda de carga de 1 a 200 kp de rango y una imprecisión del 0.1% escala completa, sería aquella que, en las peores condiciones, cometería un error equivalente al 0.1 % del valor absoluto observado. Así, si la celda de carga registra un valor de 50 kp, el valor real estará comprendido, con un 95 % de probabilidades, entre 49.95 y 50.05 kp. Si el valor registrado fuera de 2 kp, el valor real se encontraría ente 1.998 y 2.002 kp.

La fiabilidad de un test viene condicionada, en primer lugar por la imprecisión con que pueden ser obtenidas las medidas, hasta el extremo que un test muy impreciso será muy poco fiable. Las diferentes causas de imprecisión o incertidumbre se han clasificado en (García de la Chica 1991):

1) Imprecisión debida a la magnitud que se mide.
2) Imprecisión debida al instrumento de medida.
3) Imprecisión debida a las correcciones.
4) Imprecisión debida al procedimiento de medida.

1.2.3.1. Imprecisión debida a la magnitud que se mide.

Las características de magnitud a medir tienen una gran influencia sobre el proceso metrológico. Así, es muy diferente intentar cuantificar una longitud, una masa o, por ejemplo, el volumen de un gas. La imprecisión vinculada a la magnitud que se mide se ha asociado a:

- *La inestabilidad de la magnitud a medir.* En ocasiones, la magnitud a medir no permanece constante, sino que varía con el tiempo. Esta condición se da en todas las variables biológicas y se la conoce como **variabilidad biológica**.

- *La influencia de las condiciones externas.* La mayoría de las magnitudes son sensibles a las condiciones ambientales en que se efectúa la medición. Así, por ejemplo, es bien conocida la influencia que tiene la presión y la temperatura sobre el volumen de los gases, etc.

- *Por características intrínsecas de la variable medida.* Por ejemplo, la medición del grosor de un pliegue graso, no sólo viene determinada por la adiposidad del mismo, sino que también depende de la elasticidad del tejido celular subcutáneo y su contenido de agua. Como consecuencia de lo anterior, la medición de una variable relacionada con la condición física comporta además del **error tecnológico** (imprecisión de los instrumentos) otro, generalmente mayor, debido a la variabilidad biológica. El error debido a la variabilidad biológica obedece al comportamiento oscilante de los parámetros biológicos, muchos de ellos sometidos a mecanismos de servocontrol. El desplazamiento de una variable del punto de equilibrio desencadena una serie de respuestas para ajustarla nuevamente a su nivel de equilibrio. Pero los mecanismos que tratan de restablecer el equilibrio provocan un ligero desplazamiento de la variable más allá del punto de equilibrio, lo que a su vez genera una respuesta de servocontrol en sentido contrario, y así sucesivamente, hasta alcanzar un nivel de oscilación mínimo modulado por las perturbaciones ambientales.

Existen variables biológicas que oscilan con frecuencias muy altas como, por ejemplo, la presión arterial y otras que oscilan con frecuencias mucho más bajas, como ocurre con la ritmicidad que exhibe la concentración plasmática de estradiol a lo largo del ciclo sexual femenino. Además, una misma variable puede mostrar oscilaciones de alta y baja frecuencia superpuestas, tal es el caso de la testosterona que expresa numerosas oscilaciones en 1 hora, muestra un ritmo circadiano (con valores máximos al amanecer y mínimos al anochecer) y otro ultradiano (las concentraciones de testosterona en plasma son más elevadas en primavera-verano que en otoño-invierno) (Campbell y col. 1982; Riad-Fahmy y col. 1982; Touitou y col. 1990).

Lamentablemente desconocemos las características oscilatorias de numerosas variables biológicas, siendo este desconocimiento aún mayor en el ámbito de las variables relacionadas con el ejercicio físico. Aparte de las fluctuaciones periódicas que

experimentan ciertas variables biológicas relacionadas con el rendimiento deportivo, existen otros factores que pueden alterar su nivel. Entre ellos se han citado las condiciones ambientales (viento, temperatura, humedad, luminosidad, etc.), factores relacionados con los materiales y superficies en las que tiene lugar la actividad deportiva, el grado de condición física, la periodización de los entrenamientos y factores psicológicos (Kuipers y col. 1985).

1.2.3.2. Imprecisión debida al instrumento de medida.

Guarda relación con la incertidumbre con que se ha calibrado el instrumento de medida y con la influencia que puedan ejercer las condiciones ambientales sobre el comportamiento del instrumento de medida. Por ejemplo, una cinta métrica metálica con una escala en milímetros, puede emplearse para medir longitudes siempre y cuando podamos asumir un error superior a 1 milímetro y las condiciones de temperatura sean más o menos estables durante las mediciones. Una temperatura muy elevada provocaría una subestimación de la longitud real, mientras que una temperatura muy baja conduciría a una sobrestimación. En general, los instrumentos de medición vienen provistos de instrucciones que indican cuáles son las condiciones de uso. Además, en ocasiones es posible aplicar coeficientes de corrección, para contrarrestar desviaciones debidas a diferencias en parámetros ambientales. Tal es el caso de los sistemas destinados a medir el consumo máximo de oxígeno, los cuales son muy sensibles a los cambios de las condiciones medioambientales.

Como norma general, es recomendable que la precisión de los instrumentos de medida sea un orden de magnitud superior a la precisión deseada en la medición. Por ejemplo, si queremos medir longitudes con una imprecisión de 1 mm, deberíamos utilizar un instrumento con una imprecisión igual o inferior a 0.1 mm.

1.2.3.3. Imprecisión debida a las correcciones.

Existen numerosas correcciones que pueden ser introducidas en la medida, como las citadas anteriormente debidas al efecto de la temperatura sobre una determinada magnitud. Por lo que respecta a la valoración de la condición física en campo es muy importante medir la velocidad del viento. Posteriormente, se pueden corregir los resultados obtenidos o bien, repetir los tests en condiciones de viento aceptable (menos de 2 $m.s^{-1}$).

1.2.3.4. Imprecisión debida al procedimiento de medida.

Son numerosas las causas de imprecisión ligada al procedimiento de medida. Por ejemplo, cuando se trabaja con seres vivos en los que van a medirse diversas variables, el orden de las mediciones puede influir de forma considerable en la imprecisión con que se realiza cada medición. También puede influir en la imprecisión el tiempo empleado en el proceso

de medición (muy importante en determinadas técnicas de análisis enzimático, como por ejemplo, las que se utilizan para la determinación de la concentración de lactato en sangre). Este tipo de imprecisión debe distinguirse de las faltas o errores gruesos, los cuales son debidos a la aplicación incorrecta de la técnica de medición.

1.2.4. Los tests deben ser específicos.

La especificidad de un test guarda gran relación con la validez. Es más probable que un test muy específico sea válido, mientras que es menos probable que un test poco específico sea válido. La especificidad depende de la similitud existente entre los actos motores requeridos por el test y los solicitados por la cualidad física analizada. Por ejemplo, aunque la fuerza es importante para poder desplazarse velozmente, es más específico para medir la velocidad de carrera un test de velocidad lanzada, que un test de salto horizontal a pies juntos o un test de fuerza isométrica máxima de la musculatura extensora de las piernas.

En el caso de las habilidades deportivas, éstas tienen varios componentes, por lo que un test destinado a medir cierta habilidad deportiva debe requerir a los distintos componentes que intervienen en dicha habilidad. Así por ejemplo, la velocidad de desplazamiento de un futbolista con balón depende de la velocidad de carrera, pero también de la coordinación y de la percepción espacio-temporal. Lo específico sería medir la velocidad de desplazamiento mediante un test que requiera de la máxima velocidad de desplazamiento al tiempo que se conduce el balón.

Otro aspecto importante de la especificidad reside en que los tests no sólo deben requerir actos motores similares a los que integran la habilidad deportiva, si no que también deben solicitar a las distintas cualidades que intervienen en la habilidad deportiva y, además, en proporciones similares. Supongamos que interesa conocer cuál es la habilidad de un jugador de baloncesto como lanzador a canasta y que con este propósito se emplea un test de 10 lanzamientos. Cuando el jugador lanza desde 4 metros se observa que suele acertar un 90 % de los intentos efectuados, pero si elegimos una distancia de lanzamiento de 6 metros su tasa de acierto desciende al 15 %. Supongamos que la mayoría de los jugadores de baloncesto que tienen una tasa de acierto del 90 % cuando lanzan desde 4 metros, suelen acertar también un 40 % los lanzamientos efectuados desde 6 metros. Ante esta circunstancia, al entrenador le interesará saber por qué falla más el jugador en cuestión. ¿Le falta fuerza, tiene una menor agudeza visual, o simplemente se trata de una distancia que no ha entrenado? En cualquier caso, lo más importante es determinar cuál es la distancia media a la que tiene que lanzar este jugador durante los partidos, si nunca se le requiere que lance desde distancias superiores a 4 m tiene poco sentido evaluarlo en esa distancia. Pero además, en baloncesto los lanzamientos se efectúan desde distintas posiciones: frente al aro, desde los lados, etc. Además, en muchas ocasiones los defensores intentan impedir el lanzamiento. Estos últimos aspectos deberían ser considerados al elegir un test de lanzamiento a canasta, para que sea específico del deporte baloncesto. No

sorprende pues que sea mucho más difícil el desarrollo de tests con buena especificidad para la valoración de habilidades deportivas.

1.2.5. Los tests deben ser objetivos.

En ocasiones al entrenador le interesa determinar el grado de perfección biomecánica con el que se han pasado las vallas. A un gimnasta le preocupa determinar el grado de perfección en la ejecución de los movimientos gimnásticos. A un instructor de pilotos le puede interesar la rapidez y adecuación de las decisiones tomadas ante una situación de emergencia. En estos tres últimos casos, como criterio patrón se emplean tablas de puntuación elaboradas por expertos. La puntuación alcanzada en cada situación es otorgada por jueces. Para tratar de incrementar la validez en la puntuación asignada a cada sujeto es necesario que los jueces sean objetivos y que se promedien las puntuaciones otorgadas por varios jueces, incluso descartando los valores extremos.

La objetividad o fiabilidad de puntuación es una característica importante que deben reunir los tests. *Un test es objetivo cuando existe una gran concordancia entre las medidas obtenidas por distintos evaluadores.* Por lo tanto, un test podrá ser completamente objetivo cuando esté totalmente automatizado. Por ejemplo, un test de velocidad en natación será objetivo cuando los tiempos obtenidos dependan de la activación de sistemas de cronometraje automático. Si el cronometraje es manual, el grado de objetividad vendrá determinado por el nivel de concordancia de los tiempos registrados por los distintos jueces. No obstante, el grado de subjetividad es mucho mayor en los tests que requieren de la asignación de puntuaciones en función de criterios preestablecidos, según el grado de perfección en la ejecución de habilidades deportivas. Así, la puntuación en una competición de saltos de trampolín será objetiva cuando los valores otorgados por cada uno de los jueces sean similares.

1.2.6. Los tests deben ser sensibles.

Los tests deben ser capaces de distinguir entre distintos niveles de rendimiento. Cuanto más sensible es el test con más facilidad pueden ponerse de manifiesto pequeñas diferencias entre sujetos, o ligeros cambios en el nivel de rendimiento de un mismo sujeto. *Para que un test sea sensible es imprescindible que sea preciso*. En general, *es recomendable que la sensibilidad del test sea, al menos, 10 veces superior a la variabilidad biológica* de la variable que vaya a analizarse. De esta forma se conseguirá que el test sea más fiable. Por ejemplo ¿con qué sensibilidad hay que medir los cambios de masa corporal?. Si el peso o masa corporal presenta una variabilidad cercana al 1.5 %. Esto representaría unos 1000 g para un sujeto que pesara 70 kg, por lo tanto, lo recomendable es emplear una báscula cuya sensibilidad alcance los 100 g.

Efectivamente, para la valoración de la masa corporal se emplean básculas con una sensibilidad de 50 a 100 g. Emplear básculas de mayor precisión comportaría un dispendio económico muy superior, sin que el incremento conseguido en la sensibilidad fuera a contribuir de forma apreciable a aumentar la fiabilidad en la determinación de la masa corporal.

1.2.7. Los tests deben ser fáciles de administrar e interpretar.

Cuánto más fácil es la administración de un test menor es la probabilidad de cometer errores al administrarlo, lo que en definitiva contribuye a aumentar su fiabilidad. Por otro lado, los tests fáciles de administrar requieren menos tiempo por lo que resultan más apropiados cuando se desea medir a un gran número de personas. Además, cuanto más fácil es el test, menor suele ser el tiempo requerido para que los sujetos se familiaricen con el mismo, resultando también menor el tiempo necesario para formar a las personas que van ser responsables de su administración. Todo ello redunda en un menor coste económico del proceso de evaluación. En cualquier caso, el grado de complejidad del test debe ser adecuado al nivel y capacidad de los deportistas a evaluar. Además, los tests que resultan divertidos para el deportista y evaluador, así como aquéllos que no requieren de especiales medidas de seguridad, resultan más motivantes y facilitan el proceso de administración del test.

Otro aspecto importante hace referencia a la interpretación de los resultados. Cuanto más sencillo e inteligible, tanto mejor. Los tests difíciles de interpretar requieren evaluadores más experimentados, consumen más tiempo y comportan un mayor riesgo de error. Es importante señalar que complejidad tecnológica no es equivalente a fiabilidad. En ocasiones resulta mucho más fiable para medir algunas variables emplear un test poco instrumentado (más simple) que otro tecnológicamente muy complejo. Por ejemplo, para controlar la evolución de la capacidad de resistencia de un nadador resulta preferible efectuar un test de nado de 30 minutos y medir la distancia total recorrida, que determinar su VO_2max mediante un test efectuado en tapiz rodante o en cicloergómetro.

1.2.8. Los tests han de ser respetuosos con los derechos humanos del deportista.

Los criterios éticos que deben presidir la administración de un test destinado a la valoración de la condición física incluyen: *una explicación clara y detallada de los propósitos del test, una exposición objetiva de los riesgos físicos y psíquicos que comporta el test y la garantía a la confidencialidad de los resultados*. Especial atención merecen los menores de edad, que sólo pueden ser sometidos a determinadas pruebas con la autorización de los padres o del tutor legal.

Algunos tests requieren estrictas medidas de seguridad y protocolos de actuación para poder responder adecuadamente ante un contratiempo. Tal es el caso de los tests de esfuerzo hasta el agotamiento, destinados a determinar la capacidad de resistencia. Al respecto el Colegio Americano de Medicina del Deporte (ACSM: Guidelines for Exercise Testing and Prescription 1991) establece:

- *Nunca debe efectuarse un test cuando existen dudas en cuanto a la seguridad del test.*

- *Los tests de esfuerzo hasta el agotamiento efectuados en un laboratorio de ergometría deben realizarse siempre bajo supervisión médica.*

No obstante, se admite la realización de tests de esfuerzo hasta el agotamiento en instalaciones deportivas, por personal no médico pero adecuadamente formado, siempre y cuando los sujetos testados sean personas sanas. El estado de salud de los participantes debe ser establecido por un médico.

Hay que tener presente que los atletas y los deportistas bien entrenados poseen la determinación y la persistencia para llevar a sus sistemas hasta el máximo, incluso cuando dichos tests son efectuados sin que se halle presente un médico. Estas razones llevan a Thoden (1991) a recomendar que no se realicen tests hasta el agotamiento en deportistas que no hayan sido evaluados previamente por un médico, conocedor de la naturaleza de los tests a los que es sometido el atleta. De ahí que se haya aconsejado que los deportistas sometidos a entrenamiento de alta intensidad y que efectúan tests de esfuerzo hasta el agotamiento con cierta asiduidad, sigan un control médico periódico (Thoden 1991). De este modo, no sólo será posible disminuir el riesgo de accidentes, sino que también se podrán detectar precozmente problemas médicos que podrían disminuir el rendimiento o la "entrenabilidad" del deportista.

1.3. Procedimientos a seguir en la administración de los tests.

El primer problema con el que se enfrenta el evaluador para poder valorar una cualidad física, o una habilidad deportiva, es elegir el test, o los tests, que va a emplear de entre la gran variedad de tests descritos en la literatura. La elección del test a emplear ha de basarse en los siguientes criterios:

- Que el test sea válido y fiable, cuando se administra a una población similar al grupo de sujetos que queremos analizar.

- Que se conozcan adecuadamente todos los aspectos relacionados con la administración y la interpretación del test.

-Que se disponga de la infraestructura, medios técnicos y humanos suficientes para poder administrar el test en las condiciones requeridas.

- En el caso de requerir un control longitudinal, es necesario garantizar que se podrá administrar el test en condiciones similares, en las valoraciones posteriores.

Con excesiva frecuencia, se observa cómo entrenadores, u otros profesionales del deporte, que desean valorar longitudinalmente la evolución del rendimiento, cambian de test durante el período de control, con la finalidad de emplear otro supuestamente "mejor". Puede que los datos obtenidos con el segundo test sean más válidos y fiables, pero al cambiar de test se dificulta enormemente la interpretación prospectiva y retrospectiva. Así pues, es preferible utilizar pocos tests y no cambiar de procedimiento ni test si no es por razones científicamente probadas de mayor validez, fiabilidad, etc.

De este modo, el evaluador cada vez domina mejor los tests que emplea, va acumulando datos que le permiten comparar con facilidad a unos sujetos con otros y puede mantener un control longitudinal más fiable.

Una vez elegido el test, han de programarse las sesiones de medición en función del número de sujetos y de las características del test, atendiendo a las siguientes recomendaciones con carácter general:

- *Familiarizar adecuadamente a los sujetos con el test.* Dar las instrucciones pertinentes a los sujetos. Por ejemplo, la mayoría de los tests de condición física requieren que los sujetos estén descansados antes de pasar el test, que no hayan comido en las 3-4 horas anteriores, que no estén enfermos o convalecientes de una enfermedad, que no tengan dolor muscular tardío (“agujetas”), etc.

- *Comprobar que los instrumentos funcionan adecuadamente* y preparar la infraestructura necesaria. Preparar el sistema de registro de datos con suficiente antelación.

-*Comprobar que se reúnen los requisitos de seguridad* necesarios para poder administrar el test.

- *Verificar que los sujetos satisfacen las condiciones necesarias para que puedan realizar el test.*

- Si el test a administrar no establece un *calentamiento estandarizado*, se ha de decidir y programar el tipo de calentamiento a utilizar. Si el test no produce un grado importante de fatiga, es conveniente efectuar ensayos de prueba antes del test definitivo. Estos ensayos se introducen en la parte final del calentamiento y contribuyen a incrementar la fiabilidad del test.

- Durante la administración del test es conveniente tomar nota de todas aquellas *incidencias* que pudieran haber afectado a la administración del test y al rendimiento de los sujetos.

- *Planificar los ejercicios o actividades a desarrollar como enfriamiento a la finalización de las mediciones*.

- *Tratar de analizar los datos recogidos cuanto antes*. Un análisis temprano es ventajoso puesto que el evaluador aún recuerda la mayoría de las incidencias acontecidas durante la administración del test y posibilita una retroalimentación rápida, tanto al evaluador como a los sujetos testados.

1.4. Control de calidad.

Aunque la mayoría de los fabricantes aportan datos sobre la imprecisión de los instrumentos de medida, los ensayos realizados para obtener dicha imprecisión se han efectuado en condiciones óptimas que difícilmente pueden ser reproducidas en el laboratorio de Fisiología y aún menos en la pista. Por lo tanto, en el mejor de los casos, los datos de precisión aportados por el fabricante de instrumentos metrológicos, establecen la imprecisión mínima con que realizaremos posteriormente nuestras mediciones. Lamentablemente, sólo excepcionalmente los instrumentos son suministrados acompañados de un certificado de calibración externa, expedido por un laboratorio metrológico oficial, con **trazabilidad** reconocida. Esto último exige gran cautela al comparar mediciones realizadas en diferentes laboratorios de Fisiología y constituye una de las principales razones aducidas para recomendar la utilización siempre del mismo instrumento. Cuando no es posible emplear el mismo instrumento, éste deber ser reemplazado por otro similar de imprecisión conocida, calibrado mediante patrones certificados por un laboratorio de metrología oficial (es decir, con trazabilidad). De esta forma, los mismos patrones pueden ser medidos por ambos instrumentos. Posteriormente, pueden obtenerse ecuaciones y factores de corrección, que permiten transformar una medición obtenida mediante un instrumento en el valor "correcto" o en su defecto en el valor que se habría obtenido con el otro instrumento. Actualmente, el error de medida debido a la imprecisión instrumental es mucho menor que el achacable a la variabilidad biológica (Coggan y Costill 1984). Generalmente, el error tecnológico puede ser mantenido en niveles aceptables si se siguen adecuadamente en las recomendaciones del fabricante en el mantenimiento de los aparatos, en la calibración y en la técnica de medición.

Es conveniente disponer de un plan de control de calidad que incluya verificaciones periódicas de la fiabilidad y validez de los procedimientos utilizados en la evaluación de

la condición física, especialmente de aquéllos más susceptibles a la pérdida de fiabilidad y validez con el paso del tiempo. Por ejemplo, los analizadores de gases y los ergómetros pueden perder tanto validez como fiabilidad debido a múltiples factores: pérdida de estanqueidad de las conducciones, deterioro de los sensores de CO_2 y O_2, fatiga mecánica de los materiales, etc.

1.5. Síntesis de ideas fundamentales.

- La **medición** es el proceso por el cuál se recoge información cuantitativa o cualitativa, mientras que la **evaluación** consiste en utilización de mediciones para emitir un juicio de valor y adoptar decisiones.

- Los procedimientos que se emplean para efectuar las mediciones reciben el nombre de **tests** . Los tests deben ser válidos y fiables.

- La **validez** es una estimación del grado de veracidad del test, de tal manera que se considera que un test es válido cuando realmente mide lo que pretende medir. Para que un test sea válido es necesario que primero sea fiable, además tendrá que ser específico y exacto.

- La **fiabilidad** es una medida de la reproductibilidad o repetibilidad del test. Para que un test sea fiable es necesario que sea preciso, estable y objetivo.

- Un test es **objetivo** cuando existe una gran concordancia entre las medidas obtenidas por distintos evaluadores.

- Ningún test es fiable al 100 % o perfecto, ya que todo proceso de medición lleva asociado un **error**. La magnitud del error puede ser estimada por procedimientos estadísticos.

- La validez puede ser establecida por procedimientos lógicos (**validez de contenido**) y estadísticos (**validez estadística** o **validez correlacional**). Existen dos clases de validez estadística: la validez concurrente y la validez predictiva, ambos tipos de validez se determinan mediante el test de correlación de Pearson. La principal diferencia entre la validez concurrente y la validez predictiva reside en el momento en que se efectúan las mediciones. Así, las medidas se obtienen simultáneamente para efectuar una validación concurrente. En el caso de la validez predictiva, las medidas correspondientes al test objeto de análisis se obtienen antes que las medidas predichas.

- La fiabilidad de un test puede ser determinada estadísticamente. Entre los indicadores de fiabilidad más utilizados se encuentran los *coeficientes de fiabilidad* y los *coeficientes de variación*. Los coeficientes de fiabilidad se han clasificado en dos grandes categorías:

coeficientes interclase (basados en el cálculo del coeficiente de correlación de Pearson) y **coeficientes intraclase** que se obtienen por análisis de la varianza (ANOVA).

- Como principales ventajas, el coeficiente de correlación intraclase permite emplear en el cómputo del coeficiente de fiabilidad más de dos mediciones por persona, al tiempo de toma en consideración el valor medio de las medidas efectuadas, así como su dispersión (desviación estándar), por lo que proporciona una estimación más fidedigna de la fiabilidad que otros procedimientos.

- Podemos considerar como aceptables aquellos tests de condición física que presenten coeficientes de variación inferiores al 10%. Cuanto mayor sea el coeficiente de variación de un test de mayor tendrá que ser la magnitud de los cambios producidos por el entrenamiento (o cualquier otra intervención), para que puedan ser detectados por el test en cuestión.

- Las principales causas de imprecisión residen en las características de la magnitud analizada, en las limitaciones del instrumento de medida, en las correcciones aplicadas y en procedimiento de medida. Todas las magnitudes analizadas en seres vivos presentan **variabilidad biológica**.

- Lo ideal sería que la variabilidad, o error, achacable a todo el proceso de medición sea del orden de 10 veces inferior al error producido por la propia variabilidad biológica.

1.6. Aplicaciones prácticas.

- Cuanto más cercano a 1 (ó -1) se encuentre el valor del coeficiente de correlación de Pearson, mayor será la validez del test analizado. Cuando el coeficiente de correlación de Pearson es inferior a 0.6 podemos considerar que el grado de validez es insuficiente.

- Lo ideal sería que los tests fueran la vez reproducibles, exactos y precisos, no obstante, si son muy reproducibles es posible corregir los errores.

- Dado el gran número de factores que pueden influir en la reproductibilidad de un test, es conveniente determinarla experimentalmente en las condiciones habituales de trabajo.

- Para aumentar la fiabilidad de un test se pueden emplear mediciones múltiples, tomando como valor representativo de la medición efectuada el valor máximo, o el valor medio, de cada conjunto de medidas repetidas. La utilización del valor medio reduce el error debido a la falta de reproductibilidad, ya sea por errores de medición y/o por variabilidad biológica.

- La realización de múltiples mediciones puede provocar, por sí misma, una mejora en el rendimiento por un efecto aprendizaje y/o por efecto entrenamiento. El efecto aprendizaje se manifiesta, especialmente, en tests relativamente complejos como suelen ser los tests que pretenden medir las capacidades coordinativas o habilidades deportivas específicas. El efecto entrenamiento tiene lugar cuando la repetición del test mejora la capacidad de rendimiento del sujeto evaluado. Esta última situación puede darse cuando la cualidad valorada es muy "entrenable".

- Cuanto más fiable sea el test, más próximo a 1 resultará el valor del coeficiente de fiabilidad. La mayoría de las cualidades físicas pueden ser medidas con coeficientes de fiabilidad de 0.80 a 0.95.

- Los instrumentos de medida deben ser elegidos en función de la sensib ilidad y precisión requeridas. El rango de medición del instrumento debe abarcar, al menos, el rango de valores que puede adoptar la variable analizada.

- Los tests han de ser fáciles de administrar y de interpretar. Cuánto más fácil es la administración de un test menor es la probabilidad de cometer errores al administrarlo, lo que en definitiva contribuye a aumentar su fiabilidad y a reducir su coste económico. En cualquier caso, el grado de complejidad del test debe ser adecuado al nivel y capacidad de los deportistas a evaluar.

- Los criterios éticos que deben presidir la administración de un test destinado a la valoración de la condición física deben contemplar: una explicación clara y detallada de los propósitos del test, una exposición objetiva de los riesgos físicos y psíquicos que comporta el test y la garantía a la confidencialidad de los resultados. Además, nunca debe efectuarse un test cuando existen dudas en cuanto a la seguridad del test.

- Se conseguirán resultados más fiables si la administración del test sigue las siguientes reglas: familiarizar adecuadamente a los sujetos con el test, comprobar que los instrumentos funcionan correctamente, preparar la infraestructura necesaria con antelación suficiente, comprobar que se reúnen los requisitos de seguridad, verificar que los sujetos satisfacen las condiciones necesarias para que puedan realizar el test, tomar nota las incidencias durante la administración del test, estandarizar adecuadamente las actividades previas a la administración del test y analizar los datos recogidos cuanto antes.

- Es conveniente disponer de un plan de control de calidad que incluya verificaciones periódicas de la fiabilidad y validez de los procedimientos utilizados en la evaluación de la condición física, especialmente de aquéllos más susceptibles a la pérdida de fiabilidad y validez con el paso del tiempo.

Bibliografía

Baumgartner TA, Jackson AS (1987) Reliability and objectivity. En: Baumgartner TA, Jackson AS (eds) Measurement for Evaluation in Physical Education and Exercise Science. Wm C Brown Publishers, Dubuque, Iowa, pp 87-117

Calbet JAL, Dorado García C, Ferragut Fiol C, Chavarren Cabrero J (1998) Causas de error y variabilidad en la determinación del déficit máximo de oxígeno acumulado. Archivos de Medicina del Deporte 15:47-54
Campbell IT, Walker RF, Riad-Fahmy D, Wilson DW, Griffiths K (1982) Circadian rhythms of testosterone and cortisol in saliva: effects of activity-phase shifts and continuous daylight. Chronobiologia 9:389-396.
Coggan AR, Costill DL (1984) Biological and technological variability of three anaerobic ergometer tests. Int J Sports Med 5:142-145.
García de la Chica J (1991) Apuntes del Curso Superior de Calibración y Metrología. Junio de 1991. Asociación Española para la Calidad.
Groser M (1992) Entrenamiento de la Velocidad. Martínez Roca, Barcelona.
Kroll W (1962) A note on the coefficient of intraclass correlation as an estimate of reliability. Res Quart 38:412-419
Kuipers H, Verstappen FT, Keizer HA, Geurten P, van Kranenburg G (1985) Variability of aerobic performance in the laboratory and its physiologic correlates. Int J Sports Med 6:197-201.
Lopez Calbet JA (1993) Valoración Fisiológica de la Condición Física en Ciclistas Altamente entrenados. En: Departamento de Historia y Teoría de la Educación. Universidad de Barcelona, Barcelona.
Morrow JR, Jackson AW, Disch JG, Mood DP (1995) Norm-referenced measurement. En: Morrow JR, Jackson AW, Disch JG, Mood DP (eds) Measurement and evaluation in human performace. Human Kinetics, Champaign, Illinois, pp 77-101
Noakes TD (1988) Implications of exercise testing for prediction of athletic performance: a contemporary perspective. Med Sci Sports Exerc 20:319-330.

Riad-Fahmy D, Read GF, Walker RF, Griffiths K (1982) Steroids in saliva for assessing endocrine function. Endocr Rev 3:367-395.
Safrit M (1981) Evaluation in Physical Education. Prentice-Hall, Englewood Clifs, New Jersey.
Sharkey BJ (1990) Understanding muscular fitness. En: Physiology of fitness. Human Kinetics, Champaign, Illinois, pp 59-69. Physiology of fitness. Human Kinetics, Champaign, Illinois, pp 59-69
Thoden JS (1991) Testing aerobic power. En: MacDougall JD, Wenger HA, Green HJ (eds) Physiological testing of the high-performance athlete. Human Kinetics Books, Champaign, Illinois, pp 107-173
Touitou Y, Motohashi Y, Reinberg A, Touitou C, Bourdeleau P, Bogdan A, Auzeby A (1990) Effect of shift work on the night-time secretory patterns of melatonin,

prolactin, cortisol and testosterone. Eur J Appl Physiol Occup Physiol 60:288-292