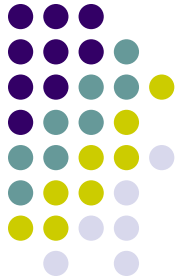
La calidad de los instrumentos de evaluación.

Dr. Pedro A. Díaz Rojas

- Profesor Titular, especialista de II grado en Histología y en Administración de Salud. Máster en Educación Médica. Doctor en Ciencias de la Educación Médica.
- Correo electrónico: pdiaz@infomed.sld.cu

EVALUACIÓN DEL EDUCANDO

Evaluación del aprendizaje o
Evaluación para el aprendizaje



CALIDAD DE INSTRUMENTOS

CUALIDADES DE UNA PRUEBA

VALIDEZ
FIABILIDAD
OBJETIVIDAD
EQUILIBRIO
EQUIDAD
ESPECIFICIDAD
DISCRIMINACION
EFICACIA
TIEMPO
LONGITUD

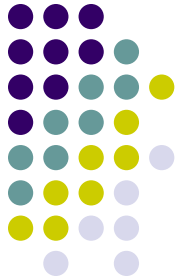
ERRORES FRECUENTES

FUTILIDAD
INEXACTITUD
AMBIGÜEDAD
CONSERVADURISMO
COMPLEJIDAD
PODER SUGESTIVO

¿Qué sería lo ideal?

- Medición de objetivos importantes.
- Comprobación de la capacidad de comprensión y de aplicación de principios.
- Comprobación de la capacidad de pensar críticamente y de resolver problemas nuevos
- Comprobación de la capacidad para seleccionar hechos y principios relevantes integrándolos para la solución de problemas complejos.

CALIDAD DEL INSTRUMENTO



1. ¿Mide el instrumento lo que realmente se debe medir?

Validez

Confiabilidad

2. ¿Con qué precisión y estabilidad se mide lo que estamos midiendo?

3. ¿Los Resultados de un Grupo son comparables con los de los otros Grupos?

Generalizabilidad

Otros aspectos:

- Objetividad
- Eficacia
- Pertinencia
- Discriminación
- Equilibrio
- Equidad
- Dificultad
- Certidumbre
- Consistencia interna

Grupos de Factores

L. J. Cronbach

1. Características generales y duraderas del examinado.
2. Características duraderas y específicas del individuo.
3. Características generales y momentáneas del examinado.
4. Características temporales y no generales del examinado.

1. Características Generales y duraderas del examinado.

- **Habilidades generales.**
- **Habilidades para comprender las instrucciones.**
- **Habilidades para resolver problemas.**
- **Actitudes, emociones y hábitos del evaluado.**

2. Características duraderas y específicas del individuo.

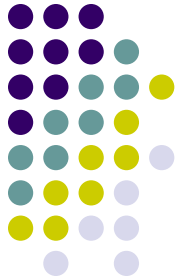
- Conocimientos y habilidades que requieren los problemas del instrumento.
- Actitudes, emociones y hábitos del evaluado.

3. Características generales y momentáneas del examinado.

- **Salud, fatiga y tensión psíquica.**
- **Motivación de ser evaluado.**
- **Efectos del calor, luz, ventilación, comodidad…**
- **Actitudes, emociones y hábitos.**

4. Características temporales y no generales del examinado.

- Cambios por fatiga o motivación.
- Fluctuaciones en la atención, coordinación o patrones de juicio.
- Fluctuaciones en la memoria.
- Experiencias en los conocimientos y habilidades requeridos.
- Buena suerte al “adivinar” las respuestas.



CONFIABILIDAD

CONFIABILIDAD

Expresión cuantitativa de la reproducibilidad con la que un instrumento mide un mismo atributo.

Estabilidad de los resultados:
- En el tiempo.
- Entre profesores.

Consistencia Interna.

FIABILIDAD

Constancia con que un instrumento mide una variable determinada.

REPLICABILIDAD

Imposibilidad de aplicar un test un número infinito de veces.

Imposibilidad de someter las respuestas a un número infinito de calificadores.

FACTORES QUE MODIFICAN LA FIABILIDAD

- Mayor interrelación de los items.
- Mayor posibilidad de acierto azaroso.
- Mayor introducción de elemen-tos extraños o capciosos.
- Inclusión de elementos que pro-ducen falsas interpretaciones:
  - -Imprecisión de los términos.
  - -Extensión exagerada del item.
  - -Uso palabras no conocidas.
  - -Estructura defectuosa.
  - -Instrucciones inadecuadas.

- Mayor número de items.
- Menor rango en dificultad de items.
- Mayor objetividad en calificación.
- Mayor homogeneidad en la prueba.

PRUEBA

- Imprecisión en la respuesta.
- Interrupciones durante el test.
- Fraude.

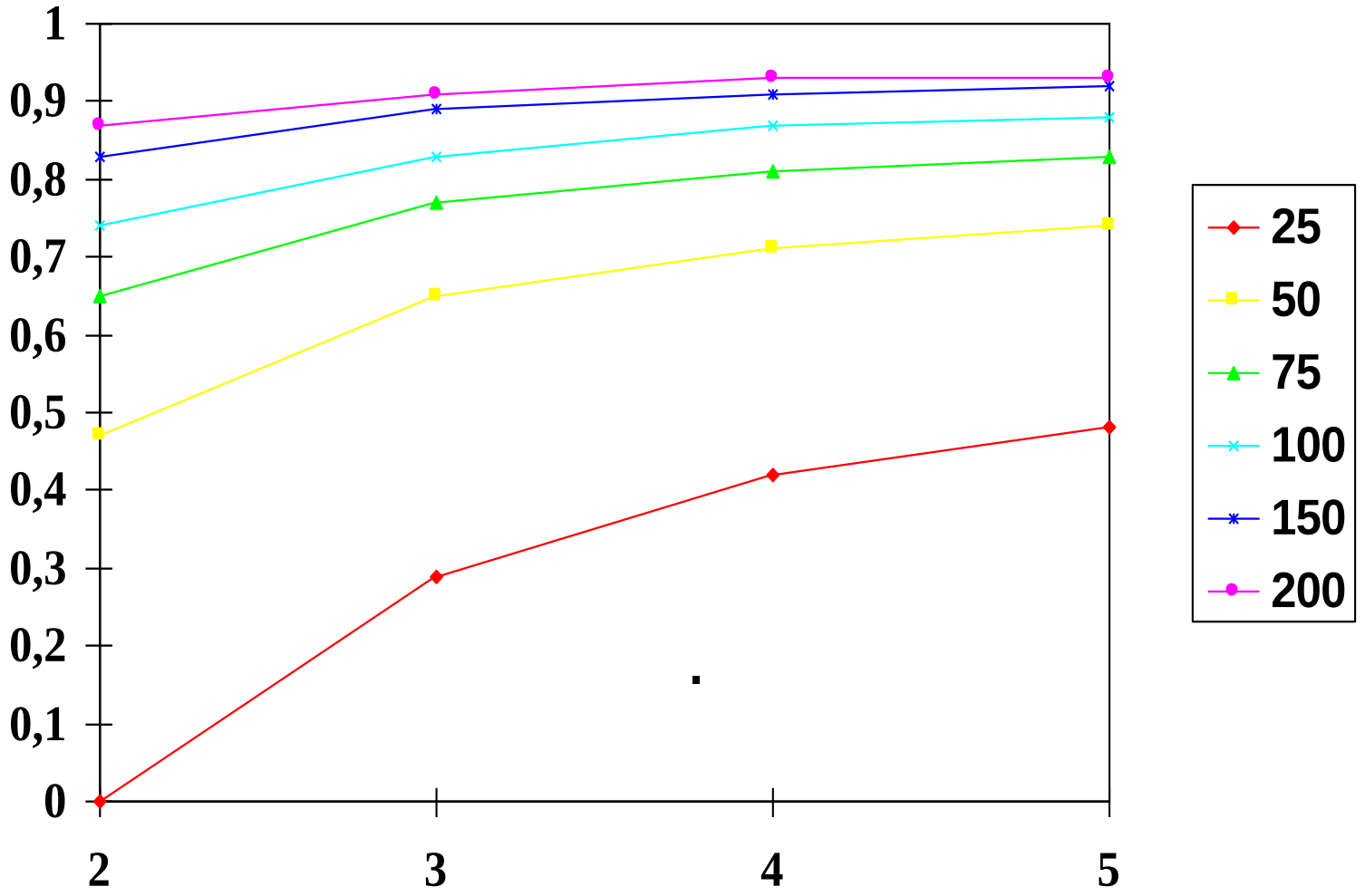
ALUMNO

- Nivel de motivación elevado.
- Velocidad en la realización.

- Efecto de "halo"
- Diversidad de criterios.
- Ausencia de clave de respuesta.
- Cansancio físico y mental.

PROFESOR

- Fórmula para determinar el efecto que la variación del número de items produce en la fiabilidad esperada



Coeficientes de confiabilidad

- Coeficiente theta (θ).
- Coeficiente omega (ω).
- Coeficiente alfa max (α_0) .
- Coeficiente de Kuder Richarson.
- COEFICIENTE ALFA (A) DE CRONBACH.

Coeficiente â de Cronbach

- **Es el más utilizado mundialmente.**
- **Es el mejor para estimar el error de muestreo.**
- **Proporciona una medida efectiva de la consis-tencia interna del instrumento evaluativo.**
- **Puede emplearse ante diferentes formatos de preguntas.**

0,60 – 0,80

CALIFICACIÓN DE LAS PREGUNTAS

ESTUDIANTES	PI	P2	P3	P4	CALIFICACIONES TOTALES
No. 1	20,7	20,8	19,0	23,1	83,6
No. 2	13,5	13,3	12,3	13,5	52,6
No. 3	16,8	15,0	15,1	17,3	64,2
No. 4	16,4	18,8	16,3	16,5	68,0
No. 5	18,8	16,7	20,6	19,2	75,3
No. 6	10,9	10,4	11,1	10,0	42,4
No. 7	04,3	07,1	05,6	06,2	23,2

coeficiente $\hat{\alpha}$

$$\hat{\alpha} = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k} s_i^2}{s_x^2}\right]$$

VARIANZA DE LA i-ÉSIMA PREGUNTA

VARIANZA DE LA CALIFIC TOTAL

CANTIDAD DE PREGUNTAS

VARIANZA

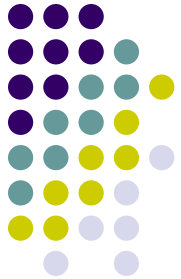
$$S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\left(\bar{x}\right)^2\right]$$

NÚMERO DE DATOS

SUMA DEL CUADRADO DE LOS DATOS

PROMEDIO DE LOS DATOS

DATOS: 12; 10; 15; 13; 11; 14; 14



VALIDEZ

VALIDEZ

Muestra hasta que punto
un examen mide "realmente"
lo que se espera que deba medir

Tipos de Estudios de Validez

Validez de Contenido

Validez de Criterio

Validez de Construcción

Validez Funcional

Validez de Contenido

NO SE DETERMINA ESTADÍSTICAMENTE

RESULTA DEL JUICIO DE EXPERTOS QUE ANALIZAN LA REPRESENTATIVIDAD

PARA ELLO

IDENTIFICAR QUE EXISTA UNA MUESTRA REPRESENTATIVA DE

COMPORTAMIENTOS
AREAS DE CONTENIDOS

Validez de Contenido

Si constituye una **MUESTRA REPRESENTATIVA** de los conocimientos y habilidades **ESENCIALES**

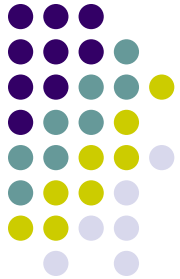
Dominio Cognitivo

Objetivos Educacionales

Marco de Referencia

Criterio de Expertos

Validez de Contenido



Etapas

1. Definición del dominio de competencias a evaluar en correspondencia con los objetivos.
2. Selección del panel de expertos calificados en dicho Dominio.
3. Esquema para correlacionar las preguntas con el dominio de competencias y los objetivos.
4. Colectar y procesar el análisis.
5. Informe, conclusiones y recomendaciones.

Validez de Contenido

Consideraciones prácticas

1. ¿Deben ser ponderados los objetivos para reflejar su importancia?
2. ¿Cómo se puede estructurar el proceso de correlación?
3. ¿Qué aspectos de la pregunta deben ser analizados?
4. ¿Cómo se deben concluir los resultados?

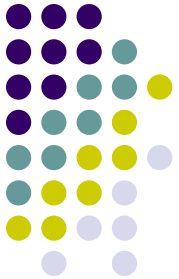
Validez de Contenido

¿Cómo concluir el análisis?

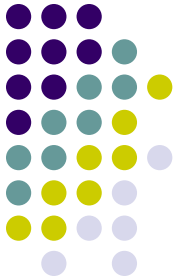
- **Porcentaje de Ítems que se correlacionan con los objetivos.**
- **Porcentaje de Ítems que correlacionan con núcleos de contenidos básicos.**
- **Correlación entre la importancia ponderada de los objetivos y el número de Ítems que los miden.**
- **Porcentaje de objetivos NO evaluados por ningún Item.**

GENERALIZABILIDAD

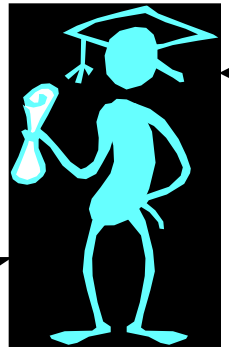
Análisis de los componentes del error de medición y determinar la contribución de cada uno de ellos en el error total calculado.



CALIDAD DE LOS ITEMS



ÍNDICE DE DIFICULTAD



DIFERENCIAR

RENDIMIENTO

DIFÍCILES

FÁCILES

Índice de dificultad.

Se calcula según la siguiente fórmula:

$$D = \frac{A}{Nt}$$

Donde:

D = Índice de dificultad del inciso.

A = Número de aciertos del inciso.

Nt = Número total de presentados en el examen.

Índice de dificultad de la temática.

Se calcula el grado de dificultad el cual se corresponde con el % de aprobados con respecto al total de examinados, considerando la media de errores cometidos en la temática como el número de desaprobados en ella, como se aprecia en la fórmula:

$$Dt = \frac{\frac{\Sigma\,(At1 + At2 + \ldots + Att)}{nt}}{Nt}$$

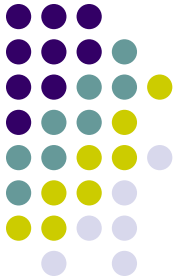
Donde:

Dt = Índice de dificultad de la temática.

At = Número de respuestas correctas de los incisos de la temática.

nt = Número de incisos de la temática.

Nt = Número total de presentados en el examen.



GRADO DE DIFICULTAD

Grado de dificultad.

$$GD = \frac{GS + GI}{N} * 100$$

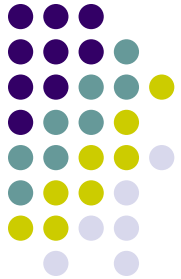
Se define como el porcentaje de estudiantes, tanto del grupo superior global como del inferior global, que contestaron correctamente el reactivo

Donde:

= Número de estudiantes, en el grupo superior global, que contestaron correctamente el reactivo.

= Número de estudiantes, en el grupo inferior global, que contestaron correctamente el reactivo.

= Número total de estudiantes que presentaron la prueba.



DISCRIMINACIÓN

Discriminación.

- Si la prueba y un ítem miden la misma habilidad o competencia, podemos esperar que quien tuvo una puntuación alta en todo el test deberá tener altas probabilidades de contestar correctamente el ítem.
- También debemos esperar lo contrario, es decir, que quien tuvo bajas puntuaciones en el test, deberá tener pocas probabilidades de contestar correctamente el reactivo.
- Así, un buen ítem debe discriminar entre aquellos que obtuvieron buenas calificaciones en la prueba y aquellos que obtuvieron bajas calificaciones.
- El máximo valor del índice de discriminación de un reactivo tiene una estrecha relación con el grado de dificultad del mismo, ya que un reactivo con valor extremo de grado de dificultad (muy alto o muy bajo) en realidad no discrimina, puesto que es resuelto por la gran mayoría de estudiantes o por casi ninguno respectivamente.

Procedimientos para analizar la discriminación.

1. Ordene los exámenes desde la puntuación más alta hasta la más baja.

2. Separe la tercera parte correspondiente a los exámenes con calificaciones más altas y el mismo número de exámenes con puntuaciones bajas

3. Con cada ítem, cuente el número de estudiantes del grupo superior que eligió cada opción. Haga lo mismo con el grupo inferior. Aplique la fórmula,

Índice de discriminación del reactivo *i.*

$$Di = \frac{\textit{GA de aciertos} - \textit{GB de aciertos}}{\textit{Ni grupo mayor}}$$

Donde:

Di = Indice de discriminación del reactivo **i.**

GAaciertos = Número de aciertos en el reactivo ***i*** del 27% de personas con las puntuaciones más altas en el test.

GBaciertos = Número de aciertos en el reactivo ***i*** del 27% de personas con las puntuaciones más bajas en el test.

Ngrupomayor = Número de personas en el grupo más numeroso (GA o GB).

Interpretación de D.

Entre más alto es el índice de discriminación, el reactivo diferenciará mejor a las personas con altas y bajas calificaciones.

Si todas las personas del GA contestan correctamente un reactivo y todas las personas del GB contestan incorrectamente, entonces

D = 1 (valor máximo de este indicador);

si sucede lo contrario,

D = -1 (valor máximo negativo);

si ambos grupos contestan por igual,

D = 0 (valor mínimo dediscriminación).

Poder de discriminación de los reactivos según su valor D.

D	Calidad	Recomendaciones
> 0.39	Excelente	Conservar
0.30 - 0.39	Buena	Posibilidades de mejorar
0.20 - 0.29	Regular	Necesidad de revisar
0.00 - 0.20	Pobre	Descartar o revisar a profundidad
< -0.01	Pésima	Descartar definitivamente

Norma discriminativa.

Es un criterio que permite determinar el valor óptimo que debe presentar un reactivo en su índice de discriminación, de acuerdo al grado de dificultad del mismo.

El valor numérico de la norma discriminativa (ND) de un reactivo se determina utilizando la siguiente expresión:

$$ND = \begin{cases} 0.3 * GD & si \quad 27\% \leq GD \leq 73\% \\ 100 - GD & si \quad 73\% \leq GD \leq 100\% \end{cases}$$

Relación discriminativa.

La norma discriminativa es utilizada para elaborar un criterio que indique si un reactivo es aceptable o no según el valor de su índice de discriminación. Este criterio se denomina *Relación Discriminativa* del reactivo y se define como el cociente entre el índice de discriminación del reactivo y su norma discriminativa:

$$RD = \frac{ID}{ND}$$

Resumen.

El valor numérico de los cuatro parámetros hasta aquí revisados (grado de dificultad (), índice de discriminación (ID), norma discriminativa (ND) y relación discriminativa (RD) pueden variar en dependencia del método utilizado para dividir a la población en grupo superior e inferior globales.

Relación discriminativa.

Si el valor de *RD* es mayor que 1, significa que el índice de discriminación es mayor que la norma y, por tanto el reactivo es aceptable. Por el contrario, si es menor a 1, se recomienda analizar el reactivo, en cuanto a contenido y redacción. Si el valor de es inferior a 0.6, el reactivo se considera desechable.

Coeficiente de Discriminación.

El **coeficiente de correlación biserial** (r_{pbis}) se calcula para determinar el grado en que las competencias que mide el test también las mide el reactivo. El r_{pbis} proporciona una estimación de la correlación producto-momento de Pearson entre la calificación total de la prueba y el continuo hipotético del reactivo, cuando éste se dicotomiza en respuestas correctas e incorrectas.

Correlación del punto biserial o Coeficiente de Discriminación.

$$r_{pbis} = \frac{X_1 - X_0}{S_x} * \sqrt{\frac{n_1 n_0}{n(n-1)}}$$

Donde:

X_1 = Media de las puntuaciones totales de aquellos que respondieron correctamente el ítem.

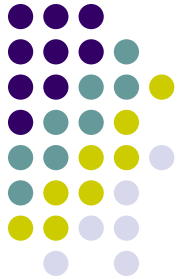
X_0 = Media de las puntuaciones totales de aquellos que respondieron incorrectamente el ítem.

S_X = Desviación estándar de las puntuaciones totales.

n_1 = Número de casos que respondieron correctamente el ítem.

n_0 = Número de casos que respondieron incorrectamente el ítem.

$n = n_1 + n_0$



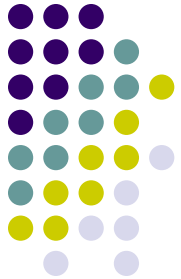
DISTRACTORES

Eficacia de los distractores.

- Determinamos la eficacia de las respuestas de distracción, comparando el número de estudiantes de los grupos superior e inferior que eligieron cada opción incorrecta.

- Una buena respuesta de distracción resultará atractiva para más estudiantes del grupo inferior que del grupo superior.

- Cuando un distractor no es seleccionado por ningún estudiante, decimos que no funcionó.



PASOS PARA EL ANÁLISIS

Pasos:

1. **Revisar la tabla de distribución contenido-tiempo, elaborada con el P1, según temas, núcleos de contenidos básicos y esencialidades.**
2. **Codificar todos los ítems del examen.**
3. **Agregar a la tabla de distribución contenido-tiempo la cantidad de ítems que le corresponden.**
4. **Tabular los resultados de los ítems, en código aciertos-desaciertos o correctos-incorrectos.**
5. **Al cierre de la columa de una pregunta se pone la nota de la pregunta y se cierra con la nota del examen.**

Pasos:

6. **Sobre la base del patrón internacional de calidad de exámenes se elabora la frecuencia teórica esperada.**
   - **5% de incisos fáciles.**
   - **20% de incisos medianamente fáciles.**
   - **50% de incisos de dificultad media.**
   - **20% de incisos medianamente difíciles.**
   - **5% de incisos difíciles.**

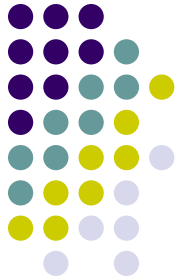
Pasos:

7. **Determinación del Índice de Dificultad de cada uno de los incisos y de las temáticas del programa analítico de la asignatura exploradas en el examen.**
8. **Se cálcula la frecuencia observada a partir de los resultados reales del examen.**
9. **Para ello se utiliza el índice de dificultad y el criterio de expertos. Un ejemplo:**
   - Menos de 40% de dificultad, se consideran difíciles.
   - Entre 41 y 50% de dificultad, se consideran medianamente difíciles.
   - Entre 51 y 80% de dificultad, se consideran de dificultad media.
   - Entre 81 y 90% de dificultad, se consideran medianamente fáciles, y
   - Entre 91 y 100% de dificultad, se consideran fáciles.

Pasos:

10. **Determinación de la relación que tiene el fondo de tiempo de cada una de las temáticas en el plan calendario con el número de incisos de cada una de ellas presentes en el examen y su grado de dificultad.**
11. **Aplicación de las priebas de confiabilidad y discriminación.**
12. **Toma de decisiones y elaboración de definiciones y estrategias.**

IMPORTANCIA DEL ANÁLISIS DE LA CALIDAD DE LOS INSTRUMENTOS



- Informaciones sobre el rendimiento de los estudiantes que permiten focalizar la atención sobre posibles problemas del aprendizaje e incidir en su seguimiento a través de las comisiones verticales y horizontales.
- Datos cuantitativos que permiten la identificación de ítems con deficiencias técnicas y por tanto de exámenes con dificultades.
- Desarrolla la capacidad del docente para elaborar buenos ítems, al ejercitar la reestructuración de aquellos que son deficientes.