

TESTING-QUEST: AUDITORÍA DE TEST Y CUESTIONARIOS

Jordi Renom Pinsach
Pablo Sánchez Bermejo

Departamento de Psicología Social y Psicología Cuantitativa

TESTING-QUEST: AUDITORÍA DE TEST Y CUESTIONARIOS

Jordi Renom Pinsach
Pablo Sánchez Bermejo

Departamento de Psicología Social
y Psicología Cuantitativa

Índice

INTRODUCCIÓN	7
CAPÍTULO 1. PRIMEROS PASOS CON TESTING-QUEST	9
1.1. Acceso a la aplicación.....	9
1.2. Operativa básica.....	9
CAPÍTULO 2. LOS ARCHIVOS DE DATOS.....	13
2.1. Estructura y ejemplos de archivos	13
2.1.1. Ejemplo del archivo de datos BIO.....	14
2.1.2. Ejemplo del archivo de datos ALEC	15
2.1.3. Ejemplo del archivo de datos CAL.....	15
2.2. Ejemplos disponibles en la versión docente de Testing-Quest.....	16
CAPÍTULO 3. ANÁLISIS GLOBALES	17
3.1. Salud del test	17
3.1.1. Test Health para ítems de elección múltiple.....	17
3.1.2. Test Health para ítems binarios y graduados.....	19
3.1.3. Distribución de los problemas del test	19
3.2. Gráficos	20
3.2.1. Reliability.....	20
3.2.2. Items Map	21
3.2.2.1. Eje X para ítems dicotómicos	21
3.2.2.2. Eje X para ítems graduados.....	22
3.2.2.3. Eje Y para todo tipo de ítems	23
3.2.2.4. Interpretación por sectores del Mapa de Ítems	26
3.2.3. Puntuación directa versus puntuación ponderada	28
3.2.4. Puntuación directa versus omisión	29
3.2.5. Puntuación directa versus coherencia	33
3.2.6. Puntuación directa versus MCI	36
3.2.7. Distribución de puntuaciones.....	37
CAPÍTULO 4. ANÁLISIS DE LOS ÍTEMS, LA TABLA DE ÍTEMS	39
4.1. Indicadores para ítems de alternativa múltiple	39
4.2. Indicadores para ítems abiertos o binarios y graduados.....	41
4.3. Gráfico de frecuencias de respuesta del ítem	43
4.3.1. Estructura e interpretación del gráfico	44
4.4. Gráfico de discriminaciones	47
4.4.1. Estructura e interpretación del gráfico	48
4.5. Gráfico de perfiles de respuesta.....	49
4.5.1. Estructura e interpretación.....	49
4.5.1.1. Ítems dicotómicos binarios.....	49
4.5.1.2. Ítems dicotómicos con varias alternativas de respuesta.....	51
4.5.1.3. Ítems de puntuación graduada	55

CAPÍTULO 5. ANÁLISIS DE LOS SUJETOS. LA TABLA DE EXAMINANDOS	57
5.1. Indicadores de la tabla	57
CAPÍTULO 6. PROCESO DE ANÁLISIS DEL EJEMPLO CAL	59
6.1. Estructura de los análisis de ejemplo	59
6.2. El ejemplo CAL	59
6.3. Secuencia de análisis	60
6.4. Resumen	69
CAPÍTULO 7. PROCESO DE ANÁLISIS DEL EJEMPLO BRITISH	71
7.1. El ejemplo BRITISH	71
7.2. Secuencia de análisis	72
7.3. Resumen	84
CAPÍTULO 8. PROCESO DE ANÁLISIS DEL EJEMPLO BOSS	85
8.1. El ejemplo BOSS	85
8.2. Secuencia de análisis	86
8.3. Resumen	90
CAPÍTULO 9. PROCESO DE ANÁLISIS DEL EJEMPLO QUANT	91
9.1. El ejemplo QUANT	91
9.2. Secuencia de análisis	93
9.3. Resumen	101
CAPÍTULO 10. PROCESO DE ANÁLISIS DEL EJEMPLO MOOD	103
10.1. El ejemplo MOOD	103
10.2. Secuencia de análisis	104
10.3. Resumen	110
CAPÍTULO 11. PROCESO DE ANÁLISIS DEL EJEMPLO MC24	111
11.1. El ejemplo MC24	111
11.2. Secuencia de análisis	112
11.3. Resumen	115
CAPÍTULO 12. PROCESO DE ANÁLISIS DEL EJEMPLO BANK	117
12.1. El ejemplo BANK	117
12.2. Secuencia de análisis	119
12.3. Resumen	131
CAPÍTULO 13. SOLUCIONES DE LOS EJEMPLOS ANALIZADOS	133
13.1. Respuestas del ejemplo CAL	133
13.2. Respuestas del ejemplo BRITISH	140
13.3. Respuestas del ejemplo BOSS	150
13.4. Respuestas del ejemplo QUANT	154
13.5. Respuestas del ejemplo MOOD	161
13.6. Respuestas del ejemplo MC24	166
13.7. Respuestas del ejemplo BANK	168
GLOSARIO	179

INTRODUCCIÓN

Este manual está pensado para los usuarios del programa Testing-Quest con el objetivo de ofrecer una guía de trabajo para el análisis de las cualidades métricas de test y cuestionario mediante las rutinas de trabajo del programa.

Testing-Quest se ha diseñado como una herramienta profesional de análisis de instrumentos psicométricos, pero también tiene una finalidad formativa. Por ello, incorpora una versión de uso abierto que permite analizar libremente una serie de ejemplos de instrumentos representativos. Esta versión resulta de utilidad tanto para estudiantes como para docentes en actividades prácticas sobre evaluación educativa y psicológica mediante test, cuestionarios y exámenes en cursos de psicología, educación y ciencias sociales.

La versión formativa de Testing-Quest no precisa registro ni instalación. Es de acceso directo y manejo intuitivo desde cualquier dispositivo y sistema operativo conectado a la Red (www.testing-quest.com).

Los primeros cinco capítulos del manual describen las prestaciones del programa, así como los resultados proporcionados en un análisis estándar. Los siguientes ponen en práctica el proceso de análisis empleando los datos de test y cuestionarios de la lista de ejemplos que incluye la web de Testing-Quest.

Todos estos ejemplos corresponden a situaciones reales de diversa complejidad. Los primeros entran con más detalle en la secuencia de instrucciones para obtener los resultados. Más adelante, la navegación por la aplicación ya no es la prioridad. El objetivo pasa a ser la adquisición de una autonomía en la interpretación y la toma de decisiones. Con este fin, cada ejemplo plantea una serie de preguntas de respuesta abierta que guían el análisis y que hay que ir respondiendo en el propio manual. Para resolverlas es conveniente trabajar en un grupo reducido y disponer de documentación sobre los conceptos manejados. Como soporte complementario se incluye un glosario final con los principales conceptos relacionados con Testing-Quest.

En caso de duda, y como último recurso, el capítulo 14 proporciona todas las soluciones a las preguntas de los análisis de ejemplo.

Las rutinas de análisis con Testing-Quest no son exclusivas para usuarios del programa. El aprendizaje adquirido es extrapolable a otras aplicaciones psicométricas de acceso más complejo y restringido. Esto hace que sea especialmente idóneo para el ámbito docente como medio de aprendizaje en aulas de informática o como usuario individual.

PRIMEROS PASOS CON TESTING-QUEST

1.1. Acceso a la aplicación

Testing-Quest (T-Q) es una aplicación web de libre uso que se encuentra incorporada en la opción «Análisis» del menú principal de la página web www.testing-quest.com, que ofrece la posibilidad de analizar las cualidades psicométricas de un test.

T-Q es una derivación del programa Metrix Engine (1992, 2003), del que mantiene algunos aspectos básicos de funcionamiento. La versión docente aquí descrita permite analizar con libertad 25 ejemplos variados de test y cuestionarios. Para el análisis de otras pruebas se precisa un servicio profesional mediante consulta previa en la página web.

Para efectuar un análisis con T-Q solo se necesita un archivo con las respuestas de los examinandos en un test o cuestionario, tabuladas en una estructura determinada y en un formato de hoja de cálculo. Para la interpretación de los resultados, es conveniente, aunque no indispensable, disponer del original del test (texto o contenido de los ítems) a fin de asociar los resultados numéricos y gráficos con el contenido de las preguntas. T-Q admite test con ítems de puntuación dicotómica (binarios, alternativa múltiple, etc.) y también graduada (Rating Scales, tipo Likert).

1.2. Operativa básica

En la versión docente el análisis es automático. Después de seleccionar cualquiera de los ejemplos, tras breves segundos, es posible consultar los resultados. La operativa del programa es la misma en ambas versiones, tanto docente como profesional, con solo algunas limitaciones en la primera.

La figura 1 muestra la ventana con los 25 archivos ordenados por nombre. Su contenido (datos, estructura, etc.) no es accesible ni editable (no visible), pero se permite realizar análisis y ensayos en plenas condiciones. La columna «Type» indica la modalidad de ítem: «Multirrespuesta» equivale a alternativa múltiple y binarios; «Graduado», a escala graduada (tipo Likert). La columna «Info» indica las dimensiones (sujetos x ítems) de la matriz de datos del ejemplo.

Una vez cargado el archivo (al clicar sobre su nombre), aparece una ventana de resultados con un menú vertical lateral (izquierda), desde el que se accede a tres bloques de resultados. La figura 2 corresponde a los primeros resultados del ejemplo «QUANT».

Al cargar un archivo, es importante no clicar ninguna opción del menú hasta que el indicador superior de procesamiento desaparezca (☺). De lo contrario, el análisis puede quedar interrumpido. La calidad de la conexión a la Red también puede prolongar el tiempo de proceso, o incluso reportar algún error del servidor.

Testing-Quest		
Name	Type	Info
CAL	Multirespuesta	100x30
BRITISH	Multirespuesta	100x58
BOSS	Grado	100x20
BIO	Multirespuesta	657x15
BANK	Multirespuesta	100x100
AUDITCLIN	Multirespuesta	766x110
ALEC	Grado	100x50
ORT	Multirespuesta	66x30
NP6B	Multirespuesta	3462x30
MOOD	Grado	97x58
METODO	Multirespuesta	92x25
LCC	Multirespuesta	96x100
IEMOCION E	Multirespuesta	108x31
MC24	Grado	474x24
IEMOCION D	Multirespuesta	108x32
IEMOCION C	Multirespuesta	108x22
IEMOCION B	Multirespuesta	108x33
IEMOCION A	Multirespuesta	108x20
ICE	Grado	100x36
HLEC	Grado	HLEC
ESTR	Grado	100x50
DT8B	Multirespuesta	824x40
CP7D	Multirespuesta	4242x40
CP7C	Multirespuesta	4255x40
QUANT	Multirespuesta	700x29

Figura 1. Lista de matrices de ejemplos.

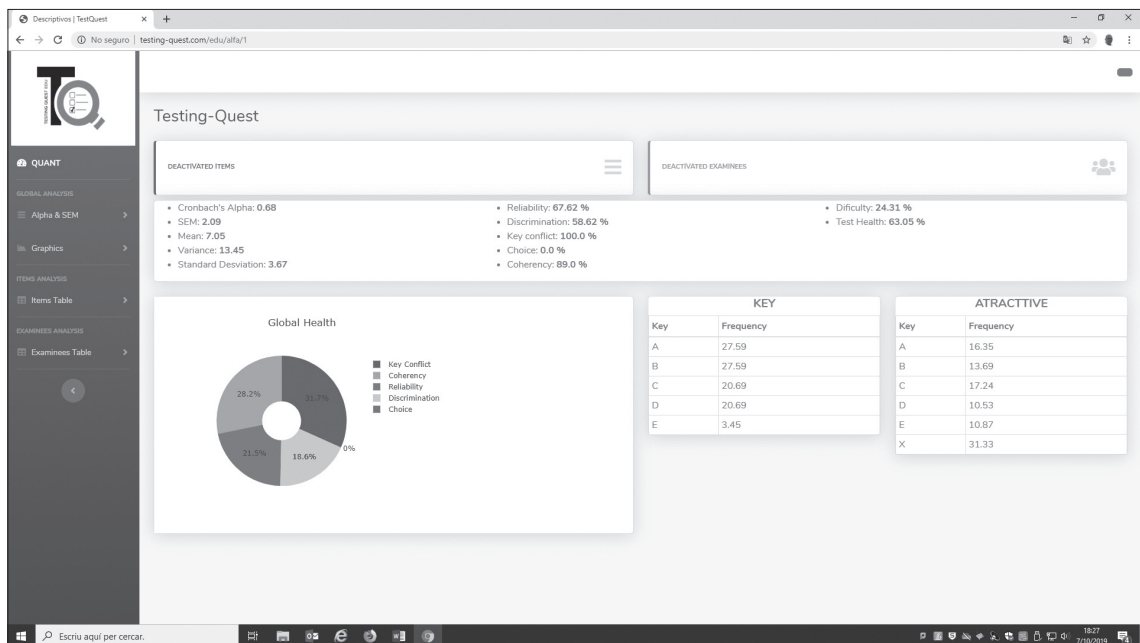


Figura 2. Ventana inicial de resultados para el ejemplo «QUANT».

Al margen de esta ventana general, el resto de resultados se agrupa en los siguientes tres bloques:

- «GLOBAL ANALYSIS»: incluye gráficos que informan del funcionamiento general del test.
- «ITEMS ANALYSIS»: concierne exclusivamente a los ítems y, si es el caso, a sus alternativas de respuesta. Constituye el núcleo principal del análisis, donde se verifica de manera numérica y gráfica la idoneidad de cada ítem, de cada una de sus alternativas, de las respuestas en blanco (omisiones) y de posibles conflictos entre alternativas (en caso de ítems de alternativa múltiple). Toda esta información se organiza en una tabla.

- «EXAMINEES ANALYSIS»: afecta a los examinandos. En general, se trata de otra tabla con puntuaciones e indicadores numéricos del conjunto de examinandos.

En cualquier momento de un análisis, se pueden hacer cambios y volver a iniciar el proceso. Durante el análisis de un test, el programa puede inducir a modificar la clave de corrección, a desactivar ítems y examinandos, a recodificar respuestas, etc. Se trata de un proceso cíclico de revisión del test original que puede repetirse varias veces hasta converger en una versión de la prueba que proporcione los mejores indicadores de salud.

Durante este proceso es recomendable contrastar los progresivos resultados gráficos y numéricos con el contenido (redactado) de los ítems del test. Todos los cambios derivados del análisis cuantitativo deberían justificarse con las respectivas interpretaciones cualitativas sobre el contenido de las preguntas afectadas. En todo proceso de análisis de test es importante hallar la causa del mal funcionamiento de estos ítems (detalles de la redacción, estructura, disposición, etc.). T-Q ayuda a identificar posibles anomalías del test, pero es importante confirmarlas mediante su revisión cualitativa.

LOS ARCHIVOS DE DATOS

2.1. Estructura y ejemplos de archivos

Aunque la versión docente de T-Q no permite el acceso a los datos, es importante conocer el formato con el que trabaja. Como aparece en la figura 3, los archivos analizados por T-Q deben adecuarse a una estructura de cuatro sectores o tipos de información. En todos los casos, tienen el formato de hoja de cálculo Excel o compatible.

- Sector A de datos personales: las columnas de la izquierda son opcionales y se destinan a recopilar datos de los examinandos. Tienen un título en el encabezamiento situado en la fila 4 o siguientes. Los datos deben ser numéricos, y siempre debe haber, como mínimo, una columna de este tipo. En caso de que no se disponga de estos datos, se ha de crear una columna con el número de orden de los sujetos.
- Sector B con datos de los ítems: las tres primeras filas superiores del archivo son obligatorias y recopilan tres datos de cada pregunta del test.
 - Fila 1: corresponde a la alternativa correcta de cada ítem (letras mayúsculas). En ítems de escala graduada, se refleja su polaridad positiva (+) o negativa (-).
 - Fila 2: indica a qué escala o subtest pertenece cada ítem (1, 2, 3, etc.). En la versión formativa del programa, todos los ítems pertenecen a la misma y única escala, codificada siempre como 1.
 - Fila 3: indica el número de alternativas de respuesta o, si son ítems graduados, los grados de puntuación de cada pregunta. En esta versión, como todos los ítems pertenecen a una sola prueba, comparten el mismo número de alternativas o grados. Si no existen datos del sector C, esta fila 3 coincidirá (a la izquierda) con los títulos o etiquetas de las columnas de datos personales del sector A.
- Sector C con datos descriptivos de los ítems: la fila 4 y las siguientes bajo el bloque B son opcionales y se destinan a datos complementarios que describen cada ítem. Son datos externos a la prueba. Pueden ser valoraciones codificadas numéricamente sobre la dificultad de cada ítem (estimada por el autor), referirse al contenido (categoría, tema, etc.), si se ha empleado antes en otros test, etc. Tras el análisis, estos datos se pueden comparar con los proporcionados por el programa para cada ítem. Es recomendable, como mínimo, que haya una fila con este tipo de datos (será la cuarta línea). En este caso, a la izquierda de esta fila 4, se encontrarán los títulos o etiquetas de las variables personales de las columnas del sector A.
- Sector D con las respuestas de los examinandos: cada celda de una fila recopila la respuesta a un ítem (columna) de un examinando. Las respuestas pueden ser de diferentes tipos.
 - Ítems de dos o más alternativas con una sola correcta o que puntúa: se tabularán como letras mayúsculas (A, B, C, etc.).
 - Ítems de puntuación graduada: se tabulan números mediante números (1, 2, 3, ...), que corresponden a la respuesta o grado marcado por el examinando. No es necesario re-

codificar (invertir) los valores en los ítems negativos. La aplicación lo hará durante el análisis.

- Ítems de respuesta abierta, sin alternativas, valorados como acierto o error o como respuesta que puntúa y que no puntúa: se tratan como ítems de dos alternativas, de manera que A es el acierto (o respuesta que puntúa) y B, el error (o respuesta que no puntúa). En este caso, la primera fila del bloque B tendrá solo letras A. No hay que confundir esta codificación con la de ítems de dos alternativas. En los abiertos, en la primera fila del sector B siempre habrá letras A, mientras que en pruebas de dos alternativas (Verdadero-Falso), puede haber letras A o B según la que puntúe en cada ítem.
- Ítems de respuesta abierta, sin alternativas, valorados de manera graduada: un ejemplo son los exámenes de respuesta en que se puntúa de manera graduada. En estos casos, se tabulan como si fueran ítems graduados (1, 2, 3, etc.) y conservan el signo + en la primera fila del sector B. En este tipo de pruebas no tiene sentido plantear ítems negativos.

En todos los archivos, la respuesta en blanco (omisión) y la doble marca se codifican como X.

Las columnas de datos B, C y D deben encajar.

	línea 1	DATOS B
	línea 2	
	línea 3	
	VDI opcional	DATOS C
DATOS A (mínimo 1)	DATOS D Multi/Binarios: letras Graduados: números	

Figura 3. Estructura completa de un archivo de datos.

2.1.1. Ejemplo del archivo de datos BIO

BIO es un archivo con las respuestas de 657 sujetos a una prueba de 15 ítems con 5 alternativas de respuesta sobre conocimientos en biotecnología. Incluye 2 columnas con datos tipo A (sobre implicación política y religiosa de cada sujeto), 3 filas con datos tipo B, una sola fila de datos C (código de quién es el autor 1, 2, 3 de cada pregunta) y, por último, datos D con las respuestas de los examinandos.

3.1. Salud del test

El primer bloque de resultados incluye 7 ventanas, una con indicadores numéricos y otras 6, gráficas. Todos estos resultados son generales, ya que no afectan a un ítem o sujeto concreto, sino al conjunto de respuestas de todos los examinandos a todos los ítems.

3.1.1. Test Health para ítems de elección múltiple

La primera ventana, «Test Health», ofrece una visión general del funcionamiento del test. La información se recopila en diferentes recuadros, 4 en el caso de pruebas de alternativa múltiple y 2 para el resto del test. La figura 7 muestra los 4 recuadros de la ventana de salud del ejemplo BIO (alternativa múltiple).

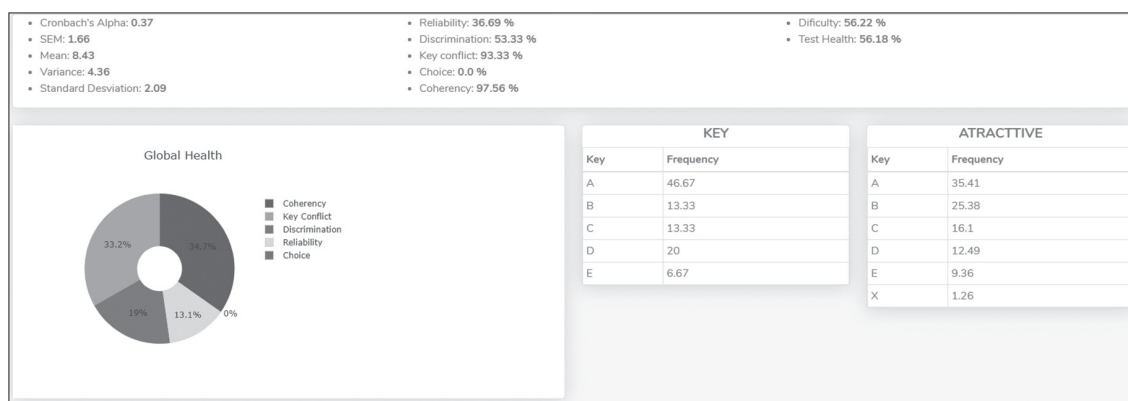


Figura 7. Indicadores de salud del ejemplo BIO.

El recuadro superior informa de los siguientes 12 indicadores:

- «Cronbach's Alpha»: coeficiente de fiabilidad conocido como de Consistencia Interna, que oscila entre 0 y 1; es recomendable aceptar valores superiores a 0,8. Este coeficiente es una estimación (aproximación) del coeficiente de fiabilidad de equivalencia en el caso de que existiese una forma paralela del test analizado. Por ello, el coeficiente Alpha es uno de los más empleados cuando no se dispone de una forma paralela o de un re-test del test analizado.
- «SEM»: indica la imprecisión de los resultados del test. Es un dato más concreto que el coeficiente Alpha, puesto que permite estimar el margen de error de la puntuación obtenida por cualquier examinando. Como regla básica, basta con sumar y restar dos veces (redondeo de $z = 1,96$) el valor de SEM ($\pm 2SEM$) a la puntuación del examinando. La puntuación verdadera de ese individuo se encontrará con un nivel de confianza del 95 %, es decir, en algún punto del intervalo generado por $\pm 2SEM$. Cuanto mayor sea este intervalo, menos preciso será el test. Por tanto, cuando se hallan valores altos de Alpha, los de SEM se reducen.

Alpha y SEM son de ayuda cuando hay que tomar decisiones vinculantes sobre la puntuación del examinando (apto-no apto, se ajusta-no se ajusta a un perfil, diagnóstico de un posible riesgo, etc.). Por ejemplo, una puntuación de 48 en un test de 100 ítems, que aparentemente queda por debajo de un valor de corte 50 (decisión de apto/no apto), se interpretará de un modo diferente si se considera como muy exacta (de entrada sería un no apto) o si se tiene en cuenta SEM. En este segundo caso, un valor de SEM = 4 puede llevar a que la amplitud del intervalo de imprecisión haga considerar la posibilidad de que la puntuación del examinando sea superior ($48 + 2SEM = 56$, que podría ser apto). Lo mismo puede producirse a la inversa ($48 - 2SEM = 40$). Con un nivel de confianza del 95 %, la verdadera puntuación de este examinando se encontraría entre 40 y 56 puntos.

En el ejemplo «BIO», el valor de Alpha (0,37) está muy por debajo de lo aceptable. En cuanto a SEM ($1,66 \times 2 = 3,32$), implica un intervalo de imprecisión de $\pm 3,32$, que equivale a 6,64 puntos sobre el rango total de 0 a 15 puntos que puede conseguir un examinando en esta prueba. Este intervalo equivale casi al 50 % del rango, lo cual hace que sea muy cuestionable la utilidad de este test.

- «Mean»: promedio de las puntuaciones obtenidas por todos los sujetos en el test.
- «Variance»: varianza de las puntuaciones conseguidas por todos los sujetos en el test. Es un indicador de variabilidad y dispersión de las puntuaciones.
- «Standard Desviation»: desviación estándar de las puntuaciones obtenidas por todos los sujetos en el test. Es la raíz cuadrada de la varianza.

Los siete indicadores restantes sirven para calcular el nivel de salud de la prueba. Los valores se calculan a partir de datos de la tabla de ítems (capítulo 4).

- «Reliability»: el coeficiente Alpha se convierte en un valor de 0 a 100.
- «Discrimination»: porcentaje de ítems que discriminan por encima de 0,3.
- «Key conflict»: porcentaje de ítems sin aviso de problemas en su clave.
- «Choice»: porcentaje de ítems con homogeneidad de atractivo entre sus alternativas erróneas (distractores).
- «Coherency»: porcentaje de sujetos con el índice MCI (Modified Caution Index, o Índice de Precaución Modificado) superior a 0,5. El índice MCI de un examinando informa de la coherencia de sus respuestas. Al margen de si un individuo obtiene una puntuación alta o baja, se espera que esta provenga de un patrón coherente. Por tanto, los valores de MCI más interesantes son los cercanos a 0.
- «Difficulty»: porcentaje general de aciertos de todos los ítems del test. Es un dato que complementa los anteriores, pero que no interviene en el cálculo de salud.
- «Test Health»: porcentaje medio o global de los cinco indicadores anteriores. Expresa la salud global del test.

El recuadro «KEY» muestra los porcentajes de veces en que las alternativas A, B, C, etc. son ciertas según la clave de corrección. Los porcentajes se calculan sobre el total de ítems del test. Estos valores permiten comprobar si la clave de aciertos del test se ha distribuido uniformemente entre las alternativas. Cuando una o más alternativas predominan sobre las otras, puede causar un aumento de la probabilidad de acierto de los ítems, ya que las alternativas que entran en juego se reducen.

El recuadro «ATTRACTIVE» informa de cómo se reparten los porcentajes de elecciones de los examinandos entre las alternativas A, B, C, etc. en el total del test. El conjunto de estos datos indica si existe alguna tendencia a la hora de escoger las alternativas del test. Como norma general, tanto estos valores como los de «KEY» deberían distribuirse uniformemente en sus respecti-

vas listas. Al inicio de un análisis, los datos de «KEY» y «ATTRACTIVE» ayudan a explicar posibles efectos en las respuestas al test, que luego se manifiestan en ítems concretos.

3.1.2. Test Health para ítems binarios y graduados

Para test de ítems binarios, sin alternativas, los indicadores de esta ventana se reducen. No se calculan «Key Conflict», «Choice» ni los recuadros inferiores que afectan solo a test con alternativas.

Por último, en pruebas con ítems de escala graduada, los indicadores se limitan a los siguientes:

- «Score»: media global de las respuestas, expresada en el rango de la escala de respuesta (de 1 a 4, de 1 a 5, etc.).
- «Reliability»: Alpha se convierte en un valor de 0 a 100.
- «Discrimination»: porcentaje de ítems que discriminan por encima de 0,3.
- «Variability»: porcentaje de ítems del test con desviación típica superior a un valor mínimo que expresa escasa variabilidad de respuestas. Este valor mínimo varía según el número de grados.

Tabla 1. Varianzas mínimas en función del número de grados de respuesta.

Grados	Mínimo
3	0,3
4	0,4
5	0,5
6	0,6
7	0,7
8	0,8
9	0,9
10	1
11	1,1

3.1.3. Distribución de los problemas del test

Independientemente del tipo de ítems en todas las ventanas de salud, existe el recuadro «TEST HEALTH», donde aparece un círculo dividido en sectores que resumen la distribución de los problemas del test.

Cada sector indica el valor relativo de cada problema en relación con el conjunto de problemas que presenta la prueba. La interpretación del círculo siempre es relativa. Ya sea un test con muchos problemas o con pocos, el círculo expresa cómo se distribuyen estos respecto al total de problemas. Por tanto, un test con pocos problemas puede mostrar un círculo similar a otro con muchos problemas siempre y cuando su reparto sea similar. La cantidad de problemas aparece en el recuadro superior de la ventana, mientras que el círculo señala cuáles pueden ser las acciones más directas para mejorar el test.