Capítulo 18

UNA INTRODUCCIÓN A LA TEORÍA DE LA GENERALIZABILIDAD

Iwin Leenen, Olivia Espinosa Vázquez

"No he fracasado, he encontrado 10,000 formas que no funcionan."
Thomas A. Edison

INTRODUCCIÓN

Como se señaló en el capítulo 16 de este libro, la teoría clásica de los tests (TCT) se fundamenta en una base axiomática comprendida en la ecuación $X = T + E$, donde X y T son variables que representan a la puntuación observada y la puntuación verdadera, respectivamente, asociadas con una medición particular considerada para una población de sustentantes. Por otro lado, E representa el error asociado con la medición y técnicamente "explica" por qué la puntuación observada de un sustentante no coincide perfectamente con su puntuación verdadera. (Nótese que de la ecuación básica resulta que $E = X - T$.) A partir del supuesto de que la correlación entre T y E es igual a 0 (o, más general, que el error de medición es independiente de la puntuación verdadera), se deriva de la ecuación básica el siguiente resultado fundamental:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2. \tag{1}$$

En palabras, la varianza observada (es decir, las diferencias entre sustentantes que la medición revela) se parte en dos: varianza verdadera (diferencias entre sustentantes que se deben a diferencias en sus puntuaciones verdaderas) y varianza error (diferencias atribuibles a factores que perturban la medición de una manera no sistemática). La Ecuación 1 directamente lleva a la definición de confiabilidad en el marco de la TCT:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2}\,, \tag{2}$$

es decir, la confiabilidad es la proporción de la varianza observada atribuible a la puntuación verdadera.

Desde su concepción en la primera mitad del siglo XX, la TCT ha permanecido como el modelo más utilizado para construir y evaluar instrumentos psicológicos. En este capítulo presentamos una extensión a la TCT, que se conoce como la Teoría de la Generalizabilidad (Teoría G), que tiene sus origines en las publicaciones de Lee Cronbach y colegas (1963, 1972). La Teoría G se diferencia de la TCT en el sentido de que distingue entre múltiples fuentes de error y, de ahí, por una manera diferente (a la presentada en la Ecuación 1) de particionar la varianza observada. Un objetivo esencial de la Teoría G consiste precisamente en estimar la contribución de las distintas fuentes que influyen en la medición, es decir, la importancia que tienen en la varianza de las puntuaciones observadas.

La estructura del capítulo es como sigue: en la siguiente sección introducimos dos casos que representan, cada uno, una evaluación típica en un contexto universitario y que utilizaremos como guía para ilustrar los conceptos que se desarrollan en el resto de este capítulo. Después, presentamos el marco conceptual de la Teoría G, con sus definiciones más importantes. Puesto que la manera en que se concreta la Teoría G depende de las características y el diseño de la evaluación específica a la que se aplica, las siguientes secciones elaboran la Teoría G en el contexto de los dos casos introducidos en la siguiente sección, que llevan a desarrollos relativamente sencillos en el marco de la Teoría G. En la sección subsiguiente describimos de una manera muy general aspectos de diseños avanzados, más complejos. Concluimos el capítulo con algunos comentarios sobre relaciones entre la Teoría G y otros enfoques psicométricos.

EJEMPLOS DE GUÍA

A continuación, describimos dos ejemplos de exámenes típicos en un contexto universitario. El examen A evalúa a los estudiantes de la materia de Cálculo Diferencial e Integral de la carrera de Matemáticas. Participan 68 estudiantes en el examen, el cual consiste en 20 preguntas que son integrales definidas que el sustentante tiene que resolver. Cada pregunta se califica únicamente considerando los puntajes de 0 y 1 (respuesta incorrecta y correcta, respectivamente). La calificación en el examen es la proporción (o el porcentaje) de las 20 integrales que el sustentante resuelve correctamente. Los datos de este examen se pueden estructurar como se muestra en la parte superior de la Tabla 1.

Tabla 1. Estructura de los datos observados para los Exámenes A y B

Examen A

		Preguntas (*i*)								
		1	2	3	4	5	⋯	19	20	$\bar{X}_{pI}$
Sustentantes (*p*)	1	0	1	1	0	0	⋯	1	1	0.55
	2	1	1	1	1	0	⋯	1	1	0.80
	3	0	1	0	0	0	⋯	1	1	0.70
	4	1	1	1	1	0	⋯	1	1	0.80
	5	0	1	1	1	0	⋯	1	1	0.85
	⋯	⋯	⋯	⋯	⋯	⋯	⋱	⋯	⋯	⋯
	67	1	1	1	1	1	⋯	1	1	1.00
	68	0	0	0	0	0	⋯	0	1	0.25
	$\bar{X}_{Pi}$	0.71	0.91	0.74	0.72	0.31	⋯	0.91	0.96	0.74

Nota: Los datos completos se encuentran en el archivo suplementario Examen A. csv. Los símbolos $\bar{X}_{pI}$ y $\bar{X}_{Pi}$ representan la media de la fila (del sustentante *p*) y de la columna (de la pregunta *i*), respectivamente.

Examen B

	Evaluadores (*j*)													
	Evaluador 1: Preguntas (*i*)							Evaluador 2: Preguntas (*i*)						
	1	2	3	4	5	6	1	2	3	4	5	6	$\bar{X}_{pIJ}$
Sustentantes (*p*) 1	7	8	7	6	8	7	6	7	6	7	7	8	7.0
2	10	9	6	7	7	7	7	7	8	8	8	8	7.7
3	9	9	8	8	8	8	7	8	7	9	6	8	7.9
4	9	10	9	8	8	10	9	7	9	10	9	10	8.9
5	9	9	9	8	8	8	8	7	8	9	9	9	8.4
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
30	8	10	8	8	10	9	8	9	8	9	8	8	8.6
31	9	10	9	8	8	9	9	9	7	10	9	9	8.8
$\bar{X}_{Pij}$	8.5	6.9	7.6	7.3	7.5	8.1	8.9	7.1	8.2	7.2	8.1	8.4	7.8

Nota: Los datos completos se encuentran en el archivo suplementario Examen B.csv. Los símbolos $\bar{X}_{pIJ}$ y $\bar{X}_{Pij}$ representan la media de la fila (del sustentante *p*) y de la columna (de la pregunta *i* y el evaluador *j*), respectivamente.

El Examen B es una evaluación que aplica un profesor de la licenciatura en Filosofía a los 31 estudiantes de su grupo de la asignatura de Ética. Este examen consiste en seis preguntas que

requieren que los sustentantes reflexionen brevemente sobre ciertos temas (por ejemplo, con una crítica sobre cierta teoría o una comparación de diferentes perspectivas éticas). Tienen que responder a cada pregunta con un breve ensayo (de una página aproximadamente). Para asignar una calificación a las respuestas en cada pregunta, el profesor pide la ayuda a dos de sus estudiantes de doctorado (que trabajan con él como ayudantes de profesor); después de revisar junto con ellos los criterios para calificar las respuestas, les pide que ambos valoren, de manera independiente y utilizando una escala de 0 a 10, la respuesta en cada una de las seis preguntas de cada estudiante. (El profesor es consciente de que el juicio sobre la calidad de la reflexión que ha mostrado el sustentante en sus respuestas del examen hasta cierto punto es subjetivo; esta es la razón por la que pide una calificación independiente por sus ayudantes.) La calificación en el examen es el promedio en los 12 puntajes que se tiene para el sustentante (seis puntajes del primer evaluador más seis puntajes del segundo). Los datos resultantes se estructuran como se muestra en la parte inferior de la Tabla 1.

CONCEPTOS CENTRALES DE LA TEORÍA DE LA GENERALIZABILIDAD

La Teoría G consiste en un marco de referencia conceptual y una metodología que permite al investigador desentrañar múltiples fuentes de error y estimar su contribución relativa a las mediciones realizadas. Los orígenes de esta teoría se encuentran en la TCT y en el análisis de la varianza (ANOVA, por sus siglas en inglés).

En la Teoría G se describen las mediciones que se realizan a las personas en los términos de las condiciones bajo las cuales se han observado. Estas condiciones forman las *facetas*. En el Examen A, las mediciones son los puntajes de los sustentantes en las 20 preguntas del examen; estas 20 preguntas son las 20 *condiciones* de la *faceta* "preguntas" (la cual es la única faceta considerada para este examen). El Examen B, en el cual se observan los puntajes otorgados a las respuestas de los 31 individuos por cada uno de los dos evaluadores en las seis preguntas del examen, presenta un diseño con dos facetas: la faceta "preguntas" (con seis condiciones), y la faceta "evaluadores" (con dos condiciones). En ambos ejemplos, los sustentantes en el examen son el *objeto de la medición* (queremos conocer de cada sustentante su nivel en el constructo evaluado por el examen) y no se consideran una faceta; este término se reserva para referirse a una fuente de error, es decir, a un factor que causa variabilidad en las distintas mediciones realizadas a un sustentante. En este sentido, las preguntas del examen son una faceta, por ejemplo, porque difieren en dificultad y, así, son una causa de la variabilidad en los puntajes en las distintas preguntas. De igual manera, los dos evaluadores en el Examen B definen una faceta, ya que pueden diferir en su juicio sobre las respuestas de los sustentantes, lo cual lleva a diferencias en los puntajes observados.

No solo el número de facetas consideradas define el diseño, sino también la manera en que se combinan (entre sí y con las personas). En el Examen A, los 68 sustentantes responden a las mismas 20 preguntas, lo cual se conoce como un *diseño cruzado,* denotado como "personas×preguntas". Al contrario, un *diseño anidado* se presentaría cuando cada sustentante en el examen respondiera a un conjunto de preguntas únicas, es decir, distintas de las

preguntas de los demás. Un ejemplo de este tipo de diseño ocurre en los exámenes orales, donde los sustentantes realizan su examen por turnos y cada sustentante recibe, por ejemplo, tres preguntas, diferentes de las preguntas de los demás (para evitar que los últimos sustentantes del examen aprendan las preguntas de los que pasaron primero). Este diseño se denota como "reactivos : personas" ("reactivos anidados en personas"). En diseños con dos o más facetas, ciertas facetas pueden estar cruzadas y otras anidadas. En el Examen B, por ejemplo, tenemos un diseño de dos facetas totalmente cruzadas: para cada sustentante, sus respuestas en cada una de las seis preguntas son calificadas por cada uno de los dos ayudantes del profesor titular ("personas×preguntas×evaluadores"), llevando a 12 mediciones para cada sustentante. Si, al contrario, el profesor titular hubiera decidido que su primer ayudante evaluara las preguntas 1, 2 y 3 de cada sustentante y el segundo las preguntas 4, 5 y 6, entonces tendríamos solo seis observaciones por persona y un diseño con las preguntas anidadas en la faceta de evaluadores, que se denota como "personas × (preguntas : evaluadores)".

Los lectores familiarizados con ANOVA se habrán dado cuenta de las similitudes en la terminología que se utiliza para esta técnica y la que acabamos de definir para la Teoría G. Por ejemplo, también ANOVA distingue entre diseños cruzados y anidados; por otro lado, "factores" (o "variables independientes") en ANOVA son conceptualmente similares a las "facetas" en la Teoría G; y los niveles de cada factor en ANOVA corresponden a las condiciones de las facetas en la Teoría G. Veremos en las secciones que vienen que un análisis en el marco de la Teoría G implica un ANOVA a los datos observados. Cabe mencionar que, para comprender a cabalidad los análisis en las siguientes secciones, sirve tener una introducción a ANOVA, como se encuentra, por ejemplo, en los libros de Pardo y San Martín (2010) y Tejedor (2019).

En el caso del Examen A observamos, para cada sustentante, sus respuestas en las 20 preguntas que conforman el examen. No se duda de que el profesor podría haber utilizado otras 20 integrales como preguntas del examen. El conjunto de todas las posibles observaciones admisibles –en este ejemplo, los puntajes en todas las preguntas que el profesor considera aceptables para incluir en el examen– se llama en la Teoría G el *universo*. Idealmente, para determinar la habilidad matemática de un sustentante en el ejemplo del Examen A se observarían sus respuestas (y puntajes) en todas las preguntas del universo y se calcularía, en este universo de preguntas, su calificación (porcentaje de respuestas correctas). Esta calificación es la *puntuación universo* del sustentante. Es obvio que, por razones prácticas, solo se pueden observar las respuestas en una *muestra* de preguntas y utilizamos, para conocer la habilidad matemática de los sustentantes, su *puntuación observada* en esta muestra de preguntas. Esto significa que hacemos una generalización a partir de la puntuación observada para conocer la puntuación universo. La precisión con la que las puntuaciones observadas pueden generalizarse a la puntuación universo se llama la generalizabilidad y, como se ilustrará en las siguientes secciones, se cuantifica a través del *coeficiente de generalizabilidad.*

Si se aplican las ideas anteriores al Examen B, el universo se define igualmente como todas las observaciones admisibles para el grupo de sustentantes considerado. En este caso, el universo se define por dos facetas: no solo consideramos todas las posibles preguntas que

según el profesor titular son adecuadas para incluirse en el examen, sino también a todos los posibles evaluadores que podría haber involucrado para calificar las respuestas de los sustentantes. Las puntuaciones universo, en este caso, son las puntuaciones que se habrían obtenido si todos los evaluadores admisibles hubieran calificado las respuestas de los sustentantes en todas las preguntas admisibles para el examen. Como esto no es viable en la práctica, se hace una generalización a partir de la puntuación observada que se calcula con base en la muestra de 12 mediciones disponibles para cada sustentante.

Un análisis en el marco de la Teoría G típicamente consiste en dos estudios. En primera instancia, se lleva a cabo el *estudio de generalizabilidad* (Estudio G), el cual estima la contribución a la varianza en las mediciones observadas de (a) las habilidades evaluadas de los sustentantes, junto con (b) las distintas facetas (fuentes de error) consideradas. Posteriormente, se utilizan los resultados del Estudio G en el *estudio de decisión* (Estudio D) para diseñar de la mejor manera posible un examen con un propósito particular, o, de manera más general, un instrumento de medición. El Estudio D evalúa diferentes diseños para un examen a la luz de los objetivos, las limitaciones prácticas y las decisiones precisas que se quieren tomar a partir de los puntajes observados en el examen, minimizando de esta manera el error de medición y maximizando la confiabilidad del examen. En las siguientes secciones continuamos con los ejemplos guía (diseños con una faceta y dos facetas, respectivamente) e ilustramos cómo se llevan a cabo los Estudios G y D, así como los resultados que generan.

DISEÑO CON UNA FACETA

En esta sección elaboramos el ejemplo del Examen A, en el cual figura una faceta: la de las preguntas. Primero elaboramos el modelo teórico (explicando cómo se particiona la puntuación y la varianza observada en diferentes componentes) y posteriormente, con base en este modelo, realizamos e ilustramos los Estudios G y D para este ejemplo.

Descomposición de la Puntuación Observada

La medición que se realiza a la persona p a través de la pregunta i la representamos por X_{pi}. En la frase anterior, p e i se refieren, de una manera general, a cualquier persona de la población de personas que se consideren como posibles participantes al examen y a cualquier pregunta del universo de preguntas que sean admisibles para este examen, respectivamente. Para un diseño cruzado con una faceta el puntaje X_{pi} se descompone en cuatro componentes de la siguiente manera:

$$X_{pi} = \mu + \alpha_p + \beta_i + \varepsilon_{p\times i,e}, \tag{3}$$

$$\text{donde } \alpha_p \equiv \mu_p - \mu, \tag{3a}$$

$$\beta_i \equiv \mu_i - \mu, \tag{3b}$$

$$\varepsilon_{p\times i,e} \equiv X_{pi} - \mu_p - \mu_i + \mu. \tag{3c}$$

El primero de estos componentes, μ, representa a la media global de todas las mediciones que resultan de aplicar todas las preguntas del universo a todas las personas de la población. En la Ecuación 3, μ recoge todas las influencias constantes, es decir, el efecto de todos los factores que siempre tienen la misma influencia en la medición X_{pi}, independientemente de la persona o de las preguntas o de otros factores con un efecto variable (como el error de medición). Por ejemplo, considerando que el profesor es el único que califica las respuestas al Examen A y suponiendo que en cada medición (de cualquier persona y de cualquier pregunta) aplica los mismos criterios con la misma rigidez, el profesor es uno de los efectos que contribuyen al efecto global μ.

Con respecto al segundo componente, α_p, la Ecuación 3a lo define como la diferencia entre μ_p y μ, donde μ_p es el puntaje universo de la persona p; en otras palabras, μ_p es el promedio de los puntajes que obtendría esta persona si le administráramos todas las preguntas del universo. Por consiguiente, α_p, que representa el efecto de la persona p, indica en qué medida el puntaje universo de esta persona se encuentra arriba de la media global (si es positivo) o abajo (si es negativo). Nótese que, por las definiciones anteriores, el promedio de los efectos α_p de todas las personas de la población es igual a 0 (ya que los efectos positivos y negativos contrarrestan).

De manera análoga, se define el tercer componente, β_i, como la diferencia entre μ_i y μ, donde μ_i es el promedio de los puntajes en la pregunta i (o bien, en caso de preguntas dicotómicas como en el Examen A, el porcentaje de personas que la acertarían) si se administrara a todas las personas de la población. De esta manera, μ_i se puede interpretar como la dificultad (o, más bien, la facilidad) de la pregunta i: preguntas con la μ_i más alta son más fáciles; si μ_i es más baja, la pregunta es más difícil. El efecto de la pregunta i, β_i, indica que esta pregunta es más fácil (si es positivo) o más difícil (si es negativo) que la dificultad promedio de las preguntas. En promedio, los efectos β_i de todas las preguntas en el universo es igual a 0.

Por último, el cuarto componente en la Ecuación 3, $\varepsilon_{p\times i,e}$, recoge el efecto residual; es decir, el efecto de todos los factores que influyen en la medición X_{pi}, excepto el efecto global μ, el de la persona α_p, y el de la pregunta β_i. Además del efecto del error de medición (es decir, los factores con una influencia no sistemática que no se contemplaron explícitamente en el diseño), el efecto residual incluye también el efecto de la interacción entre la persona y la pregunta. Es la razón por la que, para este diseño cruzado con una faceta, se incluyen como subíndices de ε, tanto $p\times i$ (la interacción) como e (el error de medición). De la definición en la Ecuación 3c sigue que, para cualquier pregunta i, el promedio de los efectos residuales $\varepsilon_{p\times i,e}$ entre todas las personas de la población es 0, así como que el promedio de los $\varepsilon_{p\times i,e}$ entre todas las preguntas del universo es 0 para cualquier persona p.

Descomposición de la Varianza Observada

Como acabamos de explicar, el primer componente en la Ecuación 3 es constante, igual para todas las mediciones en el universo de preguntas y la población de personas. Por otro lado, el componente α_p varía: tiene un valor distinto para diferentes personas. Esto quiere decir que podemos definir la varianza σ^2_α, la cual representa la variabilidad con respecto al com-

ponente α_p en la población de personas. Una interpretación alternativa de σ_α^2 es que corresponde con la variabilidad de los puntajes universo entre las personas de la población. De forma similar, el componente β_i varía entre las distintas preguntas del universo y esta variabilidad se cuantifica en la varianza σ_β^2. Se puede interpretar esta varianza como la variabilidad en la dificultad entre todas las preguntas del universo. Finalmente, se define la varianza σ_ε^2, que representa la variabilidad residual (por la variabilidad en el error de medición y el efecto de interacción, considerando las distintas combinaciones de persona-pregunta en la población de personas y el universo de preguntas). Se puede derivar del modelo introducido en la Ecuación 3 que

$$\sigma_X^2 = \sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\varepsilon^2. \tag{4}$$

En palabras, la varianza en la colección de todas las mediciones (de todas las personas de la población a quienes se administraron todas las preguntas del universo) se particiona en tres: varianza que se debe a las personas (porque difieren en su nivel general, representado por su puntaje universo), varianza debida a las preguntas (porque difieren en su grado de dificultad) y varianza debida a otros factores (incluyendo la interacción entre personas y preguntas). Las tres varianzas se llaman componentes de varianza en la colección de mediciones X_{pi} consideradas en este diseño.

Estudio de Generalizabilidad

Debe ser claro que las Ecuaciones 3 y 4 son parte de un modelo teórico y que los componentes de la puntuación y varianza observada (μ, μ_p, α_p, σ_α^2, etc.) son cantidades teóricas que en la práctica son desconocidas. Con base en los datos observados, como los que se muestran en la Tabla 1, podemos obtener estimaciones de estos componentes. En particular, el Estudio G se centra en la estimación de los componentes de varianza de la Ecuación 4.

El primer paso del Estudio G consiste en realizar un ANOVA con base en el modelo correspondiente al diseño del estudio (que se describió en las subsecciones anteriores). La Tabla 2 resume los resultados principales del ANOVA realizado a los datos del Examen A que se necesitan para estimar los componentes de varianza. En particular, las primeras cuatro columnas presentan los resultados que cualquier programa de análisis estadístico generaría para el ANOVA a estos datos. (El archivo suplementario Script_Análisis.R realiza los análisis en el entorno del software R). La última columna contiene las estimaciones de los componentes de varianza, las cuales se derivan a partir de las medias cuadráticas (MS, por sus siglas en inglés) mostradas en la penúltima columna, aplicando las siguientes fórmulas:

$$\hat{\sigma}_{\alpha}^{2} = \frac{MS_{\alpha} - MS_{\varepsilon}}{n_i} = \frac{1.0142 - 0.1259}{20} = 0.0444 \quad (5a)$$

$$\hat{\sigma}_{\beta}^{2} = \frac{MS_{\beta} - MS_{\varepsilon}}{n_p} = \frac{2.2373 - 0.1259}{68} = 0.0310 \quad (5b)$$

$$\hat{\sigma}_{\varepsilon}^{2} = MS_{\varepsilon} = 0.1259 \quad (5c)$$

En estas fórmulas, n_i y n_p son el número de preguntas y el número de personas, respectivamente, en los datos que se utilizan para la estimación. El símbolo ^ encima de los componentes de varianza indica que se trata de estimaciones (que se distinguen de los valores teóricos de la Ecuación 4).

Puesto que (las estimaciones para) los componentes de varianza dependen del esquema de calificación que se utilizó para las preguntas (en este caso, puntajes dicotómicos, 0 y 1), los valores obtenidos son difíciles de interpretar. Es la razón por la que se calcula el porcentaje de varianza que cada componente contribuye a la varianza (total) observada σ_X^2. (Estos porcentajes se muestran también en la última columna de la Tabla 2.) Por ejemplo, para la varianza σ_α^2, la cual es la varianza de los puntajes universo de las personas, se estima que contribuye 0.0444 / (0.0444 + 0.0310 + 0.1259) ≈ .221 (es decir, el 22.1%) a la varianza total en las mediciones (los puntajes 0 y 1, considerando el universo de preguntas y la población de personas). Los resultados del Estudio G enseñan que la contribución de la varianza residual (62.5%) es relativamente grande en comparación con la contribución de la varianza entre personas (22.1%) y entre preguntas (15.4%).

Estudio de Decisión

Hasta este momento, el modelo y análisis del Examen A se ha enfocado en la descomposición de la medición X_{pi}; es decir, el puntaje de la persona en una pregunta individual del universo. Sin embargo, el profesor no toma decisiones sobre sus estudiantes con base en sus puntajes en una pregunta, sino en sus promedios (o porcentajes de respuestas correctas) de todas las preguntas que conforman el examen. Este promedio de una persona p en las 20 preguntas del examen la denotamos como $\bar{X}_{pI}$ (con i mayúscula para aclarar que se trata no de una pregunta, sino de un conjunto de preguntas). La última columna de la Tabla 1 muestra los puntajes $\bar{X}_{pI}$ de los sustentantes.

El profesor podría haber elegido otras 20 preguntas del universo para el examen; en este caso se tendría un *examen aleatoriamente paralelo*; seguramente, las mediciones $\bar{X}_{pI}$ en este examen paralelo hubieran sido diferentes. Incluso, en vez de haber construido un examen de 20 preguntas, el profesor podría haber optado por un examen de un número diferente de preguntas; de manera general, el Estudio D responde a la siguiente pregunta sobre los

exámenes de n_i preguntas,[1] que se seleccionen de manera aleatoria del universo de preguntas: ¿con qué precisión permite la puntuación observada $\bar{X}_{pI}$ (el promedio en las n_i preguntas del examen) generalizarse al universo de todos los exámenes aleatoriamente paralelos de n_i preguntas? Este universo se llama el *universo de generalización* del Estudio D.

En el Estudio D, se consideran entonces las puntuaciones $\bar{X}_{pI}$ en exámenes paralelos de n_i preguntas en vez de los puntajes X_{pi} en preguntas individuales; con respecto a la varianza de estas puntuaciones $\bar{X}_{pI}$ (la variabilidad en las puntuaciones promedio en el universo de exámenes paralelos y en la población de personas) se puede derivar que:

$$\sigma^2_{\bar{X}} = \sigma^2_{\alpha} + \frac{\sigma^2_{\beta}}{n_i} + \frac{\sigma^2_{\varepsilon}}{n_i}, \tag{6}$$

donde las varianzas σ^2_{α}, σ^2_{β}, σ^2_{ε} se definen de manera idéntica que anteriormente en la Ecuación 4. El término σ^2_{β}/n_i indica la varianza con respecto a la dificultad (el puntaje promedio en las n_i preguntas de todas las personas en la población) entre todos los distintos exámenes aleatoriamente paralelos en el universo.[2] De la misma manera, $\sigma^2_{\varepsilon}/n_i$ se refiere a la varianza residual, otra vez, considerando a todos los exámenes aleatoriamente paralelos de n_i preguntas en la población de personas. Nótese que, al comparar las Ecuaciones 4 y 6, las puntuaciones universo de las personas (σ^2_{α}) tienen una contribución relativamente más grande a la varianza de los puntajes promedio en los exámenes (dado que la varianza de la dificultad y del efecto residual es más pequeña en el caso de conjuntos de preguntas en vez de preguntas individuales).

Para derivar el coeficiente de generalizabilidad, es importante precisar el objetivo de las decisiones que se desean tomar con base en las puntuaciones $\bar{X}_{pI}$ de los distintos sustentantes en el examen. Se distingue entre dos tipos de objetivos: primero, el profesor podría utilizar los puntajes en el examen para situar el desempeño de cada sustentante *relativo* al desempeño de los otros sustentantes. Un ejemplo de esta situación se presentaría si el profesor, con base en los puntajes del examen, quisiera seleccionar a los ocho mejores estudiantes para posteriormente invitarles a participar en la primera etapa de las Olimpiadas Mexicanas de Matemáticas (ya que solo un máximo de ocho personas de su grupo puede inscribirse). El objetivo principal en este caso es construir un *ranking* de los sustentantes en el examen y se utilizarán las puntuaciones en el Examen A para decisiones relativas. El

---

1 Lo anterior quiere decir que, a pesar de que el número de preguntas que el profesor incluyó en el Examen A es 20, el Estudio D permite contemplar qué pasaría (o hubiera pasado) si el examen consistiera en un número distinto de preguntas. Entonces, n_i para el Estudio D puede ser diferente de 20. Esto contrasta con el Estudio G, donde n_i siempre es el número de preguntas en los datos que se utilizan para el ANOVA.

2 Los lectores familiarizados con la estadística inferencial habrán notado la analogía con el resultado bien conocido que la varianza de la media muestral es igual a la varianza de la variable original entre n, el tamaño de la muestra: $\sigma^2_{\bar{X}} = \frac{\sigma^2_X}{n}$.

coeficiente de generalizabilidad para decisiones relativas para el diseño actual se estima a través de:

$$\hat{\rho}^2_{\text{rel}} = \frac{\hat{\sigma}^2_{\alpha}}{\hat{\sigma}^2_{\alpha} + \frac{\hat{\sigma}^2_{\varepsilon}}{n_i}}. \tag{7}$$

Introduciendo en esta fórmula los resultados del Estudio G (véase la Tabla 2), se obtiene directamente la siguiente estimación del coeficiente de generalizabilidad para el examen de 20 preguntas que el profesor aplicó:

$$\hat{\rho}^2_{\text{rel}} = \frac{0.0444}{0.0444 + \frac{0.1259}{20}} = .876.$$

Tabla 2. Tabla con los resultados del análisis de varianza y el estudio de generalizabilidad aplicados a los datos del Examen A

Efectos	Sumas cuadráticas	Grados de libertad	Medias cuadráticas	Estimaciones de los componentes de varianza
Personas (α_p)	67.949	67	1.0142	$\hat{\sigma}^2_{\alpha} = 0.0444$ (22.1%)
Preguntas (β_i)	42.508	19	2.2373	$\hat{\sigma}^2_{\beta} = 0.0310$ (15.4%)
Residual ($\varepsilon_{p\times i,e}$)	160.242	1273	0.1259	$\hat{\sigma}^2_{\varepsilon} = 0.1259$ (62.5%)

Nota. El archivo suplementario Script_Análisis. R contiene el código que permite llevar a cabo los cálculos en el entorno del *software* R para los resultados mostrados.

En cuanto a la interpretación de este coeficiente, cabe mencionar que valores más cercanos a 1 indican que las diferencias observadas entre los distintos participantes en el examen reflejan relativamente más diferencias entre sus puntajes universo. Para el resultado obtenido para el Examen A, casi el 88% de las diferencias observadas en el examen se deben a diferencias entre las personas (el resto, se atribuye al error de medición y/o la interacción particular entre personas y las preguntas del examen).

De manera alternativa, el objetivo del profesor podría consistir en determinar para cada persona si su desempeño alcanza cierto nivel mínimo, definido a partir de cierto criterio o norma; más concreto, el profesor puede estar interesado en saber de cada estudiante si tiene el bagaje suficiente para pasar al curso avanzado de Cálculo Diferencial e Integral del siguiente semestre e identificar a aquellos estudiantes que necesitan ejercicios adicionales

para pasar al siguiente semestre. Por ejemplo, podría ser que los estudiantes que dominaran menos del 75% de las preguntas en el universo (es decir, estudiantes con un puntaje universo menor de 75%), según la opinión del profesor, necesitan hacer una tarea adicional. En este caso, el profesor no está interesado en el desempeño de cada estudiante en comparación con otros sustentantes, sino en su nivel absoluto de desempeño; en el marco de la Teoría G, se dice que las puntuaciones en el examen se utilizan para tomar decisiones absolutas. El *coeficiente de generalizabilidad para decisiones absolutas* para el diseño actual se estima por:

$$\hat{\rho}^2_{\text{abs}} = \frac{\hat{\sigma}^2_\alpha}{\hat{\sigma}^2_\alpha + \frac{\hat{\sigma}^2_\beta}{n_i} + \frac{\hat{\sigma}^2_\varepsilon}{n_i}}, \tag{8}$$

Io cual, introduciendo los resultados del Estudio G, produce la siguiente estimación para el Examen A:

$$\hat{\rho}^2_{\text{abs}} = \frac{0.0444}{0.0444 + \frac{0.0310}{20} + \frac{0.1259}{20}} = .850.$$

Para este coeficiente, valores más cercanos a 1 indican que las diferencias entre el puntaje observado en el examen de los respectivos sustentantes y un criterio absoluto reflejan más diferencias entre la puntuación universo de los sustentantes y este criterio. En el caso del Examen A, 85% de estas diferencias entre el puntaje (porcentaje de respuestas correctas) en el examen y el criterio de 75% para decidir que el estudiante debe hacer una tarea adicional reflejan diferencias entre su puntaje universo y este mismo criterio.

Concluimos esta sección con dos comentarios. Primero, uno puede preguntarse cuál es la lógica de tener fórmulas distintas para decisiones relativas y absolutas. Comparando las Ecuaciones 7 y 8, se nota que la única diferencia es que la primera no incluye en el denominador la varianza del grado de dificultad entre los distintos exámenes aleatoriamente paralelos (σ^2_β/n_i). La razón es que, si solo nos interesa saber qué tanto Juan es mejor que Pedro, entonces no importa si el examen (que ambos contestaron) es un examen fácil o difícil. Si el examen fuera fácil, ambos tendrían un puntaje más alto en comparación con su puntaje en un examen difícil, pero la Teoría G espera que la *diferencia* entre ambos sea la misma en ambos casos, independientemente de la dificultad del examen. Retomando el ejemplo anterior: se espera que los ocho estudiantes con mejores puntajes serían los mismos, tanto si el examen fuera fácil como si fuera difícil. Por otro lado, si el interés está en una decisión absoluta, sí es importante la dificultad del examen: obviamente, si el examen resultara más difícil, los sustentantes no alcanzarían de igual manera el criterio de 75% de respuestas correctas. Este razonamiento explica la diferencia entre los coeficientes de generalizabilidad para decisiones relativas y absolutas.

Segundo, las Ecuaciones 7 y 8 muestran de una manera muy clara el efecto de aumentar (o disminuir) el número de preguntas (n_i) en el examen. El lector puede verificar que, si el Examen A consistiera en 50 en vez de 20 preguntas, los coeficientes de generalización aumentarían. En este sentido, las fórmulas de la Teoría G aquí presentadas se relacionan con la famosa fórmula de Spearman-Brown que expresa la relación entre la longitud de una prueba y la confiabilidad en la TCT.

DISEÑO CON DOS FACETAS

En esta sección, describimos el análisis para un diseño cruzado con dos facetas, elaborando el ejemplo del Examen B. En esta elaboración, nos enfocaremos en estos elementos para los cuales el diseño con dos facetas difiere o añade algo más en comparación con el diseño con una faceta. Mantenemos el mismo orden para el desarrollo que en la sección anterior: empezamos con el modelo en que se basan los análisis, incluyendo la descomposición de la puntuación observada y la varianza observada, y posteriormente delineamos los pasos del Estudio G y el Estudio D.

Descomposición de la Puntuación y la Varianza Observada

En el caso del Examen B consideramos, de manera genérica, el puntaje X_{pij} de la persona p en la pregunta i otorgado por el evaluador j (donde, como comentamos anteriormente, las preguntas y los evaluadores que se reconocen como adecuados para este examen definen el universo de mediciones admisibles para la población de personas). En el diseño cruzado de dos facetas, X_{pij} se expresa como sigue:

$$X_{pij} = \mu + \alpha_p + \beta_i + \gamma_j + (\alpha\beta)_{pi} + (\alpha\gamma)_{pj} + (\beta\gamma)_{ij} + \varepsilon_{p\times i\times j,e}, \qquad (9)$$

$$\begin{aligned}
\text{donde } \alpha_p &\equiv \mu_p - \mu,\\
\beta_i &\equiv \mu_i - \mu,\\
\gamma_j &\equiv \mu_j - \mu,\\
(\alpha\beta)_{pi} &\equiv \mu_{pi} - \mu_p - \mu_i + \mu,\\
(\alpha\gamma)_{pj} &\equiv \mu_{pj} - \mu_p - \mu_j + \mu,\\
(\beta\gamma)_{ij} &\equiv \mu_{ij} - \mu_i - \mu_j + \mu,\\
\varepsilon_{p\times i\times j,e} &\equiv X_{pij} - \mu_{pi} - \mu_{pj} - \mu_{ij} + \mu_p + \mu_i + \mu_j - \mu.
\end{aligned}$$

La media global μ se define como el promedio de la medición entre todas las personas de la población y todas las preguntas y todos los evaluadores del universo. Por otro lado, μ_p es el puntaje universo de la persona *p*; es decir, el promedio de sus puntajes en todas las preguntas, evaluadas por todos los evaluadores. Similar al diseño con una faceta, α_p es el efecto de la persona *p* e indica cómo el puntaje universo de la persona *p* se relaciona con la media global. De manera análoga, se definen μ_i, μ_j, β_i y γ_j: μ_i es la dificultad de la pregunta *i* (el promedio de los puntajes en esta pregunta de todas las personas de la población, considerando los juicios de todos los evaluadores del universo, así que valores más altos en μ_i indican una pregunta más fácil) y μ_j la exigencia del evaluador *j* (que es el promedio de todos sus puntajes otorgados a las respuestas de todas las personas de la población en todas las preguntas del universo, de tal manera que los evaluadores con μ_j más alta son menos exigentes). Las β_i y γ_j, similar a α_p, indican la diferencia de la pregunta *i* y el evaluador *j*, respectivamente, con la media global μ (la cual se puede interpretar como la dificultad promedio de todas las preguntas, así como la exigencia promedio de todos los evaluadores en el universo).

Como diferencia importante con el diseño descrito en la sección anterior, el diseño cruzado con dos facetas incluye varios términos de interacción (separados del efecto residual). En primera instancia, la interacción $(\alpha\beta)_{pi}$ da cuenta del efecto de la *combinación particular* de la persona *p* y la pregunta *i*. Por ejemplo, puede ser que, a pesar de que la pregunta *i* es difícil (con un efecto β_i negativo) y que la persona *p* tiene un nivel debajo de la media de la población (con un efecto α_p negativo), la pregunta *i* en particular le resulta más fácil a la persona *p* (por las experiencias previas de la persona). Las interacciones $(\alpha\gamma)_{pj}$ y $(\beta\gamma)_{ij}$ se interpretan de una manera similar: $(\alpha\gamma)_{pj}$ alude a la posibilidad que los evaluadores pueden ser más exigentes con unas personas que con otras. (Así, por ejemplo, en caso de favoritismos de ciertos evaluadores a ciertos sustentantes, se espera un efecto de interacción $(\alpha\gamma)_{pj}$ importante.) El efecto de interacción $(\beta\gamma)_{ij}$ significa que un evaluador califica las respuestas en ciertas preguntas de manera más o menos exigente que aquellas en otras preguntas (por ejemplo, porque está más o menos familiarizado con el tema elaborado en la pregunta o porque no entendió cabalmente los criterios de evaluación para juzgar las respuestas). La definición de estos efectos de interacción incluye a las medias μ_{pi} (el promedio de los puntajes otorgados por todos los evaluadores en el universo a la respuesta en la pregunta *i* dada por la persona *p*), μ_{pj} (el promedio de los puntajes entre todas las preguntas del universo, respondidas por la persona *p* y valoradas por el evaluador *j*) y μ_{ij} (el promedio de los puntajes de todas las personas en la pregunta *i* valoradas por el evaluador *j*). Por último, el efecto residual en el diseño cruzado con dos facetas incluye el efecto de la interacción triple entre personas, preguntas y evaluadores (que es difícil de interpretar y tiene menos importancia práctica), mezclado con el error de medición.

Como siguiente paso, se considera la variabilidad entre todas las mediciones X_{pij} (es decir, de todas las personas en la población y de todas las preguntas y todos los evaluadores en el universo), expresada por la varianza σ_X^2, la cual con base en las definiciones anteriores se particiona en siete componentes de varianza como sigue:

$$\sigma_X^2 = \sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_{\alpha\beta}^2 + \sigma_{\alpha\gamma}^2 + \sigma_{\beta\gamma}^2 + \sigma_\varepsilon^2. \quad (10)$$

En palabras, la varianza total de las mediciones se explica por diferencias entre las personas con respecto a su puntaje universo (σ_α^2), diferencias entre preguntas con respecto a su grado de dificultad (σ_β^2), diferencias entre evaluadores con respecto a su nivel de exigencia (σ_γ^2), diferencias entre las combinaciones persona-pregunta ($\sigma_{\alpha\beta}^2$), diferencias entre las combinaciones persona-evaluador ($\sigma_{\alpha\gamma}^2$), diferencias entre las combinaciones pregunta-evaluador ($\sigma_{\beta\gamma}^2$) y diferencias residuales (σ_ε^2). En la siguiente sección explicamos cómo se estiman estos componentes de varianza a partir de datos observados.

Estudio de Generalizabilidad

Para el Estudio G realizamos un ANOVA, con base en el modelo de la Ecuación 9, a los datos empíricos para el Examen B que se muestran en la parte inferior de la Tabla 1. El resultado de este análisis se encuentra en las primeras cuatro columnas de la Tabla 3. La última columna contiene las estimaciones de los componentes de varianza, que en este caso se calcularon a partir de las medias cuadráticas de la siguiente manera:

$$\hat{\sigma}_\alpha^2 = \frac{MS_\alpha - MS_{\alpha\beta} - MS_{\alpha\gamma} + MS_\varepsilon}{n_i \cdot n_j} = \frac{6.0443 - 0.8036 - 0.5023 + 0.4505}{6 \cdot 2} = 0.4324$$

$$\hat{\sigma}_\beta^2 = \frac{MS_\beta - MS_{\alpha\beta} - MS_{\beta\gamma} + MS_\varepsilon}{n_p \cdot n_j} = \frac{23.9914 - 0.8036 - 1.1849 + 0.4505}{31 \cdot 2} = 0.3622$$

$$\hat{\sigma}_\gamma^2 = \frac{MS_\gamma - MS_{\alpha\gamma} - MS_{\beta\gamma} + MS_\varepsilon}{n_p \cdot n_i} = \frac{8.4301 - 0.5023 - 1.1849 + 0.4505}{31 \cdot 6} = 0.0387$$

$$\hat{\sigma}_{\alpha\beta}^2 = \frac{MS_{\alpha\beta} - MS_\varepsilon}{n_j} = \frac{0.8036 - 0.4505}{2} = 0.1766$$

$$\hat{\sigma}_{\alpha\gamma}^2 = \frac{MS_{\alpha\gamma} - MS_\varepsilon}{n_i} = \frac{0.5023 - 0.4505}{6} = 0.0086$$

$$\hat{\sigma}_{\beta\gamma}^2 = \frac{MS_{\beta\gamma} - MS_\varepsilon}{n_p} = \frac{1.1849 - 0.4505}{31} = 0.0237$$

$$\hat{\sigma}_\varepsilon^2 = MS_\varepsilon = 0.4505$$

En la Tabla 3 se muestran también las contribuciones relativas (como porcentajes) de cada componente de varianza a la varianza total. Se estima que el 29% de la varianza total en las mediciones se debe a diferencias entre (los puntajes universo de) las personas. También las diferencias entre (las dificultades de) las preguntas explican una parte importante (24.3%) en la varianza total. Además, el componente ($\sigma^2_{\alpha\beta}$) tiene una contribución relativamente grande (11.8%) a la varianza total, lo cual apunta a una interacción importante entre las personas y las preguntas, es decir, que las personas difieren con respecto a cuáles preguntas les resultan más difíciles. Por otro lado, las diferencias entre los evaluadores en cuanto a su exigencia son relativamente pequeñas (solo explican el 2.6% de la varianza total) y también las interacciones de los evaluadores con las personas ($\sigma^2_{\alpha\gamma}$) y con las preguntas ($\sigma^2_{\beta\gamma}$) son diminutas (0.6% y 1.6%, respectivamente), por lo que se concluye que la exigencia de los evaluadores no varía mucho entre diferentes personas o para diferentes preguntas.

Tabla 3. Resultados del análisis de varianza y el estudio de generalizabilidad aplicados a los datos del Examen B

Efectos	Sumas cuadráticas	Grados de libertad	Medias cuadráticas	Estimaciones de los componentes de varianza	
Personas [α_p]	181.33	30	6.0443	$\hat{\sigma}^2_{\alpha} = 0.4324$	(29.0%)
Preguntas [β_i]	119.96	5	23.9914	$\hat{\sigma}^2_{\beta} = 0.3622$	(24.3%)
Evaluadores [γ_j]	8.43	1	8.4301	$\hat{\sigma}^2_{\gamma} = 0.0387$	(2.6%)
Personas × Preguntas [$(\alpha\beta)_{pi}$]	120.54	150	0.8036	$\hat{\sigma}^2_{\alpha\beta} = 0.1766$	(11.8%)
Personas × Evaluadores [$(\alpha\gamma)_{pj}$]	15.07	30	0.5023	$\hat{\sigma}^2_{\alpha\gamma} = 0.0086$	(0.6%)
Preguntas × Evaluadores [$(\beta\gamma)_{ij}$]	5.92	5	1.1849	$\hat{\sigma}^2_{\beta\gamma} = 0.0237$	(1.6%)
Residual ($\varepsilon_{p\times i\times j,e}$)	67.58	150	0.4505	$\hat{\sigma}^2_{\varepsilon} = 0.4505$	(30.2%)

Nota. El archivo suplementario Script_Análisis.R contiene el código que permite llevar a cabo los cálculos en el entorno del *software* R para los resultados mostrados.

Estudio de Decisión

Al tomar decisiones sobre los sustentantes a partir de sus puntajes en el Examen B, es oportuno tener en cuenta que habrían sido igualmente apropiados otras seis preguntas y otros dos evaluadores del universo de preguntas y evaluadores admisibles. Esto quiere decir que el universo de generalización del Estudio D para este caso es el universo de todos los exámenes aleatoriamente paralelos de seis preguntas y dos evaluadores.

Si es nuestro interés tomar decisiones relativas con respecto a los sustentantes, el coeficiente de generalizabilidad para este diseño cruzado de dos facetas se obtiene por:

$$\hat{\rho}^2_{\text{rel}} = \frac{\hat{\sigma}^2_{\alpha}}{\hat{\sigma}^2_{\alpha} + \frac{\hat{\sigma}^2_{\alpha\beta}}{n_i} + \frac{\hat{\sigma}^2_{\alpha\gamma}}{n_j} + \frac{\hat{\sigma}^2_{\varepsilon}}{n_i \cdot n_j}}, \tag{11}$$

lo cual con las estimaciones de los componentes de varianza sacadas del Estudio G resulta en:

$$\hat{\rho}^2_{\text{rel}} = \frac{0.4324}{0.4324 + \frac{0.1766}{6} + \frac{0.0086}{2} + \frac{0.4505}{6 \cdot 2}} = .858.$$

Nótese que en la Ecuación 11, el denominador no incluye la varianza en la dificultad de las preguntas (σ^2_{β}), ni la varianza en la exigencia entre evaluadores (σ^2_{γ}), ni la varianza debida a la interacción entre estas dos facetas $(\sigma^2_{\beta\gamma})$. Efectivamente, la dificultad global de las preguntas o la exigencia general de los evaluadores no importa al tomar decisiones relativas: si el examen paralelo consistiera, por ejemplo, en preguntas más difíciles y/o las respuestas fueran valoradas por evaluadores más exigentes, el *orden* relativo entre los sustentantes no cambiaría (aunque todos los sustentantes tendrían puntajes más bajos). Por otro lado, las interacciones entre personas y preguntas $(\sigma^2_{\alpha\beta})$ y personas y evaluadores $(\sigma^2_{\alpha\gamma})$ sí entran como varianza error para la toma de decisiones relativas: si resulta que cierta pregunta es difícil para una persona, pero fácil para otra, entonces sí las preguntas incluidas en el examen afectarían el orden entre los sustentantes, igual que la selección de evaluadores del universo afectaría este orden en el caso de que ciertos evaluadores fueran menos exigentes para unos sustentantes que para otros.

El coeficiente de generalizabilidad para decisiones absolutas en este diseño se calcula como:

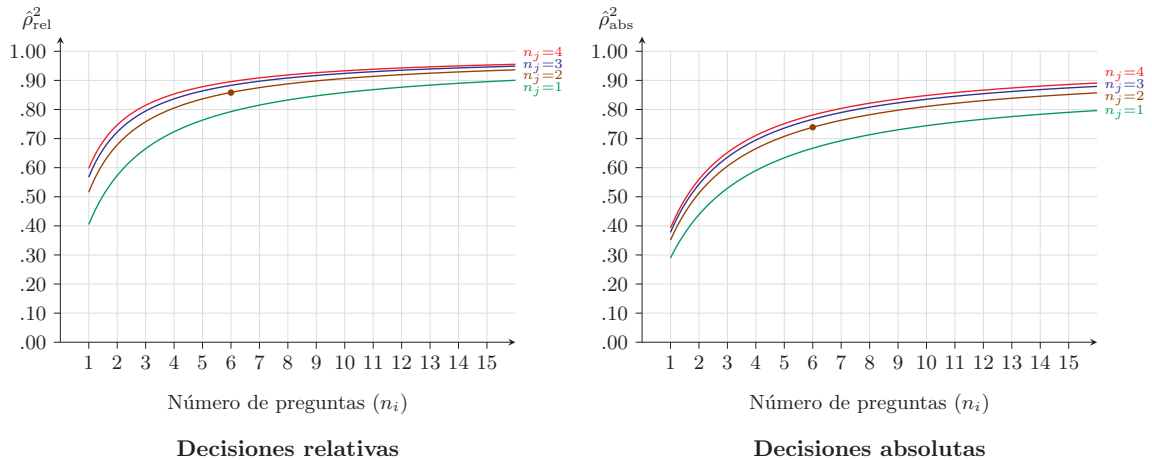
$$\hat{\rho}^2_{\text{abs}} = \frac{\hat{\sigma}^2_{\alpha}}{\hat{\sigma}^2_{\alpha} + \frac{\hat{\sigma}^2_{\beta}}{n_i} + \frac{\hat{\sigma}^2_{\gamma}}{n_j} + \frac{\hat{\sigma}^2_{\alpha\beta}}{n_i} + \frac{\hat{\sigma}^2_{\alpha\gamma}}{n_j} + \frac{\hat{\sigma}^2_{\beta\gamma}}{n_i \cdot n_j} + \frac{\hat{\sigma}^2_{\varepsilon}}{n_i \cdot n_j}} \tag{12}$$

y para el ejemplo del Examen B es igual a:

$$\hat{\rho}^2_{\text{abs}} = \frac{0.4324}{0.4324 + \frac{0.3622}{6} + \frac{0.0387}{2} + \frac{0.1766}{6} + \frac{0.0086}{2} + \frac{0.0237}{6 \cdot 2} + \frac{0.4505}{6 \cdot 2}} = .739.$$

Para las decisiones absolutas, cuando el objetivo es saber si el sustentante alcanza cierto nivel de dominio, la dificultad de las preguntas y la exigencia de los evaluadores seleccionados sí afectan las decisiones: en exámenes con preguntas más difíciles y evaluadores más exigentes, los sustentantes requieren un nivel más alto para "aprobar" (es decir, para alcanzar el nivel mínimo de dominancia). En este caso, se observa una diferencia más grande (en comparación con el ejemplo anterior del diseño con una faceta) entre los coeficientes para decisiones relativas y absolutas, lo cual principalmente se debe a la variabilidad relativamente grande de las preguntas con respecto a su dificultad (σ_β^2) y el número relativamente bajo de preguntas en el examen. El efecto del número de preguntas (n_i) tanto como del número de evaluadores (n_j) en los coeficientes de generalizabilidad para este ejemplo se muestran en la Figura 1. Esta información permitirá al profesor titular de la materia tomar decisiones sobre la organización de su examen en el futuro, por ejemplo, si conviene más aumentar el número de preguntas o bien el número de evaluadores. Es claro que el número de preguntas tiene un efecto más importante en la generalizabilidad de los puntajes derivados de su examen. Por ejemplo, si el profesor quisiera tener un examen con un índice de generalizabilidad de 0.80 para tomar decisiones absolutas, se puede leer en la Figura 1 que le serviría aumentar el número de preguntas en el examen de seis a nueve.

Figura 1. Coeficientes de generalizabilidad para decisiones relativas (panel izquierdo) y absolutas (panel derecho) en función del número de preguntas (n_i) y el número de evaluadores (n_j) para el Examen B



OTROS ASPECTOS DE LOS DISEÑOS EN LA TEORÍA DE LA GENERALIZABILIDAD

En las secciones anteriores ilustramos los análisis para dos diseños típicos y relativamente sencillos en el marco de la Teoría G. La diferencia entre ambos diseños es fundamentalmente el número de facetas consideradas; obviamente, diseños multifacéticos generalmente llevan a análisis más complejos, pero conducen a resultados e interpretaciones más ricas. A continuación, examinamos brevemente otros dos aspectos que son relevantes al diseñar estudios de generalizabilidad.

En la sección de *Aspectos Centrales de la Teoría de la Generalizabilidad* ya explicamos la diferencia entre facetas cruzadas y facetas anidadas. Cruzar todas las condiciones de las facetas en un diseño multifacético muchas veces lleva a un número demasiado alto de combinaciones que, por las limitaciones logísticas, prácticas o éticas, no se pueden llevar a cabo. Por ejemplo, un estudio en el marco de la Teoría G del examen clínico objetivo estructurado (ECOE) que es común en las Ciencias de la Salud (véase, por ejemplo, Dizon et al., 2021; Espinosa-Vázquez et al., 2017; Trejo-Mejía et al., 2016) podría fácilmente considerar las siguientes facetas: estaciones (escenarios clínicos donde el sustentante se desempeña), examinadores (que evalúan el desempeño del sustentante), sitios (por ejemplo, clínicas) y versiones del examen (que se aplican en diferentes turnos). Es obviamente imposible que todos los evaluadores valoren el desempeño de todos los sustentantes en todas las estaciones de las distintas versiones del examen realizadas en las distintas clínicas. Más bien, los examinadores suelen estar anidados en las estaciones (cada estación tiene múltiples examinadores que solo evalúan esta estación) tanto como en los sitios (en cada clínica trabaja un equipo diferente de examinadores), mientras que las estaciones están anidadas en las versiones del examen (cada versión consiste en determinadas estaciones) pero se cruzan con los sitios (en cada sitio se lleva a cabo cada estación); por otro lado, los sustentantes están anidados en versiones y sitios (ya que cada estudiante participa solo en un turno y en un sitio). Precisamente porque múltiples fuentes de error pueden afectar el resultado de los sustentantes en los ECOEs, la Teoría G se ha convertido en el estándar de oro para estudiar la calidad psicométrica de estos exámenes.

Al desarrollar los ejemplos anteriores hemos considerado las condiciones de las facetas en los datos observados como una muestra aleatoria del universo de todas las preguntas y/o de todos los evaluadores admisibles. En lenguaje técnico se dice que se trata de *facetas aleatorias*. Aunque en la gran mayoría de las aplicaciones será apropiado considerar las facetas como aleatorias, hay ocasiones donde el objetivo del estudio no es generalizar las condiciones de cierta faceta a un universo más grande (o, más bien, que se considera que las condiciones del estudio conforman el universo completo). Un ejemplo se presentaría al incluir el turno (con las condiciones matutino y vespertino) como faceta en el estudio; probablemente el objetivo no sería generalizar a un universo que incluya otros turnos. En este caso, la faceta se considera *fija*. Cabe mencionar que la manera en que las facetas entran en el diseño (anidadas vs. cruzadas, aleatorias vs. fijas) tiene repercusiones en los análisis del Estudio G tanto como los del Estudio D. El lector interesado puede consultar las publicaciones de Brennan (2001) y Webb et al. (2006).

Aunque en los ejemplos anteriores se manejó el mismo diseño (con una o dos facetas cruzadas) tanto para el Estudio G como el Estudio D, es posible considerar distintos diseños para ambos estudios. Sin embargo, aplican restricciones; en general, son preferibles los diseños cruzados para los datos recopilados que se analizan ya que estos permiten, en el Estudio G, estimar todos los posibles componentes de varianza, mientras que, en el Estudio D, se puede evaluar cómo la anidación de una faceta en otra afecta los coeficientes de generalizabilidad. Por el contrario, no es posible estimar, a partir de datos recopilados

según un diseño anidado, los resultados que se obtuviesen en un diseño cruzado. Webb et al. (2006) describen cuáles diseños se pueden analizar en el Estudio D para cada uno en una lista de diseños para el ANOVA en el Estudio G.

COMENTARIOS FINALES

Iniciamos este capítulo retomando algunas definiciones y resultados de la TCT. Esperamos que la introducción a la Teoría G en este capítulo haya dejado claro al lector que la Teoría G no contradice la TCT, sino que la extiende y complementa. Comparando la ecuación básica ($X = T + E$) de la TCT con la descomposición de la puntuación observada en las Ecuaciones 3 y 9, directamente muestra que el puntaje universo ($\mu_p \equiv \mu + \alpha_p$) en la Teoría G conceptualmente corresponde con la puntuación verdadera (T) en la TCT y que los demás componentes en las Ecuaciones 3 y 9 se juntan en la puntuación error (E). Asimismo, hay correspondencias claras entre la descomposición de la varianza de las puntuaciones observadas en la TCT (Ecuación 2) y la Teoría G (Ecuaciones 4 y 10), así como entre la definición de la confiabilidad en la TCT (Ecuación 3) y los coeficientes de generalizabilidad (en las Ecuaciones 7, 8, 11 y 12). Lo anterior muestra que la ventaja principal de la Teoría G es que separa la influencia de las distintas fuentes de error en la medición. En general, los errores son inevitables y muy variados, pero conviene mirarlos como oportunidades que nos permitan aprender y mejorar nuestros esfuerzos, en el mismo espíritu que Thomas Edison expresó en su frase célebre mostrada al inicio del capítulo. La Teoría G ofrece las herramientas para conocer mejor los errores y, aprovechando la información proporcionada por los análisis, obtendremos mediciones más precisas y tomaremos decisiones más acertadas.

Algunos autores también han comentado sobre las relaciones entre la Teoría G y el enfoque psicométrico de la teoría de respuesta al ítem (TRI, véase el capítulo anterior). El esfuerzo más notable viene de Briggs y Wilson (2007), quienes combinan los dos enfoques en un nuevo modelo: GIRM (*Generalizability in Item Response Modeling*). Al lector interesado en conocer este y otros esfuerzos de conciliar la Teoría G y la TRI, le puede resultar útil la tesis doctoral de Choi (2013).

Agradecimiento

Los autores agradecen a José J. Naveja por la lectura crítica y comentarios realizados sobre una versión previa de este capítulo.

REFERENCIAS

Brennan, R. L. (2001). *Generalizability theory.* Springer.

Briggs, D. C., y Wilson, M. (2007). Generalizability in item response modeling. *Journal of Educational Measurement*, *44*(2), 131–155. doi: https://doi.org/10.1111/j.1745-3984.2007.00031.x

Choi, J. (2013). *Advances in combining generalizability theory and item response theory* [Tesis de doctorado no publicada]. Universidad de California en Berkeley. doi: 10.13140/RG.2.1.4458.1285

Cronbach, L. J., Gleser, G. C., Nanda, H., y Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles.* Wiley.

Cronbach, L. J., Rajaratnam, N., y Gleser, G. C. (1963). Theory of generalizability: A liberation of reliability theory. *The British Journal of Statistical Psychology*, *16*(2), 137–163. doi: 10.1111/j.2044-8317.1963.tb00206.x

Dizon, S., Malcolm, J. C., Rethans, J.J., y Pugh, D. (2021). Assessing the validity of an OSCE developed to assess rare, emergent or complex clinical conditions in endocrinology & metabolism. *BMC Medical Education*, *21*, 288. doi: 10.1186/s12909-021-02653-4

Espinosa-Vázquez, O., Martínez-González, A., Sánchez-Mendiola, M., y Leenen, I. (2017). Análisis de un examen clínico objetivo estructurado en odontología desde la teoría de la generalizabilidad. *Investigación en Educación Médica*, *6*(22), 109–118. doi: 10.1016/j.riem.2016.09.001

Pardo, A., y San Martín, R. (2010). *Análisis de datos en ciencias sociales y de la salud II*. Síntesis.

Tejedor, F. J. (2019). *Análisis de varianza: Introducción conceptual y diseños básicos* (Cuadernos de Estadística no. 3, 3ra ed.). La Muralla.

Trejo-Mejía, J. A., Sánchez-Mendiola, M., Méndez-Ramírez, I., y Martínez-González, A. (2016). Reliability analysis of the objective structured clinical examination using generalizability theory. *Medical Education Online*, *21*(1), 31650. doi: 10.3402/meo.v21.31650

Webb, N. M., Shavelson, R. J., y Haertel, E. H. (2006). Reliability coefficients and generalizability theory. *Handbook of Statistics*, *26*, 81–124. doi: 10.1016/S0169-7161(06)26004-8