XI CARACTERÍSTICAS PSICOMETRICAS DE LOS INSTRUMENTOS:CONFIABILIDAD

El investigador en ciencias sociales se enfrenta por lo general a problemas *sui generis* que conllevan, a la necesidad de construir instrumentos (cuestionarios, escalas, cédulas, etc.) especiales, ad-hoc para el problema de investigación. Por otro lado, éste, el problema de interés puede involucrar una gran cantidad de variables, por lo que es difícil que la decisión de llevar a cabo todos los pasos necesario para la construcción de un instrumento que se aboque a la medición de una sola variable (véase, por ejemplo, Nunnally, y Bernstein,1993).la alternativa más frecuente por la que opta el investigador es elaborar un instrumento por medio del cual pueda indicar o medir de la mejor manera posible y con un mínimo de tiempo, la mayor cantidad de información relevante de esas variables. En esta tarea uno de los problemas que se debe resolver es el de las características psicométricas del instrumento y que se refieren a dos aspectos fundamentales: la confiabilidad y la validez de los mismos y que frecuentemente son dejados de lado. En este capítulo se plantean los conceptos teóricos y los mecanismos prácticos para determinar la confiabilidad y la validez de los instrumentos, teniendo en cuenta los objetivos del investigador.

11.1. Confiabilidad

La confiabilidad de una prueba o instrumento se refiere a la consistencia de las calificaciones obtenidas por los mismos individuos en diferentes ocasiones o con diferentes conjuntos de reactivos equivalentes.

Es un hecho que en cualquier conjunto de medidas se encuentra que éstas varían, es decir que se producen errores de medición. La confiabilidad y la validez se ocupan, aún cuando de manera relativamente distinta, del problema del error de medición.

Como es sabido, existen dos posibles clases de error, el error constante, sistemático y el error casual, variable, debido al azar. Estos errores forman parte de cualquier medida obtenida (X):

$$x=v+e$$

Donde: v es igual a la medida verdadera más cualquier error sistemático y "e" viene a ser el error residual, casual o variable. Como puede verse el enunciado X expresa el supuesto básico psicológico referido a una sola medida y dado que la aplicación de instrumento se traduce en un conjunto de medidas, resulta más conveniente traducir el enunciado X a su forma más general:

$$\sigma x^2 = \sigma v^2 + \sigma e^2$$

Donde: σx^2 representa la varianza total; σv^2 se conoce como la varianza verdadera o sistemática porque engloba las medidas verdaderas más los errores constantes y σe^2 se refiere a la varianza debida a los errores variables.

En el sentido más amplio, la confiabilidad indica el grado con el que las diferencias individuales en las calificaciones de las pruebas se atribuyen a errores aleatorios de la medición, y el grado con el que se atribuyen a diferencias reales de las características o dominio en consideración. En términos técnicos, la confiabilidad señala qué proporción de la varianza total de las calificaciones de una prueba es "varianza de error". Al respecto, es importante entender a que se refiere esa varianza de error, ya que los actores que pueden ser considerados como varianza de error para un propósito, pueden clasificarse bajo el rubro de "varianza verdadera" para otro. Por ejemplo, si se está interesado en medir fluctuaciones de estado de ánimo, entonces los cambios cotidianos en las calificaciones de una prueba qué midiera alegría, depresión, serían relevantes para el propósito de la prueba y serían por lo tanto parte de la varianza verdadera de las calificaciones. Si, por otro lado, la prueba está diseñada para medir características de personalidad más duraderas o permanentes, las mismas fluctuaciones diarias caerían bajo el rubro de varianza de error.

Sin embargo , en esencia, puede decirse que cualquier condición que es irrelevante al propósito de la prueba representa varianza de error. Vista así, puede inferirse que cuando el examinador trata de mantener las condiciones de prueba uniformes, controlando el ambiente de la situación de prueba, (las instrucciones , los límites de tiempo, el "rapport" y otros factores similares), su propósito es reducir a la varianza de error y con esto incrementar la confiabilidad de la prueba. Sin embargo, pese a las óptimas condiciones de prueba logradas, difícilmente puede obtenerse un instrumento perfectamente confiable . De aquí la importancia de que cada prueba establezca claramente su índice de confiabilidad. Tal medida de confiabilidad tiene sentido cuando la prueba es administrada bajo condiciones estándar y aplicada a Sujetos semejantes a aquellos que constituyeron la muestra normativa. De aquí que las característica de tales muestras deberán estar perfectamente especificadas, junto con el tipo de confiabilidad que se utilizó.

Puede haber por supuesto, tantas variedades o tipos de confiabilidad como haya condiciones que afecten las calificaciones de las pruebas ya que cualquiera de tales condiciones pueden ser irrelevantes para ciertos propósitos y muy relevantes para otros. No obstante la gran diversidad de tipos de confiabilidad en la

práctica se utilizan unos cuantos. Los principales son los siguientes:

11.1.1. Estabilidad Temporal

Una fuente de varianza de error obvia para la mayoría de los propósitos de la aplicación de pruebas, es la de las fluctuaciones azarosas de la ejecución, que ocurren de una sesión de prueba a otra. Estas variaciones pueden deberse en parte a la falta de control de las condiciones de prueba; a los cambios en la condición del Sujeto mismo, (enfermedad, fatiga, tensión emocional, preocupación, experiencias recientes de naturaleza agradable o desagradable, etc.). La estabilidad temporal de una prueba, depende en parte, del lapso mayor o menor que interviene entre la primera y segunda medición.

Las fluctuaciones azarosas, a corto plazo, que ocurren durante intervalos que van de unas cuantas horas a algunos meses por lo general se incluyen dentro de la varianza de error de la calificación de la prueba. Con este tipo de confiabilidad, se aconseja que sea corto dicho intervalo, más aún si se trata de niños pequeños, ya que a edades tempranas se producen cambios de desarrollo más rápidos que en los adultos. Los estudiosos del tema consideran que el intervalo de tiempo entre la aplicación de las pruebas no deberá exceder a los seis meses.

Con períodos de tiempo mayores cualquier cambio adicional en la ejecución relativa de una prueba es probable que sea acumulativo y progresivo y no tanto debido al azar. En esas condiciones los cambios encontrados caracterizan un área de conducta más amplia que la que cubre la ejecución de la prueba. Así por ejemplo, el nivel general de un individuo en aptitud verbal, compresión mecánica, o juicio artístico, puede verse apreciablemente alterado en largos períodos de tiempo debido al cúmulo de experiencias ocurridas, comunes o poco comunes durante ese tiempo. El status de un Sujeto puede haber cambiado en forma apreciable en relación a otros de su misma edad, pueden producirse cambios en el hogar, en el trabajo o en la escuela; cambios en el organismo (maduración fisiológica y psicológica, etc.)

El grado en el que tales factores pueden afectar el desarrollo psicológico del individuo plantea un importante problema para la investigación; esto sin embargo, no debe confundirse con la estabilidad de una prueba particular. Así, si se mide, por ejemplo la confiabilidad de una prueba de inteligencia, de personalidad, por lo general la prueba de la estabilidad se hace habiendo transcurrido tan sólo unas semanas. Se han hecho estudios en los que se han replicado las pruebas con intervalos de tiempo grandes, pero los resultados por lo general se discuten, o se habla de ellos, en términos de “constancia del nivel intelectual”,

predictibilidad de la inteligencia del adulto a partir de la ejecución infantil, más que en términos de la confiabilidad de una prueba determinada. El concepto de confiabilidad por lo general se restringe a cambios azarosos a corto plazo, cambios que caracterizan el comportamiento de la prueba en sí misma.

11.1.2. Muestreo de reactivos

Con toda seguridad, todos hemos pasado por la experiencia de tomar un examen en alguna materia y sentido que nos "iluminaba" la suerte porque muchos de los reactivos cubrían los temas que habíamos estudiado con más cuidado. En alguna otra ocasión, nos pudo haber sucedido lo contrario, es decir, nos encontramos con una gran cantidad de preguntas acerca de los temas que no habíamos revisado sino tan solo por "encima". Esta situación ilustra una segunda fuente de varianza de error en las calificaciones de las pruebas. ¿Hasta qué grado dependen las calificaciones de esta prueba, de los factores específicos de la selección particular de reactivos? Si un investigador diferente, trabajando en forma independiente prepara otra prueba de acuerdo con las mismas especificaciones, ¿qué tanto diferiría la calificación de un individuo en ambas pruebas?

Supóngase que se construye una prueba de vocabulario de 40 reactivos para obtener una medida general de comprensión verbal. Después se construye una segunda prueba con el mismo número de reactivos, solo que con palabras diferentes. Ambas tienen el mismo propósito, se ha tenido cuidado que en ambas, los reactivos cubran el mismo rango de dificultad.

Las diferencias en las calificaciones obtenidas por los mismos individuos en estas dos pruebas ilustran el tipo de confiabilidad que se está considerando. Debido a factores fortuitos en la experiencia pasada de diferentes individuos, la dificultad relativa de las dos pruebas variará algo de persona a persona. En esta forma, la primera prueba puede tener un mayor número de palabras desconocidas para el sujeto A que la segunda. Por otro lado, ésta puede contener demasiadas palabras desconocidas para el sujeto B. Si los dos individuos son aproximadamente iguales en sus conocimiento total de palabras (v.gr., en sus "calificaciones verdaderas"); de todas maneras B excederá a A en la primera prueba mientras que A excederá a B en la segunda. La localización relativa de estos dos sujetos en las dos pruebas será reversible, debido a la diferencia azarosa en la selección de los reactivos.

11.1.3 Homogeneidad de los reactivos

La homogeneidad de una prueba se refiere esencialmente a **la consistencia de la ejecución de todos los reactivos dentro de una**

prueba. Por ejemplo, si una prueba tiene solo reactivos de multiplicación, mientras que otra comprende reactivos de sumas, restas, multiplicaciones y divisiones, la primera probablemente tendrá más consistencia entre su reactivos que la segunda. En la segunda, la prueba más heterogénea, un sujeto puede responder mejor en la parte de las restas que en cualquiera de las otras operaciones aritméticas; otro sujeto puede responder relativamente bien en la parte de las multiplicaciones que en cualquiera de las otras operaciones aritméticas; otro, puede responder relativamente bien en la parte de la divisiones, pero en forma más pobre en sumas, restas y multiplicaciones; y así sucesivamente. Un ejemplo más extremo: una prueba con 40 reactivos de vocabulario, en contraste con otra que tiene 10 reactivos de vocabulario; 10 de relaciones espaciales, 10 de razonamiento aritmético y 10 de discriminación perceptiva. Es probable que en la segunda se encuentre poca o ninguna relación, en la ejecución del Sujeto, entre los diferentes tipos de reactivos.

Como es de esperarse las calificaciones de una prueba serán menos ambiguas cuando éstas se derivan de pruebas relativamente homogéneas. Supóngase que en la prueba de los 40 reactivos altamente heterogéneos, antes citada, los Sujetos A y B obtuvieron ambos calificación de 20. ¿Se puede concluir que las ejecuciones de estos dos Sujetos en dicha prueba fueron iguales?. Por supuesto que no. El Sujeto A pudo haber completado en forma correcta 10 reactivos de vocabulario, y los 10 reactivos de discriminación perceptual y ninguno de los reactivos de razonamiento aritmético y de relaciones espaciales. En contraste, el sujeto B pudo haber obtenido una calificación de 20 respondiendo correctamente a cinco reactivos de cada uno de los cuatro tipos diferentes de reactivos. Se podrían producir muchas otras combinaciones que arrojarían una calificación total de 20. Pero la calificación tendría un significado muy diferente al haberse obtenido de combinaciones de reactivos tan disímiles. Por otro lado, en la prueba relativamente homogénea de vocabulario, una calificación de 20 probablemente significa que el Sujeto contestó correctamente aproximadamente las primeras 20 palabras, si los reactivos estaban ordenados por grado creciente de dificultad. Pudo haber fallado en dos o tres de las palabras más fáciles y respondido correctamente a dos o tres de las más difíciles, pero tales variaciones individuales, son ligeras en comparación con aquellas encontradas en la prueba más heterogénea.

Un aspecto relevante en relación a esto se refiere al grado de relativa homogeneidad o heterogeneidad de la misma variable que la prueba trata de medir. Aunque las pruebas homogéneas son preferidas debido a que sus calificaciones permiten una interpretación relativamente poca ambigua, una única prueba homogénea, obviamente, no es un predictor adecuado de un criterio altamente heterogéneo.

En la predicción de un criterio heterogéneo, la heterogeneidad de los reactivos de una prueba no necesariamente representa la de error. Las pruebas de inteligencia tradicionales proporcionan un buen ejemplo de pruebas heterogéneas; quizá sea más deseable construir varias pruebas relativamente homogéneas, cada una midiendo una fase diferente del criterio heterogéneo. En esta forma se podría combinar una interpretación de calificaciones sin ambigüedades y un cubrimiento adecuado del criterio.

¿En qué forma difiere la homogeneidad de la adecuacidad del muestreo de reactivos? Un ejemplo extremo servirá para resaltar la diferencia; supóngase que cada uno de los reactivos de cierta prueba, mide una función diferente y no relacionada. Sería totalmente posible construir otra prueba paralela a la primera, que contuviera el mismo tipo y distribución de reactivos. Teóricamente las calificaciones de estas dos formas podrían estar muy de acuerdo, indicándose en esta forma una alta confiabilidad de la prueba en términos de muestreo de reactivos. Sin embargo, la homogeneidad de esta prueba sería cercana a cero, ya que la consistencia de la ejecución de un reactivo a otro dentro de cualquiera de las formas no sería mejor que la dada por el azar.

11.1.4 Confiabilidad del examinador y el calificador

Deberá ser claro ahora que los diferentes conceptos de confiabilidad de una prueba dependen de los factores que se coloquen bajo el término "varianza de error". En un caso, la varianza del error abarca fluctuaciones temporales, en otro, se refiere a las diferencias entre conjuntos de reactivos paralelos; y en otro más incluye cualquier inconsistencia entre los reactivos. Por otro lado, los factores excluidos de las medidas de la varianza de error son de dos tipos: a) aquellos factores cuya varianza debiera permanecer en las calificaciones, ya que son parte de las diferencias reales o verdaderas bajo consideración; y b) aquellos factores irrelevantes que pueden ser controlados experimentalmente. Por ejemplo, no es común reportar los errores de medición que resultan cuando una prueba es administrada en condiciones distractoras o con tiempos límite menores o mayores que los especificados en el manual. Sin embargo, como los errores de tiempo y de distracción serios pueden eliminarse en forma empírica de la situación de prueba, no es necesario reportar coeficientes de confiabilidad especiales correspondientes a la "varianza de tiempo", o "varianza de distracción".

En forma semejante, la mayoría de las pruebas proporciona procedimientos altamente estandarizados para la administración y calificación por lo que se puede suponer que la "confiabilidad del examinador" y la "confiabilidad del calificador" son lo

suficientemente altos para propósitos prácticos. En esta forma, no existe ninguna necesidad especial de medir estos tipos de confiabilidad. Esto es particularmente cierto para pruebas de grupo, diseñadas para ser aplicadas a grandes grupos de sujetos y para ser calificadas por máquinas. En tales pruebas sólo se debe asegurar que se **sigan** en forma cuidadosa los procedimientos prescritos. El problema es por tanto de **control empírico de las condiciones.**

Sin embargo, en ciertas pruebas individuales, el papel del examinador es mucho más complejo. Como ejemplo se puede citar la prueba de Stanford Binet, y la mayoría de las pruebas preescolares. El procedimiento de prueba en tales casos no está tan rígidamente estandarizado. Mucho depende del éxito que el examinador tenga en el establecimiento de **rapport** y en la producción de la motivación adecuada. Con frecuencia la ejecución del Sujeto necesita ser evaluada por el examinador durante el mismo proceso de administración de la prueba, ya que tal ejecución determina en qué forma procederá el durante la prueba. Bajo tales condiciones es probable que aún examinadores muy calificados obtengan a veces resultados diferentes de los mismos sujetos. Estas variaciones en la calificación constituiría la varianza de error atribuible a idiosincrasia o diferencias individuales entre los examinadores.

En pruebas en las que la idiosincrasia del examinador puede jugar una parte apreciable, es deseable obtener alguna medida de la "confiabilidad del examinador", especialmente cuando han de combinarse los resultados obtenidos por varios examinadores. Para este tipo de pruebas se deberá dar igual importancia al índice de confiabilidad de examinadores, como se le da a otros tipos de índices de confiabilidad

11.1.5. Fuentes principales de error

Ya se ha señalado que los instrumentos de medición son confiables en la medida en que son repetibles, y que cualquier influencia azarosa que tienda hacer que las medidas sean diferentes de una ocasión a la siguiente es una fuente de error. En la práctica hay muchos factores que hacen que los instrumentos de medición no sean exactamente repetibles; el número y tipo de factores depende de la naturaleza de la prueba y de cómo se utilice ésta. Ahondando en algunas de las fuentes principales de error en la medición, se pasa a exponer algunos ejemplos.

11.1.5.1 Variación dentro de una prueba

Es importante hacer una distinción entre los errores de medición que producen variación en la ejecución de un reactivo al siguiente, dentro de una prueba y los errores que se manifiesten sólo en la

variación de la ejecución en diferentas formas de una prueba aplicada en diferentes tiempos u ocasiones.

La principal fuente de error de una prueba es debido al muestreo de reactivos. De acuerdo con el modelo domino-muestra, cada persona tiene una probabilidad particular de responder en forma correcta a cada reactivo, que depende de su calificación verdadera y de la dificultad del reactivo para la gente en general. En el caso más simple, si una persona tiene una calificación verdadera promedio y todos los reactivos tienen un índice de dificultad del 0.5 para la gente en general, esa persona tiene una probabilidad de 0.5 de responder correctamente cualquier reactivo seleccionado al azar dentro del dominio. Se esperaría que respondiera, en forma correcta la mitad de los reactivos en cualquier prueba que se sacara del dominio, pero esta expectativa estaría acompañada de algún error. Entre más reactivos tuviera la prueba, menor sería el error, la misma lógica se puede extender a los reactivos que no tienen una respuesta "correcta" (respuestas de sentimiento). Por ejemplo, en reactivos que se refieran a estar o no de acuerdo con ciertas afirmaciones. Se puede pensar que cada persona tiene una probabilidad establecida de estar de acuerdo con cada afirmación, lo que a su vez llevaría a un número esperado de respuestas de acuerdo dentro de una muestra de reactivos. Dependiendo del número de reactivos en cada muestra, habría alguna variabilidad en las calificaciones de una prueba a otra prueba.

El error debido al muestreo de reactivos es totalmente predictible a partir de la correlación promedio. En consecuencia, el **coeficiente alfa** sería la medida correcta de la confiabilidad para cualquier tipo de reactivo, y la versión especial de esa fórmula, la (KR-20, Kudder - Richardson 20), para pruebas de reactivos dicotómicos (Nunnally y Bernstein, 1993).

En las pruebas de elección múltiple, la adivinación es una fuente de error de la medición. Si por ejemplo, un individuo realmente no sabe la respuesta a dos preguntas, puede contestar correctamente una y no la otra debido a que adivinó. El adivinar produce alguna variación en la ejecución de un reactivo al siguiente, y esto tiende a disminuir la confiabilidad de la prueba. El adivinar es manejado con facilidad por el modelo **dominio-muestra**. Puede pensarse que el dominio está constituido por reactivos de opción múltiple. La correlación típica entre tales reactivos permitiría una estimación de la confiabilidad de cualquier muestra de reactivos. El adivinar podría servir para disminuir la correlación típica; pero una vez que ésta fuera estimada de las correlaciones dentro de una prueba, se podría usar para estimar la confiabilidad.

Además de la adivinación, muchos otros factores producen variación en las calificaciones de un reactivo al siguiente dentro de una

prueba. Por ejemplo, a un Sujeto, le puede empezar a doler la cabeza cuando está resolviendo una prueba; esto tenderá a disminuir sus calificaciones en los reactivos que contestó cuando apareció su malestar; otra persona puede tener la intención de marcar la alternativa **a** para un reactivo en particular, y debido a un error marcar en cambio la alternativa **b**: otra puede, inadvertidamente, alterar un reactivo que podría haber contestado en forma correcta; a la mitad de una prueba, una persona puede darse cuenta que mal interpretó las instrucciones de la forma en que se deberá responder y por no tener tiempo de regresar a los reactivos anteriores, resulta que su desempeño fue mejor en los reactivos sucesivos que en los primeros; también puede suceder que una persona que realmente sabe la respuesta a una pregunta, puede responder en forma incorrecta porque accidentalmente leyó "no es un ejemplo de" en lugar de " es un ejemplo de". El número de ejemplos para señalar factores que producen errores dentro de las pruebas, podría formar enormes listas, pero lo que interesa es dejar claro las causas de variación en las pruebas.

Hasta cierto punto, pueden estimarse los errores de calificación para una prueba. En las pruebas objetivas, los errores de calificación son puramente mecánicos, pero como tienden a disminuir las correlaciones entre los reactivos, caen dentro del campo del modelo dominio-muestra. En algunas pruebas la calificación es principalmente subjetiva, como por ejemplo, en los exámenes de tema o ensayo y en la mayoría de las pruebas proyectivas. El elemento de error de medición en este caso está causado por las fluctuaciones en la norma de calificación de un calificador individual, y por las diferencias en las normas de diferentes calificadores. para el calificador individual, tales errores e manifiestan dentro de una prueba si cada reactivo es calificado Independientemente de los otros reactivos. Por ejemplo, en un examen de temas el instructor puede calificar todas las respuestas a la pregunta número uno; después calificar todas las respuestas a la pregunta dos, y así sucesivamente. Si tales calificaciones son independientes, la corrección promedio entre los reactivos puede usarse para obtener una estimación exacta de la confiabilidad.

Todos los errores que ocurren dentro de una prueba pueden ser fácilmente abarcados por el modelo dominio muestra. Las suposiciones del modelo pueden extenderse al caso donde las influencias situacionales son "asignadas" azarosamente a los reactivos. En esta forma, no solamente a cada persona se le administraría una muestra aleatoria de los reactivos del dominio si no que cada reactivo estaría acompañado por un conjunto azaroso de factores situacionales. Así, el que una persona pase o no cualquier reactivo obtenido al azar del dominio es función en parte, de la frecuencia con que un reactivo sea escogido y en parte, de la frecuencia de los factores situacionales que acompañan al reactivo.

Todas estas fuentes de error tenderán a disminuir la correlación promedio entre los reactivos de una prueba, pero la correlación promedio es todo lo que se necesita para estimar la confiabilidad.

11.1.5.2. Variaciones entre pruebas

Si se administran formas alternativas, equivalentes o paralelas de una prueba con un intervalo de dos semanas de tiempo transcurrido entre ellas, casi nunca correlacionarán en forma perfecta los dos conjuntos de calificaciones. El modelo dominio-muestra proporciona una predicción de la correlación, y como se dijo anteriormente, la predicción toma en cuenta no sólo el muestreo del contenido, sino también muchas fuentes de error dentro de cada sesión de prueba. Hay sin embargo, tres fuentes de error principales que intervienen entre la administración de diferentes pruebas que no son precisamente estimadas a partir de la correlación promedio de los reactivos dentro de cada prueba. El modelo dominio-muestra es un muestreo real de los reactivos de un dominio hipotético. Dos pruebas de ortografía construidas independientemente por dos personas pueden enfatizar diferentes tipos de palabras. Entonces, la correlación entre las dos pruebas puede ser menor que la predicha a partir de la correlación promedio entre los reactivos de cada prueba. En forma semejante, forma alternativas de un instrumento que mide actitudes hacia las Naciones Unidas, pueden ser sistemáticamente diferentes en contenido, y en consecuencia la correlación entre las dos formas sería menor que la predicha por el modelo dominio-muestra.

Un segundo factor que produce variación en las calificaciones en algunas pruebas de una ocasión a la siguiente, es debido a la subjetividad de la calificación. En un examen por temas o en una prueba proyectiva, el mismo examinador puede dar clasificaciones algo diferentes a las mismas personas, y aún diferencias mayores deberán esperarse entre las calificaciones dadas por diferentes examinadores. Previamente se dijo que parte del error debido a la subjetividad de la calificación de una persona podría estimarse a partir de la correlación entre los reactivos dentro de una prueba, si los reactivos se calificaran independientemente, pero esto tocará a tan solo una parte del error. El calificador puede cambiar sus normas en alguna forma de una ocasión a la siguiente. por ejemplo, entre las dos administraciones de la prueba, el examinador puede considerar un tipo particular de respuesta como más patológico de lo que previamente había considerado. Antes se dijo que si diferentes partes de una prueba son calificadas en forma independiente por diferentes examinadores, la correlación promedio entre los reactivos sería indicativa del error involucrado al usar diferentes examinadores; pero como rara vez colaboran dos examinadores en esta forma, difícilmente se conoce la cantidad de error que existe en un examinador.

Otra fuente de variación en la ejecución de una prueba de una ocasión a la siguiente es debida al hecho de que la gente realmente cambia al respecto del atributo que se está midiendo. Una persona puede sentirse mucho mejor de una ocasión a la siguiente, puede estudiar el contenido del domino, o puede cambiar su actitud hacia las Naciones unidas. Es razonable pensar que exista cierta fluctuación en las habilidades de un día a otro, dependiendo de factores fisiológicos y ambientales. Lo mismo sucede con los estados de ánimo, autoestima, y actitudes hacia la gente y cosas. Tales cambios en la gente tenderán a hacer que las correlaciones entre las formas alternativas de las pruebas sean menores que las predichas por la correlación promedio de los reactivos de cada prueba.

11.2 Métodos Experimentales para obtener la confiabilidad de una prueba

La confiabilidad se puede definir como la "correlación entre pruebas paralelas". La definición de pruebas paralelas se expresa en términos de igualdad de medias, desviaciones estandar e intercorrelaciones.

El término confiabilidad fue introducido por Spearman en sus trabajos básicos de teoría de las pruebas (Spearman, 1904,1907,1910 y 1913). Desde entonces ha habido muchas discusiones de los varios factores que influyen sobre la confiabilidad en relación a los diferentes métodos de medida. Existen muchas formas diferentes de clasificar los factores que influyen sobre la confiabilidad y sobre los métodos para medirla. Entre ellas se cuentan los siguientes métodos:

a) el uso de pruebas paralelas
b) "retest" con la misma forma de la prueba
c) varios métodos de mitades, tales como la primera contra la segunda mitad; reactivos pares contra nones, y el método de subpruebas al azar apareadas (ya sea mitades o tercios).

Recientemente se han elaborado métodos para estimar la confiabilidad de una prueba de homogeneidad que no hacen uso de la correlación de calificaciones paralelas. En lugar de eso, estos métodos usan datos de análisis de reactivos para estimar la homogeneidad del grupo de reactivos de una prueba.

Aunque el error de medición es un concepto más básico en la teoría de las pruebas que el coeficientes de confiabilidad, se ha vuelto costumbre durante los últimos 50 años evaluar a las pruebas en términos del coeficiente de confiabilidad más que en términos del error de medición. Como existen ventajas y desventajas para cada una de estas medidas, se sugiere que se utilicen ambas para la

evaluación completa de cualquier prueba. Otis y Knollin (1921) señalaron que el error de medición es superior al coeficiente de confiabilida ya que no varía con cambios en la heterogeneidad del grupo. Kelley (1921) indicó que, aunque el error de medición no varía con la heterogeneidad del grupo, sin embargo, la unidad en la que se expresa el error de medición si varía de una prueba a la otra. Lincoln (1932) y (1933) señaló que la confiabilidad podría ser muy alta aun cuando las diferencias entre dos conjuntos de medidas fueran muy grandes.

Las pruebas o subpruebas que se correlacionen para determinar la confiabilidad de una prueba, deberán ser paralelas tanto en el sentido de que satisfagan los criterios estadísticos de pruebas paralelas (Gulliksen, 1950, capítulo 14) como en el sentido de que los reactivos requieren los mismos procesos psicológicos y el mismo tipo de aprendizaje por parte de los Sujetos. Este último criterio depende del juicio del técnico en pruebas y el experto en la materia, y será diferente para cada tipo de prueba. Aquí, se considera sólo los métodos generales del establecimiento de pruebas o subpruebas paralelas que son comunes a todos los tipos de material.

11.2.1. Uso de las formas paralelas

Para la mayoría de las situaciones, se ha encontrado que el mejor método para obtener la confiabilidad de una prueba es construir formas paralelas de la prueba y administrarlas en diferentes ocasiones al mismo grupo de sujetos. Así el método comúnmente usado será construir dos formas paralelas para este propósito; pero se sabe que con tres formas paralelas es posible hacer una evaluación más completa y además permite asegurar que las formas sean paralelas, no sólo con respecto a sus medias y varianza, sino también en relación con sus correlaciones.

Existe sólo una situación en la que el uso de formas paralelas administradas en diferentes días no es aconsejable. Esto es, cuando la habilidad que está siendo probada cambia marcadamente en el intervalo de tiempo transcurrido entre las pruebas. Por ejemplo, si se quiere determinar la confiabilidad de una prueba de mecanografía administrando una forma a un grupo el lunes y otra forma el viernes, el método no funcionaría si el grupo estuviera practicando (y por lo tanto aumentando rápidamente su habilidad mecanográfica)durante el intervalo del tiempo transcurrido.

En la misma forma, este método no es adecuado si la primera prueba se da cuando los Sujetos están en excelente “condición” y la segunda se aplica cuando la habilidad de los Sujetos ha disminuido, por falta de práctica durante la semana transcurrida entre ambas aplicaciones.

El mismo tipo de consideración se aplica por ejemplo, a cualquier prueba de destreza física o habilidad o dominio muscular. Las dos administraciones de la prueba no pueden usarse para estimar la confiabilidad de la prueba si existe una buena razón para creer que los sujetos han mejorado o desmejorado en la variable que está siendo medida.

Para la mayoría de las pruebas de logro escolar y habilidad mental, es razonablemente fácil estar seguros de que los sujetos no han cambiado realmente en forma marcada durante el período que interviene entre las dos pruebas (Gulliksen 1950, pag. 195). Para otros tipos de ejecución, de los cuales las habilidades atléticas de varios tipos son un buen ejemplo, es muy difícil mantener a un grupo en un estado de excelencia uniforme. Es probable que la habilidad se deteriore con la falta de práctica, y/o pueda mejorar o pueda estancarse con la misma. En tales casos todo el "error de medición" no puede ser atribuido a la prueba. Mucho de lo que se manifiesta en la prueba estadística como error de medición es de hecho una variabilidad real de la habilidad. Sin embargo, desde otro punto de vista se debe reconocer que la medición de algunas habilidades es en extremo poco confiable (independiente de la causa de esta falta de confiabilidad); en ese caso, al utilizar cualquier de tales mediciones, se deben tratar, como se tratarían las mediciones muy poco confiables.

Sin embargo, si se está manejando un periodo de tiempo durante el cual la habilidad medida o la variable no cambiará en forma sistemática para los diferentes miembros del grupo, y se está trabajando con un grupo de Sujetos bajo condiciones tales que no es probable que la habilidad o variable cambie, la utilización de las diferentes formas de la prueba es el método más adecuado para indicar la confiabilidad.

Deberá de señalarse que las posibilidades de error anotadas arriba pueden detectarse con facilidad. Si el grupo se ha mejorado o se ha deteriorado, la media será más alta o baja en la segunda ocasión. Si alguna persona han mejorado y otras han desmejorado, la desviación estándard con toda probabilidad cambiará. Un conjunto complicado de influencias en el que algunas personas mejoran y otras se deterioran en tal forma que la media y la desviación estándar del grupo permanezcan iguales, es una posibilidad que puede existir, pero sin duda alguna sería muy extraña o rara.

En resumen, el método de aplicar pruebas paralelas con un intervalo de tiempo entre ellas, es un método que permite que las fuentes de error relevantes influyan sobre el coeficiente de confiabilidad. Si se utilizan las pruebas estadísticas de medias y desviaciones estandar, y si se satisfacen, el método es uno que

puede usarse en forma rutinaria con relativamente poco temor de que factores irrelevantes y no detectados estén produciendo que el coeficiente de confiabilidad obtenido sea uno espuriamente alto o espuriamente bajo.

Se debe notar, que el método de las formas paralelas es válido para las pruebas de velocidad. Una prueba de velocidad es una prueba compuesta de reactivos, muy fáciles. Reactivos tan fáciles que podrían ser contestados todos por todo el grupo si se les permitiera tiempo para hacerlo. Por ejemplo, un conjunto de suma de dos dígitos dados a alumnos de segundo de secundaria se aproximaría a ser una prueba de velocidad. Si se va a obtener un buen rango de calificaciones en tal prueba, es necesario que se tengan un gran número de reactivos, y establecer un tiempo límite tan corto que únicamente los mejores del grupo terminen, si es que lo hacen. En tal prueba, es importante el efecto de la práctica de una vez a la siguiente. A menos de que condiciones tales como cantidad de práctica y el uso de "ejercicio previo" estuvieran cuidadosamente estandarizados, no sería posible que se tuviera la misma media y varianza en las formas paralelas para el grupo. Sin embargo, si las medias y las varianzas son iguales,uno puede estar razonablemente en lo cierto al decir que la intercorrelación de las dos formas es una aproximación razonable del coeficiente de confiabilidad que debería tener la prueba.

La confiabilidad de formas paralelas también se puede obtener administrando ambas formas en la misma sesión. Una vez más, en algunas pruebas, puede haber una marcada diferencia en la ejecución debido al hecho de que la aplicación de la primera prueba influyó a la segunda, por ejemplo, si es una prueba de velocidad de sumas de dos dígitos, es probable que para muchas personas, particularmente las peores, la calificación en la segunda prueba sea mucho mejor debido a la práctica obtenida en la primera prueba. Por supuesto que esto puede detectarse fácilmente en los resultados ya que la media sería mayor para la segunda forma. Existen otras pruebas para las cuales la ejecución en la segunda forma es muy probable que sea peor que la ejecución en la primera. Cualquier prueba que produzca fatiga o cansancio en los sujetos claramente entra dentro de esta categoría, y una vez más, tal fatiga podría ser detectada con facilidad a partir de los resultados. El promedio sería menor para la segunda prueba que para la primera.

Si las anteriores, más bien obvias y fácilmente detectables, dificultades no estuvieran presentes, la dificultad principal con la confiabilidad obtenida mediante la administración sucesiva de formas paralelas, es que es muy alta. Esto se debe al hecho de que no hay posibilidad de que la variación debida a la variabilidad normal diaria disminuya la correlación entre las formas paralelas. Woodrow (1932) en su estudio de la variabilidad cotidiana demostró

que existen variaciones de un día a otro en la ejecución de las pruebas.

Algunos otros autores han señalado que algunas veces una baja correlación entre dos formas paralelas de una prueba indica que la prueba es una medida inestable de un rasgo estable; en otras ocasiones tal correlación baja, puede surgir de una medición estable de un rasgo inestable. La inestabilidad, ya sea en la prueba o en el rasgo, resultaría en una baja correlación entre las formas paralelas. Métodos para determinar la inestabilidad de un rasgo como algo diferente de la inestabilidad de un instrumento han sido sugeridas por Paulsen (1931), Thouless (1936) y (1939), Preston (1940). Se puede entonces concluir que si las formas paralelas de una prueba son aplicadas en el mismos día y que si se satisface el criterio estadístico de las pruebas paralelas, específicamente el de igualdad de medias y desviaciones estándard, la confiabilidad obtenida es probablemente más alta que la que hubiera sido obtenida si se hubiera permitido que la variabilidad de un día a otro hubiera afectado a la confiabilidad.

Hablando en términos generales, entonces, el uso de dos o tres formas paralelas administradas en diferentes ocasiones es el mejor método para determinar la confiabilidad de una prueba. Sin embargo, como con frecuencia no se dispone de varias formas paralelas, y como también es difícil algunas veces asegurar la cooperación de los Sujetos durante períodos extendidos de tiempo, se considerará la posibilidad de obtener una indicación de la confiabilidad cuando sólo se dispone de una forma de la prueba.

11.2.2 "Retest"(Replicación) con la misma forma

Algunas veces, cuando las formas paralelas de una prueba no están disponibles, es posible obtener una estimación de la confiabilidad administrando la misma prueba dos veces. Por lo general, es preferible hacer esto con un intervalo de tiempo transcurrido "regularmente grande" entre ambas aplicaciones. Una vez más con este método se debe de estar pendiente al respecto del efecto de la práctica o fatiga que será detectado con facilidad en la mayoría de los casos, observando las distribuciones de las calificaciones de la prueba en la primera y segunda aplicación (administración). A parte de tal efecto, el peligro principal en esta técnica es que la confiabilidad será muy alta debido a que existe la tendencia por parte del Sujeto de repetir su ejecución previa. Esto es, si el Sujeto no conoce la respuesta a un reactivo, pero con suerte lo adivina y lo resuelve bien, es más probable que lo vuelva a adivinar, la siguiente vez y se asegure el crédito para un reactivo para el que en realidad no conoce la respuesta. En la misma forma, si comete alguna pequeña equivocación, y como resultado responde en forma incorrecta a un reactivo que en situaciones normales hubiera

contestado en forma correcta, es más probable que repita su ejecución cuando se le vuelva a administrar la prueba. Tal efecto no ocurrirá si la persona se estuviera sometiendo a una forma paralela que no contiene los mismos reactivos. En otras palabras, la ejecución en la repetición de una prueba tiende a parecerse más a la calificación original que la ejecución en una forma paralela misma prueba. Por tal razón, es aconsejable que este método de repetición de la misma prueba en diferentes ocasiones no se utilice, ya que producirá un coeficiente espuriamente alto, y no es fácil determinar el grado de error.

Puede haber excepciónes con ciertas variables, por ejemplo con discriminación perceptual, para la cual no se pueden construir formas paralelas. Por ejemplo, una prueba de discriminación de tonos o una prueba de umbrales de audición para diferentes tonos puros, podría ser replicada sin que tal efecto ocurriera. La persona simplemente juzga cada vez si oye un tono o si no oye un tono. En una prueba como ésta parece no haber una forma o manera sencilla en la que la persona pudiera repetir en forma espuria sus errores y éxitos del conjunto previo de ensayos. Sin embargo, aún en temas tan simples, con frecuencia es deseable construir varias técnicas de medición diferentes y correlacionarlas, así como obtener la confiabilidad de una prueba repetida por el uso de cada método. En general, se puede decir que aún cuando parezca que la repetición de la misma forma es todo lo que se puede hacer, está bien que el constructor de la prueba use algo de su ingenio para presentar un factor dado en diferentes formas que sean comparables **grosso modo**, y que después vea que tanto acuerdo existe, entre las diferentes pruebas. Con frecuencia se descubren cosas nuevas de la variable estudiada al ser medida en esta forma.

11.2.3 Consideraciones generales de los Métodos pos mitades

Por lo general cuando sólo se dispone de una sola forma de la prueba, la confiabilidad se determina por algún método de mitades. Esto quiere decir que los reactivos de la única forma se dividen para formar dos, cada una con la mitad de número de reactivos de la forma original. Típicamente, los sujetos no saben que la prueba va a ser calificado en dos parte, y no saben qué reactivos estarán en cuál de las mitades. El experimentador no necesita decidir y por lo general no lo hace, cómo van a dividirse los reactivos hasta que ve los resultados de la prueba. Sin embargo, desde el punto de vista de establecer procedimientos de calificación eficientes, es deseable, decidir sobre la división para formar las dos subpruebas antes de que se mande a imprimir la prueba.

Los métodos discutidos en la secciones previas (ya sea formas paralelas o reaplicación con la misma forma), proporcionan al experimentador dos componentes (conjuntos) de calificaciones. En

tal caso la confiabilidad está dada directamente por la correlación producto Momento de Pearson entre los dos conjuntos de calificaciones. Es necesario un método ligeramente modificado cuando se va a calcular la confiabilidad a partir de las calificaciones de dos subpruebas obtenidas de una sola prueba original. Un método, es correlacionar las calificaciones de las dos mitades y entonces substituir esta correlación en la fórmula de Spearman-Brown para el doble de la longitud. podemos escribir:

$$r'_{xx} = \frac{2r_{12}}{1+r_{12}}$$

donde: r'_{xx} Designa la confiabilidad de la prueba total estimada al corregir la correlación por mitades al doble de longitud.

r_{12} designa la correlación entre las dos mitades de la prueba.

Otro método de obtener la confiabilidad de la prueba total partiendo de la información contenida en los dos conjuntos de calificaciones de las subpruebas es utilizando la fórmula presentada por Rulon (1939):

$$r''_{xx} = 1 - \frac{s_d^{\,2}}{x_x^{\,2}}$$

donde: $s_d^{\,2}$ es la varianza de $x_1 - x_2$, la diferencia de las calificaciones en las dos mitades de la prueba.

$s_x^{\,2}$ es la varianza de las calificaciones en la prueba total, la suma de calificaciones de las dos mitades de la prueba: $x = x_1\ x_2$

r''_{xx} se utiliza para designar la confiabilidad de la prueba.

Es deseable no usar la correlación entre dos conjuntos de calificaciones de las subpruebas para la estimación de la confiabilidad, si no dividir la prueba total en tres o posiblemente cuatro partes, y probar la semejanza de estas partes, así como obtener la correlación entre ellas. Estas correlaciones pueden usarse en la fórmula general de Spearman-Brown:

$$R_{K\,\bar{K}} = \frac{Kr_{1\,1}}{1 + (k-1)r\,11}$$

donde: $r11$ es la confiabilidad de la unidad de prueba.

K es el número de reactivos de la prueba alargada dividida entre el número de reactivos de la unidad de prueba, y

$R_{K\,\prime}$ es la confiabilidad de la prueba alargada.

En el caso particular que se está viendo, K sería igual a 3 o 4. Al utilizar este método, se sabe que se está usando una correlación entre tres subpruebas **paralelas** como base para la obtención de la confiabilidad. Esto significa que la confiabilidad encontrada no será muy baja porque no se han escogido subpruebas no paralelas como base para estimarla.

El principal problema al usar calificaciones de subpruebas para el propósito de estimar la confiabilidad, es dividir la prueba original en subpruebas equivalentes. A continuación se considerarán algunos de los métodos para la división de la prueba en subpruebas, así como las ventajas y desventajas de cada uno.

11.2.3 Mitades o tercios sucesivos.

Dividir una prueba en mitades o tercios equivalentes no es cosa sencilla. Por ejemplo, la manera más fácil de dividir la prueba es tomar la primera mitad de la prueba contra la segunda. Con frecuencia un método como éste, no dará como resultado pruebas paralelas. Por ejemplo, si la prueba es administrada en una sola sesión y es contra reloj, cualesquiera reactivos que no se contestaran por falta de tiempo se encontrarían en la segunda mitad de la prueba. La calificación de la segunda mitad, seria menor que la de la primera. Para una prueba de velocidad compuesta de reactivos fáciles, los resultados de graficar las calificaciones de la primera mitad contra las calificaciones de la segunda serían muy peculiares. Todos los sujetos que no llegaron a la segunda mitad, tendrían una calificación de cero en ésta; independientemente de la calificación que hayan obtenido en la primera mitad. Si la prueba es una prueba de velocidad pura, en el sentido de que la mayoría de los Sujetos pueden contestar correctamente a los reactivos si se enfrentan a ellos, en forma tal que los únicos errores serían "los reactivos no intentados", cualquier persona que termina la primera mitad obtiene una calificación perfecta o casi perfecta en ésta, independientemente de su calificación en la segunda mitad. Siempre que la calificación esté en **gran parte** determinada por el hecho de que el tiempo transcurrió antes de que muchos Sujetos hayan terminado, se aproxima uno a la situación antes descrita, y la primera contra la segunda mitad no serán "mitades comparables" adecuadas para obtener una estimación del coeficiente de confiabilidad.

Se puede pensar que, si todos los Sujetos terminaran las dos terceras partes de la prueba, se podría correlacionar el primer tercio de la prueba contra el segundo, y corregir este coeficiente para el triple de longitud. Sin embargo, un método tal es válido únicamente si el último tercio es paralelo a las dos mitades apareadas obtenidas de los primeros dos tercios. Si los reactivos difíciles están al final de la prueba, es imposible de hacer cualquier adivinación plausible al respecto de lo que sucedería si el límite de tiempo fuera aumentado en forma tal que todos pudieran terminar la prueba. Aún más, tal método no da la confiabilidad de la prueba con el límite de tiempo menor. Lo que hace es estimar la confiabilidad que se tendría si el límite de tiempo fuera tal que prácticamente todos terminarán la prueba. Si el limite de tiempo es

importante, se debe usar el método de formas paralelas para estimar la confiabilidad. Si el límite de tiempo es generoso en forma tal que la mayoría de los Sujetos terminen la prueba, es posible estimar la confiabilidad partiendo de las calificaciones de las subpruebas.

Además del problema de los límites de tiempo en una prueba, debe también considerarse el problema de la dificultad de los reactivos. Muchas pruebas están construidas con los reactivos fáciles al principios, los reactivos de dificultad promedio después, y los reactivos más difíciles al final de las mismas: es claro que si los reactivos de la prueba están ordenados de acuerdo a su grado de dificultad, la primera y segunda mitad no serán "mitades comparables". Se puede ver que si una prueba contiene un número de reactivos de dificultad promedio, y es alargada añadiendo reactivos más difíciles, la confiabilidad de la prueba disminuiría a pesar del aumento en la longitud de la prueba y en el tiempo límite. Los nuevos reactivos añadidos serán contestados en una base en el azar por la mayoría de las personas; siendo así que será accidental que contesten en forma correcta o incorrecta a los nuevos reactivos. A medida que se añade un gran número de reactivos difíciles, un componente mayor de la calificación se deberá al proceso de adivinación, y este componente disminuirá la confiabilidad de la calificación de la prueba alargada. Esto de ninguna manera contradice la formulación de Spearman-Brown, sobre la relación de la longitud de la prueba y la confiabilidad (esta formulación dice que entre más larga sea la prueba es más confiable), ya que esta formulación supone que el nuevo conjunto de reactivos es paralelo a los antiguos o anteriores. Esto significa que los reactivos tienen medias, desviaciones estándar y confiabilidad iguales. Los nuevos reactivos supuestamente añadidos aquí, serían reactivos difíciles con una media más baja, y como serían contestados al azar, la confiabilidad de esta nueva parte y su correlación con la parte más fácil de la prueba, estaría más cercana a cero (Gulliksen, 1960,p.203).

Partiendo de consideraciones como éstas, se ve que el efecto de aumentar el límite de tiempo de una prueba, es difícil de predecir. Aumentar el límite de tiempo permitirá que los Sujetos al no conocer la respuesta a los reactivos más difíciles al final de la prueba procederán a adivinar las respuestas a estos reactivos y añadir al azar su calificación. Este incremento no permanecerá estable de forma a forma; por lo tanto disminuirá la confiabilidad de la prueba.

Si se desea usar la primera y segunda mitades (o los tercios sucesivos) de una prueba para calcular su confiabilidad, es posible planear una prueba en forma tal que se superen los problemas producidos por los tiempos límites y el grado de dificultad de los

reactivos: para el método de la primera contra la segunda mitad, por ejemplo, se arreglan los reactivos de la prueba de tal manera que el rango del grado de dificultad en la primera parte de la prueba sea respetado en la segunda parte. Entonces, si se da suficiente tiempo como para que todos o casi todos tengan la oportunidad de terminar la prueba, la primera y la segunda parte serán comparables.Si existe un efecto de **práctica o fatiga** a medida que el Sujeto avanza a lo largo de la prueba, pero si la prueba es administrada en dos sesiones, con tiempo entre ellas para descanso y relajación, si el grado de dificultad de los reactivos es igual en ambas sesiones, y se dan tiempos comparables para cada sesión, es probable que se pueda obtener una buena estimación de la confiabilidad correlacionando los resultados de la primera sesión contra los de la segunda.

11.2.3.2 División por reactivos nones contra pares.

Con mucho, la forma más común de mitades comparables es la división de reactivos en pares y nones. Es probable que este método nunca de un valor muy bajo para el coeficiente de confiabilidad. Si hay error siempre será en la dirección de una confiabilidad que sea espuriamente alta. Algunas veces, como se verá, la confiabilidad de pares y nones sobrestima seriamente la confiabilidad de la prueba indicada por el método de formas paralelas.

Se puede ver fácilmente que, si los reactivos están ordenados de acuerdo con su grado de dificultad, los reactivos nones tendrán aproximadamente la misma dificultad promedio y dispersión de dificultad que los reactivos pares. Si existe alguna influencia, es posible que los reactivos nones sean en promedio ligeramente más fáciles que los reactivos pares.

Sin embargo, al usar este método, se debe asegurar que no existe dependencia de un reactivo con otro. En algunas pruebas se encuentran una serie de preguntas al respecto de un tópico dado, y es algunas veces difícil decidir si los reactivos son independientes, en el sentido de que el conocer la respuesta dependa principalmente de si se ha o no estudiando el tema o si existe una dependencia espuria entre ellos. En las pruebas de ejecución, cuando el Sujeto tiene que armar o desarmar un mecanismo, y se le califica en los diversos pasos, es muy probable que exista una relación espuria, en el sentido de que el Sujeto aprende o no un cierto conjunto de actos como una unidad, mientras que el examinador para poder calificar la ejecución en forma objetiva, establece una cantidad de divisiones más bien artificiales. En casos como estos, parece que la pregunta que se debe responder es: ¿establecería usted, como constructor de exámenes, tales mitades como pruebas separadas?. En un conjunto de afirmaciones que describen las características de los conos y los

bastones del ojo, por ejemplo, es posible que el constructo de la prueba pudiera usar sólo la mitad de las afirmaciones para tener una forma corta de la prueba. Sin embargo, es muy posible que los reactivos nones no constituyan una forma paralela satisfactoria de los reactivos pares. Los reactivos deberán de inspeccionarse para asegurar que el tipo de materia que cubren y la distribución de su dificultad de una de las mitades es **grosso modo** paralela al de la otra mitad.

La correlación pares-nones también es espuriamente alta en una prueba con un límite de tiempo muy pequeño porque un gran número de sujetos no terminan la prueba. Si un Sujeto no contesta los diez últimos reactivos de la prueba, obviamente no "pasa" ninguno de ellos. En esta forma obtiene cinco puntos más de error en su calificación non y también cinco puntos más de error en su calificación par. Es muy probable que una observación cuidadosa demuestre que muchas de las confiabilidades publicadas son espuriamente altas debido a este factor. Una vez más, este tipo de error queda ilustrado en forma muy clara en las pruebas de velocidad a las que se hizo referencia con anterioridad. Si cada Sujeto tiene todos los reactivos correctos hasta donde haya llegado, el que termine diez reactivos tendrá una calificación non de cinco y una calificación par de cinco, si termina once reactivos tendrá una calificación non de seis y una calificación par de cinco, y con doce reactivos, la calificación será de seis y seis. Esto es, la calificación nones y pares serán ya sea idénticas, o la calificación non será un punto más alta que la calificación par.

Deberá notarse que la confiabilidad pares - nones puede ser muy alta, aún cuando los reactivos estén ordenados de acuerdo a su grado de dificultad si les permite a los sujetos terminar la prueba, y los reactivos sean independientes uno de otro (en el sentido de que cometiendo un error en uno de ellos no necesariamente aumenta la probabilidad de cometer un error en otro reactivo). La variabilidad debida a las variaciones de un día al otro en la habilidad, queda descontada y aún si la variación que pudiera ser causada por un efecto ligero de práctica o fatiga a medida que se progresa a lo largo de la prueba también queda descartada, si se usa el método de las formas paralelas como norma. La confiabilidad pares - nones, como se aplica por lo general a la mayoría de las pruebas, es muy probable que arroje un resultado bastante alto debido a que se pueden controlar diversas fuentes de variación y también debido a que por lo general, la mayoría de las pruebas tienen límites de tiempo logrando en esta forma, una buena proporción de la calificación ya que la mayoría de los Sujetos, no tienen oportunidad de intentar contestar los últimos reactivos.

En una prueba de velocidad en la que la calificación depende de qué tan rápido trabaja un Sujeto en el tiempo Límite dado, **no hay forma de estimar la confiabilidad** si no es aplicando una **prueba** o **forma paralela** una segunda vez. Ahora bien, diferentes métodos de medir la confiabilidad dan diferentes resultados: en general, la confiabilidad de formas paralelas es la más baja, y la pares - nones (corregida) es la más alta.

Se puede pensar que, si todos terminaron dos tercios de la prueba se podría usar una confiabilidad pares - nones en los dos primeros tercios, obtener la correlación entre estos dos, y corregirla al triple de longitud. Sin embargo, esto proporciona una estimación de la confiabilidad de la prueba total sobre la suposición de que **todos terminen** la prueba. No da una estimación del grado con el que un Sujeto alcanza la **misma tasa de velocidad** en diferentes administraciones de la prueba, y que por lo tanto, llegue al mismo punto en la prueba. No hay forma posible de estimar este factor con exactitud, excepto dando formas paralelas con **tiempos límites comparables** y bajo instrucciones estandarizadas, y cuidando el grado en el que las calificaciones sean las mismas.

11.2.3.3 Subpruebas apareadas al azar:

Si las calificaciones son obtenidas en una sola aplicación de la prueba que va a usarse para estimar la confiabilidad de la misma, es necesario considerar a esta calificación única como dividida en dos, tres o cuatro calificaciones de subpruebas equivalentes. En las secciones anteriores se ha visto que bajo ciertas condiciones las mitades o tercios sucesivos de una prueba pueden ser razonablemente consideradas como formas paralelas, mientras que bajo otras condiciones los segmentos sucesivos de una prueba, no son paralelos en forma clara. De manera semejante, asignar cada segundo o tercer reactivos a una, dos o tres sub-pruebas puede ser un buen o mal método según diferentes condiciones para obtener subpruebas paralelas.

Si una prueba está compuesta de un gran número de reactivos independientes y es administrada con un tiempo límite "normal", se puede por lo general subdividir en subpruebas paralelas. Si una prueba tiene pequeños grupos de reactivos, como por ejemplo en la mayoría de la pruebas mecánicas o en pruebas que involucran escribir un párrafo, puede ser o no posible construir una prueba que esté compuesta por subpruebas paralelas. Si se usa un tiempo límite pequeño, no hay posibilidad de obtener alguna estimación válida de la confiabilidad usando calificaciones de sub-pruebas.

Si se tienen datos de análisis de reactivos de una prueba (que tiene un gran número de reactivos independientes y un tiempo límite liberal), los reactivos deberán de ser apareados en base en los

datos del análisis de reactivos y asignados a las subpruebas. Este es un método excelente de asegurar que las subpruebas sean paralelas. Por ejemplo, supóngase que se dispone del porcentaje de personas que contesta correctamente cada reactivo (p) y se tiene también su correlación r_b biserial con el resto de los reactivos de la prueba. El mejor procedimiento para construir subpruebas paralelas es representar a cada reactivo por medio de un punto en un diagrama de dispersión, la abcisa para p y la ordenada para r_b. Para poder identificar los reactivos, cada punto, deberá estar señalado con el número del reactivo, como se demuestra en la fig. 11.1. Entonces los reactivos pueden ser apareados simultáneamente en p y r, trazando una línea alrededor de los pares, tríos o cuádruples apareados.

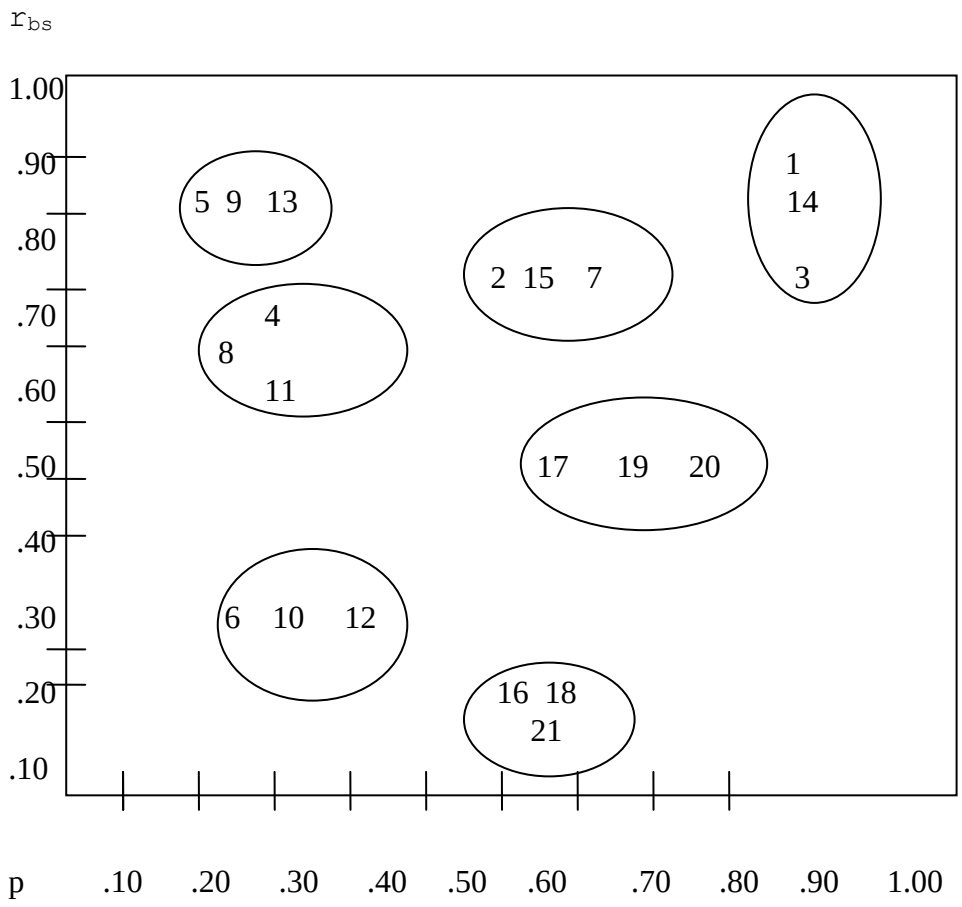

Figura 11.1 Cómo construir tres subpruebas o pruebas paralelas apareando simultaneamente reactivos sobre índices de dificultad y confiabilidad.
p = proporción de Sujetos que pasa el reactivo o lo endosa afirmativamente
r_{bs} = correlación reactivo-calificación total.

Es importante hacer notar que si la prueba es heterogénea con respecto al tipo de reactivos, o con respecto al tipo de material cubierto, es necesario aparear los reactivos en esos aspectos, así como en relación a p y r.

Se deberá entonces, asignar el azar, a un miembro de cada grupo a una subprueba dada. Por ejemplo, si solamente se forman dos subpruebas la asignación podría determinarse a través de un "volado", y asignar el reactivo con el número más pequeño del par a la forma A si "cae águila" y a la forma B si "cae sol". Al construir tres subpruebas paralelas, es necesario asignar a cada triada de reactivos a las diferentes tres subpruebas paralelas por medio de un procedimiento un poco más complicado. Por ejemplo los reactivos de cada tríada pueden identificarse a través de su número de reactivos como bajo, medianos y altos (B, M y A). Existen por lo tanto seis posibles manera de asignar estos tres reactivos, uno a cada una de las tres subpruebas. A cada uno de esos órdenes se les puede entonces asignar un número del 1 al 6 (1=BMA;2=BAM; etc.) y cada triada asignarse de acuerdo a la tirada de un dado.

Si la información del análisis de reactivos se encuentra disponible antes de que se constituya la prueba, se pueden simplificar las rutinas de calificación si los reactivos de una subprueba se ponen en primer lugar, y luego los de la segunda, etc.; o si los reactivos de las diferentes subpruebas se distribuyen con cierta secuencia en forma sucesiva a lo largo de la prueba.

Debe hacerse énfasis en que no importa en qué orden se emplean los reactivos, pero que sí es necesario permitir tiempo suficiente para que la mayoría de los Sujetos terminen casi todos los reactivos. No es posible estimar la confiabilidad de una prueba a partir de subpruebas paralelas si la calificación de la prueba se ve marcadamente influida por un tiempo límite.

11.2.4 Correlación Intraclase

Uno de los procedimientos más utilizados en la estimación del coeficiente de confiabilidad es el que se sigue en la obtención del coeficiente de correlación intraclase:

$$p = 1 - \sqrt{\Sigma\ x_1\ x_2\ /2\ n\ \sigma\ c^2}$$

donde: x_1 es el puntaje obtenido, por un sujeto particular, en la primera medición y x_2 es el puntaje obtenido por el mismo individuo en la segunda medición, n representa el número de sujetos que

respondieron al instrumento y σc^2 corresponde a la varianza de los puntajes combinados de ambas mediciones.

Dos aspectos evidentes en la fórmula de la correlación intra clase son:

-que la confiabilidad depende de la precisión de las medidas, es decir, de la menor discrepancia entre los puntajes de la primera y segunda medición:

-que la confiabilidad depende de la población medida por el instrumento, en donde, el grado de precisión es relativo a la variabilidad que se da entre todos los puntajes. Una determinada discrepancia promedio producirá un alto coeficiente de confiabilidad si la variabilidad intersujeto es grande al compararla cuando ésta es pequeña.

Un alto coeficiente de confiabilidad indica que el instrumento de medición utilizado localiza de manera precisa a cada sujeto en relación a todos los otros sujetos que también fueron medidos. Sin embargo, en dicha localización sólo puede asumirse que ésta se ha hecho sobre la dimensión de un atributo. La determinación de si ese atributo corresponde efectivamente al propósito de la medición es problema de la validez.

11.2.5 Consistencia Interna.

11.2.5.1 Alpha de Cronbach.

La ecuación Alpha de Cronbach para determinar la confiabilidad del tipo de consistencia interna inter-reactivos, es una de las deducciones más importantes de la teoría del error de medición (Nunally y Bernstein, 1993). La fórmula es la siguiente:

$$r_{kk} = k / k - 1 \ (\ 1 - \Sigma\sigma^2_i \ / \sigma^2_y \ )$$

donde, K es igual al número de reactivos que componen el instrumento;

$\Sigma\sigma^2_i$ es igual a la suma de la varianza de los reactivos (en la matriz de covarianza se obtiene sumando los elementos de la diagonal principal).

σ^2_y es igual a la varianza total, (en la matriz de covarianza se obtiene sumando los elementos de la diagonal principal más dos veces la suma de los elementos que se encuentran fuera de dicha diagonal).

Se puede derivar la misma fórmula del modelo de las pruebas paralelas, y se pueden derivar fórmulas semejantes a partir de otros modelos matemáticos de la medición del error. Esta fórmula representa la correlación que se espera de una prueba con una forma

alternativa que contenga el mismo número de reactivos. La raíz cuadrada del coeficiente Alpha es la correlación estimada de una prueba con calificaciones verdaderas sin error. Esta fórmula debería aplicarse en forma rutinaria a todas las pruebas nuevas.

El coeficiente Alpha se podrá calcular para una prueba o escala con opciones de respuesta múltiple (más de dos) siempre y cuando **todos** los reactivos de la prueba tengan el mismo número de opciones de respuesta. EL procedimiento es el siguiente:

1. Se califica cada reactivo con el peso de la opción marcada por el sujeto
2. Se califica toda la prueba, usando los pesos de los reactivos individuales: calificación total.
3. Se prepara una matriz de puntajes de acuerdo con la figura 11.2. En ella se escriben los puntajes que obtiene cada Sujetos en cada reactivo y la calificación total.

	REACTIVOS								
SUJETOS	1	2	3	4	5	.	.	.	K	C.T.
1										
2										
3										
.										
.										
.										
n										

FIGURA 11.2 MATRIZ DE CALIFICACIONES

4.Se obtiene la media para cada columna, incluyendo la de las calificaciones totales (C.T.). Se prepara entonces otra matriz de desviaciones, donde se anota en cada celdilla, la desviación de cada puntaje con respecto a su media, la de su columna.

5.Se procede entonces a elevar al cuadrado cada desviación y se suman éstas para cada columna dividiendo la suma entre el número de sujetos (n), obteniendo así la varianza de cada reactivo y la de calificación total(σx^2).

6. por último se procede a substituir los valores en la ecuación y se obtiene r_{kk}

7. Se busca el nivel de significancia de r_{kk} con N-K grados de libertad.

Cuando se efectúa la investigación de la confiabilidad de una prueba compuesta por reactivos dicotómicos (sí y no; falso-verdadero, correcto-incorrecto), el coeficiente Alpha adopta la siguiente fórmula especial:

$$r_{kk} = k/k-1 \ (1 - \Sigma \ pq \ / \ \sigma_{2y} \ )$$

donde:

$\sum pq$: la suma del producto de la proporción de sujetos que contesten una opción (correcta, o aquella que tenga el peso de 1) por la proporción de sujetos que contestaron en la otra opción (o sea 1-p).

Los pasos a seguir para la determinación de r_{kk}.en la fórmula anterior son los siguientes:

1. Se encuentra el valor p de cada reactivo, que se multiplica entonces por 1-p.
2. Se suman estos productos
3. Se calcula la varianza de las calificaciones totales (σt^2).
4. Se divide la $\sum pq$ entre σt^2.
5. Se resta este número de 1
6. Se multiplica el resultado de (5) por la proporción del número de reactivos entre ese número menos 1
7. Se determina su nivel de significancia.

Esta versión del coeficiente Alpha se conoce también como la "fórmula 20 de Kudder Richarson" (KR-20).

Por otra parte se recordará que el coeficiente de confiabilidad de cualquier prueba es la correlación promedio estimada de esa prueba con todas las posibles pruebas de la misma longitud cuando se propone que las dos pruebas miden la misma cosa. El coeficiente Alpha también se puede derivar como la correlación que se espera entre una prueba real y una forma hipotética alternativa. Si llamamos X a la prueba real y "Y" a la prueba hipotética, entonces la matriz de la varianza total para todos los reactivos se puede esquematizar como se ve en la fig. 11.3.

	X	**Y**
X	Cx	Cxy
Y	Cxy	Cy

Figura 11.3: Matriz de la varianza total.

A partir del modelo dominio-muestra se espera que el término diagonal promedio en Cx sea el mismo que en Cy y que el promedio de los elementos fuera de la diagonal en las dos matrices sea el mismo También se espera que el elemento promedio a lo largo de Cxy sea igual al promedio del elemento fuera de la diagonal en Cx. Por lo

tanto,se puede derivar el coeficiente Alpha a partir de la correlación de suma, como sigue:

$$r_{xy} = \frac{\bar{C}xy}{\sqrt{\bar{C}_x}\sqrt{\bar{C}_y}}$$

De acuerdo con el modelo, $\bar{C}x$ es aproximadamente igual a $\bar{C}y$, de manera tal que la ecuación anterior se puede volver a escribir como sigue:

$$r_{xy} = \frac{\bar{C}xy}{\bar{C}x}$$

De acuerdo con el modelo, el coeficiente promedio en Cxy (y así la suma de los coeficientes) se puede derivar de Cx. Primero sería necesario restar de Cx las varianzas de los reactivos que se encuentran en la diagonal. Después seria necesario inflar el resultado por el factor desarrollado previamente, es decir, K/(K-1), lo que nos lleva de nuevo al coeficiente alpha.

11.2.5.3 Análisis de Reactivos

El análisis de reactivos es otro procedimiento que se sigue en la búsqueda de consistencia interna. Es decir, cualquier operación que implique un análisis de la varianza de los elementos componentes de un instrumento, proporciona una estimación de la consistencia interna. Un procedimiento general y un tanto diferente al de al Alpha de Cronbach es el propuesto por Cureton (1966), que permite seleccionar reactivos que muestran correlaciones más altas con el resto de los mismos, y por lo tanto, los que se correlacionan más alto con la puntuación total. Una ventaja de este procedimiento es que es aplicable a reactivos dicotómicos o múltiples, además de que en el cálculo de las correlaciones también se prevé, la corrección necesaria que elimina el falso incremento del valor de los coeficientes que se produce cuando se incluye en el puntaje total el reactivo que se está analizando.

REFERENCIAS

Díaz Guerrero, R. y Salas, M.(1975). **El Diferencial Semántico del Idioma Español**. México: Editorial Trillas.

Edwards, A.L.(1957). **Techniques of Attitude Scale Construction**. Nueva York: Appleton-Century-Crofts.

Gulliksen, H.(1950). **Theory of Mental Tests**. Nueva York: John Wiley and Sons.

Gulliksen, H.(1950b). The reliability of speeded tests. **Psychometrika**, 15, 259-269.

Guttman, L.(1944). A basis for scaling qualitative data. **American Sociological Review**, 9, 139-150.

Guttman, L. (1945). **Questions and answers about scale analysis**. Research Branch, Information and Education Division, Army Service Forces. Report D-2.

Guttman, L.(1946). An approach for quantifying paired comparisons and rank order. **Annals of Mathematical Statistics**, 17, 144-163.

Guttman, L. (1947a). Suggestions for further research in scale and intensity analysis of attitudes and opinions. **International Journal of Opinion and Attitude Research**, 1, 30-35.

Guttman, L. (1947b). The Cornell Technique for scale and intensity analysis. **Educational Psychological Measurement**, 7, 247-280.

Kelley, T.L.(1921). The reliability of test scores. **Journal of Educational Research**,3, 370-379.

Likert, R.(1932). A technique for the measurement of attitudes. **Archives of Psychology**. No. 140.

Lincoln, E.A.(1932). The unreliability of reliability coefficients. **Archives of Psychology**, 140.

Lincoln, E.A.(1933). Reliability coefficients are still unreliable. **Journal of Educational Psychology**, 24, 235-236.

Nunnally, J.C. y Bernstein,I.J.T. (1993). **Teoría Psicométrica**. México: McGraw Hill.

Osgood, Ch.E., Suci, G.J., Tannenbaum, P.H.(1957). **The Measurement of Meaning**. Urbana: University of Illinois Press.

Ottis, A.S. y Knollin, H.E.(1921). The reliability of the Binet Scale and of pedagogical scales. **Journal of Educational Research**, 4, 121-142.

Paulsen, G.(1931). A coefficient of trait variability. **Psychological Bulletin**, 28, 218-219.

Preston, M.G.(1940). Psychophysical measurement methods. **Psychological Bulletin**, 35, 63-83.

Rulon, P.J. (1939). A simplified procedure for determining the reliability of a test by split-halves. **Harvard Educational Review**.9, 99-103.

Spearman, Ch.(1904a). The proof and measurement of association between two things. **American Journal of Psychology**, 15, 72-101.

Spearman, Ch.(1904b). "General Intelligence" objectively determined and measured. **American Journal of Psychology**, 15, 201-295.

Spearman, Ch.(1907). Demonstration of formulae for true measurement of correlation. **American Journal of Psychology**, 18, 161-169.

Spearman, Ch. (1910). Correlations calculated with faulty data. **British Journal of Psychology**, 3, 271-295.

Spearman, Ch. (1913). Correlations of sums and differences. **British Journal of Psychology**, 5, 417-426.

Thouless,R.H.(1936). Test unreliability and function fluctuation. **British Journal of Psychology**, 26, 325-343.

Thouless,R.H. (1939). The effects of errors of measurement on correlation coefficients. **British Journal of Psychology**, 29, 383-403.

Woodrow, H.(1932). Quotidian variability. **Psychological Review**, 39, 245-256.