Universidad de Puerto Rico en Aguadilla

Departamento de Matemáticas

Recomendaciones sobre el Diseño Experimental Necesario para Hacer Avalúo en el Salón de Clases

Prof. José Neville Díaz Caraballo

Primer Semestre 2025-26

I. Contexto y Motivación

El Problema Fundamental

En evaluación educativa, demostrar que existe "alguna diferencia" entre pre-test y post-test es estadísticamente trivial pero pedagógicamente insuficiente. La pregunta crítica no es:

"¿Mejoró el estudiante?"

Sino:

> "¿Mejoró el estudiante LO SUFICIENTE para justificar la intervención pedagógica?"

Contexto del Estudio

Aspecto	Descripción
Cursos	MATE 3171 Precálculo I (secciones L91 & L11) ESMA 3101 Estadística Aplicada I (secciones LA1 & LB1)
Semestre	Primer semestre 2025-26
Institución	Universidad de Puerto Rico en Aguadilla
Profesor	José Neville Díaz Caraballo
Instrumento	Pre-test y Post-test (escala 0-100 puntos)
Diseño	Medidas pareadas (mismo estudiante, antes y después)

⚠ Declaración de Limitaciones

Tipo de muestra: Muestra de conveniencia (convenience sample)

Implicación: Los resultados NO son generalizables a:

- Toda la población de estudiantes de Precálculo o Estadística
- Otros profesores o instituciones
- Otros semestres o cohortes

Validez:

Las conclusiones se limitan estrictamente a los cursos estudiados en el contexto específico descrito.

Etikan et al. (2016). "Comparison of Convenience Sampling and Purposive Sampling." American Journal of Theoretical and Applied Statistics, 5(1), 1-4.

II. Principios de Diseño Experimental

A. Aleatoriedad

Propósito fundamental: Reducir el efecto de lo que no podemos controlar.

La aleatoriedad en la asignación de tratamientos permite:

- Distribuir equitativamente factores de confusión desconocidos entre grupos
- Obtener un estimado válido del error experimental
- Concluir con confianza estadística si las diferencias son resultado de la intervención o del azar

Limitación en Nuestro Estudio: Nuestras secciones NO fueron asignadas aleatoriamente. Los estudiantes se autoseleccionaron en las secciones por factores como horario, recomendaciones, requisitos de programa, etc.

Principios de Diseño Experimental (cont.)

B. Repetición

Propósito fundamental: Estimar la variabilidad experimental y aumentar la precisión de las estimaciones.

En nuestro estudio: Tenemos repetición a través de múltiples estudiantes en cada curso (n = 36 para Precálculo, n = 28 para ESMA) y a través de dos cursos diferentes con múltiples secciones.

C. Bloquear

Propósito fundamental: El efecto de un elemento ruidoso pero controlable se cuantifica aparte.

En nuestro estudio: El diseño de medidas pareadas (pre-test y post-test del mismo estudiante) es equivalente a un diseño en bloques donde cada estudiante es su propio bloque.

III. ¿Qué es el Avalúo?

Definición Operacional

El avalúo en educación no es simplemente "dar exámenes" o "asignar notas". Es un proceso sistemático de:

Componente	Descripción
1. Medición	Cuantificar el aprendizaje estudiantil
2. Análisis	Interpretar esos datos con herramientas estadísticas apropiadas
3. Reflexión	Examinar nuestras prácticas pedagógicas a la luz de los resultados
4. Acción	Modificar nuestra enseñanza basándonos en la evidencia

La Importancia del Modelo Estadístico Correcto

> Puedes tener un grado mayor de confiabilidad si usas el modelo estadístico correcto.

Ejemplo: Muestras Pareadas vs. Independientes

Modelo	Enfoque	Consecuencia
✗ Incorrecto	Two-sample t-test	Ignora correlación → Menor poder estadístico
✓ Correcto	Paired t-test	Aprovecha correlación → Mayor poder

Matemáticamente:

$$Var(X - Y) = Var(X) + Var(Y) - 2Cov(X, Y)$$

Cuando X e Y están positivamente correlacionadas, Cov(X, Y) > 0, entonces la varianza de la diferencia es menor.

IV. Limitaciones Críticas del Avalúo

El avalúo se puede efectuar por grupo pero NO se puede generalizar.
NO se puede usar para comparar mis grupos ni comparar profesores.

A. ¿Por Qué No Podemos Generalizar?

1. **Auto-selección:** Los estudiantes eligieron matricularse en estas secciones por razones que no controlamos
2. **Ausencia de aleatorización:** No asignamos estudiantes aleatoriamente a secciones
3. **Contexto específico:** Los resultados reflejan este profesor, estos estudiantes, este semestre

B. Ejemplos Históricos de Muestreo Fallido

Caso: Landon vs. Roosevelt (1936)

La encuesta: Literary Digest envió 10 millones de cuestionarios, recibieron 2.4 millones de respuestas, predijeron victoria de Landon.

El resultado real: Roosevelt ganó con el 62% del voto popular.

¿Qué salió mal?

1. **Sesgo de muestreo:** Encuestas a propietarios de teléfonos y automóviles → gente más rica → más republicanos
2. **Sesgo de no respuesta:** Solo respondieron ~24%

Lección: Un tamaño de muestra enorme (2.4 millones) no compensa un diseño de muestreo defectuoso.

Squire, P. (1988). Why the 1936 Literary Digest Poll Failed. Public Opinion Quarterly, 52(1), 125-133.

C. Tipos de Error

1. Error Muestral

Aproximación para proporciones con 95% confianza:

```
Error ≈ 0.98 / √n
```

Nota importante: Esta fórmula se aplica a proporciones (encuestas de opinión). Para medias (como puntuaciones de exámenes), el margen de error depende de la desviación estándar: Error = t × (s/ √n).

2. Error No-Muestral

Son todos los errores que NO se deben al tamaño de la muestra: sesgo de selección, sesgo de no respuesta, sesgo de medición, etc.

Propiedad crítica: El error no-muestral NO disminuye al aumentar n.

V. La Superioridad de las Pruebas Pareadas

A. ¿Qué es una Prueba Pareada?

Una prueba pareada compara dos mediciones en las mismas unidades experimentales:

- **Unidad experimental** = Estudiante individual
- **Primera medición** = Pre-test (inicio del semestre)
- **Segunda medición** = Post-test (final del semestre)

B. La Importancia del Baseline

El pre-test establece el **baseline** - el punto de partida de cada estudiante. Esto es crucial porque:

1. Los estudiantes no empiezan igual
2. Sin baseline, no podemos medir crecimiento
3. El baseline permite comparaciones justas: cada estudiante consigo mismo

C. ¿Por Qué Son Más Poderosas?

```
Var(X - Y) = Var(X) + Var(Y) - 2Cov(X, Y)
```

Caso	Fórmula
Caso 1 - Independientes:	Cov(X, Y) = 0 → Var(X - Y) = Var(X) + Var(Y)
Caso 2 - Pareadas:	Cov(X, Y) > 0 → Var(X - Y) < Var(X) + Var(Y)

La varianza de la diferencia en el diseño pareado es MENOR que en el diseño independiente.

Consecuencias:

- Error estándar más pequeño → Intervalos de confianza más estrechos
- Estadístico t más grande → Más probabilidad de rechazar H_0 cuando hay efecto real
- Mayor poder estadístico

D. Demostración Numérica

Presunciones del modelo: Var(Pre) = 400, Var(Post) = 320, Corr(Pre, Post) = 0.60

```
Cov(Pre, Post) = 0.60 × 20 × 17.9 = 214.8
```

Diseño	Var(dif)	SD
Independiente	720	26.8
Pareado	290.4	17.0

Resultado: Error estándar 36% menor en diseño pareado

VI. Fundamentos Teóricos: Hipótesis Nula No-Cero

A. El Paradigma Tradicional (Insuficiente)

La formulación tradicional prueba si existe "algún cambio":

$$H_0: \mu_{diff} = 0$$
$$H_1: \mu_{diff} \neq 0$$

Problema: Esta prueba puede rechazar H_0 con una diferencia de 0.5 puntos si n es suficientemente grande.

Estadísticamente significativo ≠ Pedagógicamente significativo

B. El Nuevo Paradigma: MEID

MEID (Educación): Minimum Educationally Important Difference - el cambio mínimo que justifica la inversión pedagógica

- Adaptado de MCID en medicina
- **MEID = 5 puntos** (en escala 0-100)

Justificación:

- Representa el 5% de la escala total
- Media letra de diferencia en sistema de calificación tradicional
- Margen superior al error estándar esperado de medición
- Threshold conservador para declarar "efectividad pedagógica"

Jaeschke et al. (1989). Controlled Clinical Trials, 10(4), 407-415.

Norman et al. (2003). Medical Care, 41(5), 582-592.

VII. Reformulación de las Hipótesis

LA PREGUNTA CORRECTA

¿Puedo garantizar que la mejora promedio es de AL MENOS 5 puntos?

Formulación con Threshold

Para prueba unilateral de superioridad:

$$H_0: \mu_{diff} \leq \delta_0$$
$$H_1: \mu_{diff} > \delta_0$$

Donde $\delta_0 = 5$ puntos (nuestro MEID)

Interpretación:

- **H_0:** La mejora promedio es de 5 puntos o menos (intervención insuficientemente efectiva)
- **H_1:** La mejora promedio supera los 5 puntos (intervención efectivamente superior al threshold)

Estadístico de Prueba Modificado

$$t = (\bar{d} - \delta_0) / (s_d / \sqrt{n})$$

Prueba	Fórmula	Pregunta
Tradicional	$t = (\bar{d} - 0) / (s_d / \sqrt{n})$	¿Es $\bar{d}$ diferente de cero?
Con Threshold	$t = (\bar{d} - 5) / (s_d / \sqrt{n})$	¿Es $\bar{d}$ mayor que 5?

No buscamos rechazar "no hay cambio", sino rechazar "el cambio es insuficiente".

Walker & Nowacki (2011). Journal of General Internal Medicine, 26(2), 192-196.

VIII. Análisis Completo: Precálculo I (MATE 3171)

A. Datos Descriptivos Generales

Estadístico	Pre-test	Post-test
N (válido)	47	39
N* (faltante)	4	12
Media (x̄)	31.28	70.77
Media Recortada	30.47	72.86
Desv. Estándar (s)	16.23	28.78
Coef. Variación (CV)	51.91%	40.67%
Mínimo	0	0
Q1	20	50
Mediana	30	80
Q3	40	100
Máximo	80	100

CV > 50% (Pre-test): Grupo extremadamente heterogéneo. Los estudiantes llegan con niveles de preparación muy dispares.

B. Análisis de Datos Pareados (n = 36)

Estadístico	Pre-test	Post-test	Diferencia (d)
N	36	36	36
Media ($\bar{x}$)	31.67	71.11	**39.44**
Desv. Estándar (s)	17.65	28.26	27.25
Error Estándar (SE)	2.94	4.71	4.54

Intervalo de Confianza 95% para $\mu_{difference}$

```
IC 95%: (30.22, 48.66)
```

Interpretación: Con 95% de confianza, la mejora promedio verdadera en la población está entre 30.22 y 48.66 puntos.

Nota crítica: El límite inferior del IC (30.22 puntos) está MUY por encima de nuestro MEID de 5 puntos.

C. Prueba t Tradicional (H_0: $\mu_{diff} = 0$)

```
t = (39.44 - 0) / 4.54 = 8.69
```

Resultado Minitab: t = 7.58, p-value = 0.000 (altamente significativo)

Conclusión tradicional: Rechazamos H_0. Existe diferencia estadísticamente significativa entre pre y post-test.

¿Es suficiente? NO. Hemos probado que $\mu_{diff} \neq 0$, pero eso no responde si la mejora es pedagógicamente relevante.

D. Prueba t con Threshold (H_0: $\mu_{diff} \leq 5$)

```
t = (39.44 - 5) / 4.54 = 34.44 / 4.54 = 7.58

Valor crítico: t_crítico = 1.690 (α = 0.05, una cola, gl = 35)
```

Resultado: t = 7.58 > 1.690 (altamente significativo)

p-value ≈ 0.000 (prácticamente cero)

> **Conclusión:** Existe evidencia estadística MUY FUERTE de que la mejora promedio SUPERA significativamente los 5 puntos. La intervención pedagógica es altamente efectiva según nuestro criterio de efectividad mínima.

Magnitud del Efecto

La mejora observada (39.44 puntos) es:

- **7.89 veces** el MEID de 5 puntos
- Representa el **39.44%** de la escala completa (0-100)
- Equivale a casi **4 letras** de diferencia en calificación tradicional

E. Tamaño del Efecto (Cohen's d)

$$d = \bar{d} / s_d$$

$$d = 39.44 / 27.25 = 1.45$$

Magnitud de d	Clasificación (Cohen, 1988)
0.2	Efecto pequeño
0.5	Efecto mediano
0.8	Efecto grande
1.45	**Efecto MUY GRANDE**

Nuestro resultado: d = 1.45 representa un efecto MUY GRANDE en el contexto educativo. **Contexto educativo (Hattie, 2009):** d = 0.40 es el "hinge point" - el punto donde los efectos empiezan a ser visiblemente importantes. Nuestro d = 1.45 está **3.6 veces por encima** de este threshold.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.).

Hattie, J. (2009). Visible Learning: A Synthesis of Over 800 Meta-Analyses.

F. Índice de Hakes (Ganancia Normalizada)

Fórmula:

$$g = (\langle Post \rangle - \langle Pre \rangle) / (100 - \langle Pre \rangle)$$

$$g = (71.11 - 31.67) / (100 - 31.67)$$

$$g = 39.44 / 68.33$$

$$\mathbf{g = 0.577}$$

Rango de g	Clasificación (Hake, 1998)
$g < 0.30$	Ganancia Baja
$0.30 \leq g \leq 0.70$	Ganancia Media
$g > 0.70$	Ganancia Alta
g = 0.577	**Ganancia Media**

Interpretación: Un índice de Hakes de 0.577 indica que los estudiantes aprovecharon el **57.7%** del margen de mejora disponible (de 31.67 a 100 puntos).

Nota Metodológica: Cohen's d vs. Hakes

Nota metodológica: Mientras Cohen's d mide la magnitud del efecto relativa a la variabilidad del grupo, el índice de Hakes mide el aprovechamiento del potencial de crecimiento hacia el máximo posible.

> Ambas métricas son complementarias: un d = 1.45 (excepcional) con g = 0.577 (medio) indica que el grupo mejoró sustancialmente relativo a su variabilidad, aprovechando más de la mitad del margen disponible hacia el máximo de 100 puntos.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data. American Journal of Physics, 66(1), 64-74.

G. Análisis de Heterogeneidad

Coeficiente de Variación de las diferencias:

$$CV_{diff} = (27.25 / 39.44) \times 100\% = 69.1\%$$

Interpretación: Existe considerable variabilidad en cuánto mejoraron los estudiantes individuales. Algunos estudiantes mejoraron mucho más que el promedio, otros menos.

Implicación pedagógica: Aunque la intervención fue efectiva en promedio, existe heterogeneidad en la respuesta individual. Esto sugiere que algunos estudiantes se beneficiaron más que otros de las estrategias implementadas.

IX. Análisis Comparativo: Dos Cursos

A. Resumen Descriptivo por Curso

Curso	n	Media Pre	Media Post	Mejora (đ)	s_d	SE	CV Pre
Precálculo I	36	31.67	71.11	39.44	27.25	4.54	55.7%
ESMA 3101	28	32.14	64.29	32.14	37.80	7.14	81.6%

Observaciones Descriptivas Clave

1. **Nivel inicial similar:** Ambos cursos empiezan con medias de pre-test muy similares (~31-32 puntos)
2. **Mejoras sustanciales en ambos:** Precálculo = 39.44 pts, ESMA = 32.14 pts
3. **Mayor heterogeneidad en ESMA:** CV Pre = 81.6% vs 55.7% en Precálculo
4. **Mayor variabilidad en diferencias (ESMA):** s_d = 37.80 vs 27.25 en Precálculo
5. **Precálculo alcanza nivel más alto:** Media Post = 71.11 vs 64.29 en ESMA

B. Intervalos de Confianza 95%

Curso	IC 95% para $\mu_{difference}$	Amplitud IC
Precálculo I	(30.22, 48.66)	18.44 puntos
ESMA 3101	(17.49, 46.80)	29.31 puntos

Interpretación crítica:

- **Ambos ICs están completamente por encima de 5 puntos:** Evidencia fuerte de efectividad pedagógica en ambos cursos
- **IC más estrecho en Precálculo:** Estimación más precisa debido a menor variabilidad
- **ESMA tiene mayor incertidumbre:** IC casi 60% más amplio debido a mayor heterogeneidad

C. Pruebas t con Threshold (H_0: $\mu_{diff} \leq 5$)

Curso	t (δ_0=0)	p-value	t (δ_0=5)	p-value	Cohen's d
Precálculo I	8.69	0.000	7.58	0.000	1.45
ESMA 3101	4.50	0.001	3.80	0.001	0.85

Cálculo Detallado para ESMA 3101

Prueba tradicional ($\delta_0 = 0$): $t = 32.14 / 7.14 = 4.50$
Prueba con threshold ($\delta_0 = 5$): $t = (32.14 - 5) / 7.14 = 27.14 / 7.14 = 3.80$
Valor crítico: $t_{crítico} = 1.703$ ($\alpha = 0.05$, una cola, $gl = 27$)
Cohen's d para ESMA: $d = 32.14 / 37.80 = 0.85$

Interpretación de Resultados

Precálculo I:

- Rechaza fuertemente H_0: $\mu_{diff} \leq 5$ con $t = 7.58$, $p \approx 0.000$
- La mejora supera el MEID por un margen de 34.44 puntos
- Tamaño de efecto $d = 1.45$ (muy grande, excepcional en educación)
- Evidencia robusta de alta efectividad pedagógica

ESMA 3101:

- Rechaza H_0: $\mu_{diff} \leq 5$ con $t = 3.80$, $p = 0.001$
- La mejora supera el MEID por un margen de 27.14 puntos
- Tamaño de efecto $d = 0.85$ (grande según estándares de Cohen)
- Evidencia sólida de efectividad pedagógica, aunque con mayor variabilidad individual

D. Comparación de Tamaños de Efecto

Curso	Cohen's d	Clasificación Cohen	Percentil Hattie	Interpretación
Precálculo I	1.45	Muy grande	Top 7%	Efectividad excepcional
ESMA 3101	0.85	Grande	Top 20%	Efectividad alta

Contexto según Hattie (2009):

- d = 0.40 = "hinge point" (zona de efectos visibles)
- d = 0.60 = intervención típica efectiva
- d = 0.80+ = efectos grandes raramente alcanzados
- d = 1.00+ = efectos excepcionales

> Ambos cursos superan ampliamente los estándares de efectividad educativa establecidos en la literatura.

E. Índice de Hakes para ESMA 3101

Fórmula:

$$g = (\langle Post \rangle - \langle Pre \rangle) / (100 - \langle Pre \rangle)$$

$$g = (64.29 - 32.14) / (100 - 32.14)$$

$$g = 32.15 / 67.86$$

$$\mathbf{g = 0.474}$$

Curso	Cohen's d	Hakes (g)	Clasificación d	Clasificación g
Precálculo I	1.45	0.577	Muy Grande	Media
ESMA 3101	0.85	0.474	Grande	Media

Interpretación Índice de Hakes

Interpretación para ESMA 3101: Un índice de Hakes de 0.474 indica que los estudiantes aprovecharon el **47.4%** del margen de mejora disponible (de 32.14 a 100 puntos). Esta ganancia normalizada se clasifica como **Media** según los estándares de Hake (1998).

Comparación entre cursos: Ambos cursos muestran ganancias normalizadas en el rango medio (Precálculo: 57.7%, ESMA: 47.4%), lo que indica que en ambos casos los estudiantes aprovecharon aproximadamente la mitad del margen de mejora posible.
Las diferencias en Cohen's d (1.45 vs 0.85) reflejan principalmente diferencias en la variabilidad de las respuestas individuales más que en la efectividad pedagógica absoluta.

F. Diferencias Entre Cursos: Reflexión Pedagógica

¿Por qué Precálculo muestra mayor efecto que ESMA?

Las diferencias pueden reflejar:

Factor	Precálculo I	ESMA 3101
Naturaleza del contenido	Secuencial, acumulativo, estructurado jerárquicamente	Más conceptual, menos dependiente de secuencia estricta
Población estudiantil	Programa de matemáticas/ciencias	Curso de servicio (mayor diversidad de programas)
Prerequisitos	Álgebra intermedia específica	Prerequisito menos específico
Homogeneidad	Relativamente más homogéneo (CV = 55.7%)	Muy heterogéneo (CV = 81.6%)
Tipo de conocimiento	Procedimental + conceptual	Estadístico-probabilístico (mayor abstracción)

⚠ Advertencia Metodológica Crítica

NO podemos comparar estadísticamente estos dos cursos porque:

1. **Son poblaciones diferentes (no asignación aleatoria)**
2. **Contenido diferente (Precálculo ≠ Estadística)**
3. **Los pre-tests y post-tests miden constructos diferentes**
4. **Las escalas de 0-100 no son equivalentes entre cursos**

Lo que SÍ podemos decir: Ambas intervenciones pedagógicas fueron efectivas en sus respectivos contextos, superando ampliamente el criterio de efectividad mínima de 5 puntos.

Implicación práctica importante: La mayor variabilidad en ESMA NO indica menor efectividad de la enseñanza, sino mayor diversidad en la población. Algunos estudiantes de ESMA mejoraron tanto o más que los de Precálculo, pero la variabilidad es mayor.

G. Síntesis Comparativa

Conclusión Principal:

Ambos cursos demuestran efectividad pedagógica robusta y estadísticamente significativa:

1. **Precálculo I:** Efectividad excepcional con d = 1.45, mejora promedio de 39.44 puntos
2. **ESMA 3101:** Efectividad alta con d = 0.85, mejora promedio de 32.14 puntos

Las diferencias en magnitud del efecto reflejan diferencias en naturaleza del contenido, características de la población estudiantil, y homogeneidad de prerequisitos.

> Ambas intervenciones superan ampliamente el threshold de 5 puntos y se encuentran en el rango superior de efectividad educativa según estándares internacionales.

X. Justificación Pedagógica: Los 5 Puntos

A. La Objeción Inicial

Pregunta: ¿Por qué compensar con 5 puntos y no 3 o 7?

Aparente problema: La elección parece arbitraria.

B. La Contra-Respuesta

Si cuestionamos los 5 puntos de compensación por arbitrarios, debemos cuestionar también:

- ¿Por qué A = 90-100?
- ¿Por qué B = 80-89?
- ¿Por qué C = 70-79?

> La escala tradicional 90/80/70/60 también es una convención arbitraria establecida sin fundamentación empírica rigurosa.

C. La Arbitrariedad Fundamental

Toda evaluación educativa está construida sobre múltiples capas de decisiones arbitrarias del profesor.

Aspecto	Ejemplos de Decisiones Arbitrarias
Selección de Contenido	¿Por qué preguntar sobre factorización y no sobre funciones inversas? ¿Por qué 3 problemas de trigonometría y no 5?
Distribución de Puntos	¿Por qué la pregunta de límites vale 20 puntos y la de derivadas 15 puntos?
Cantidad de Preguntas	¿Por qué 8 preguntas y no 12? ¿Por qué 50 minutos de examen y no 75?
Nivel de Dificultad	¿Por qué pedir "derivar x^3" y no "derivar $e^{sen(x^2)}$"?
Orden de Presentación	¿Pregunta difícil primero (intimidar) o al final?

D. La Ecuación Real de una Nota

Lo que queremos medir:

Conocimiento del estudiante

Lo que realmente medimos:

La intersección entre el conocimiento del estudiante y ~50+ decisiones arbitrarias del profesor sobre qué preguntaste, cómo lo preguntaste, cuánto vale cada pregunta, qué dificultad elegiste, en qué orden las pusiste, etc.

E. Los 5 Puntos de Humildad al Medir

No preguntamos: ¿Son arbitrarios 5 puntos de compensación?

Preguntamos: Dado que el profesor toma ~50+ decisiones arbitrarias que afectan significativamente la nota final, ¿son 5 puntos de compensación suficientes o insuficientes?

Los 5 puntos de compensación no son una concesión generosa.

Son humildad al medir - un reconocimiento honesto de que:

1. **Toda evaluación está saturada de arbitrariedad en su construcción**
2. **El estudiante no tiene control sobre ninguna de esas decisiones arbitrarias**
3. **Esas decisiones afectan la nota tanto o más que el conocimiento real**
4. **5 puntos son mínimos comparados con la variabilidad introducida por todas esas elecciones**

F. Convergencia de Justificaciones

Los 5 puntos aparecen en este análisis con dos justificaciones independientes pero convergentes:

1. **MEID estadístico:** Threshold para declarar efectividad pedagógica (5% de la escala)
2. **Humildad al medir:** Reconocimiento de incertidumbre inherente a la medición

No es coincidencia. La arbitrariedad en la construcción del examen GENERA variabilidad de medición, que a su vez establece el límite estadístico de lo que podemos detectar confiablemente.

G. Mensaje Final

La evaluación educativa es un acto de juicio profesional, no una ciencia exacta.

Los 5 puntos de compensación no distorsionan la medición - revelan su naturaleza aproximada.

XI. Limitaciones Metodológicas Adicionales

A. Naturaleza de las Muestras

Muestras de Conveniencia: Implicaciones Críticas

Definición: Las secciones L91, L11, LA1 y LB1 son muestras de conveniencia - estudiantes que se matricularon en estas secciones específicas con este profesor específico en este semestre específico.

Lo que NO podemos concluir:

- "Este método funciona para todos los estudiantes de Precálculo o Estadística"
- "Otros profesores obtendrán los mismos resultados"
- "Estos resultados se replicarán en futuros semestres"

Lo que SÍ podemos concluir:

- "En estos cursos específicos, hubo mejora significativa superior a 5 puntos"
- "La intervención fue efectiva en este contexto particular"
- "Los resultados justifican investigación adicional con diseño más robusto"

B. Grupos Heterogéneos y Homocedasticidad

Observación: CV > 50% en ambos cursos indica grupos altamente heterogéneos.

Supuesto de homocedasticidad: Los tests paramétricos asumen varianzas relativamente homogéneas. La alta heterogeneidad puede inflar la tasa de error Tipo I.

Justificación para proceder con prueba t estándar:

- Tamaño de muestra razonable (n ≥ 28)
- Diferencias son sustanciales comparadas con variabilidad
- Resultados consistentes en ambos cursos
- La prueba t es robusta a violaciones moderadas con n > 25

Zimmerman (2004). British Journal of Mathematical and Statistical Psychology, 57(1), 173-181.

C. Distribuciones Bimodales

Evidencia de subgrupos distintos:

1. **Estudiantes preparados:** Con base sólida de prerequisitos, puntúan > 40 en pre-test
2. **Estudiantes no preparados:** Con deficiencias en prerequisitos, puntúan < 25 en pre-test

Análisis adicional recomendado:

- Estratificación por nivel de pre-test (bajo/medio/alto)
- Análisis de regresión: ¿La mejora depende del nivel inicial?
- Identificación de estudiantes "en riesgo" para intervenciones dirigidas

D. Efectos del Pre-test

Problema: El acto de tomar el pre-test puede influir en el desempeño del post-test independientemente de la instrucción.

Mecanismos: Sensibilización, práctica, memoria

Nuestro contexto:

- **Efecto observado:** d = 0.85-1.45 (mucho mayor que efecto de testing solo)
- **Tiempo entre mediciones:** ~8 semanas (reduce memoria)
- **Contenido no idéntico:** Pre y post tienen problemas diferentes

Conclusión: El testing effect puede contribuir a las mejoras, pero efectos de la magnitud observada no se explican completamente por este fenómeno (Shadish et al., 2002).

E. Datos Faltantes

Precálculo I:

- Pre-test válido = 47 (N* = 4 ausentes)
- Post-test válido = 39 (N* = 12 ausentes)
- Pareados completos = 36

ESMA 3101:

- Pre-test válido = 32 (N* = 4 ausentes)
- Post-test válido = 30 (N* = 6 ausentes)
- Pareados completos = 28

Los estudiantes con datos faltantes NO abandonaron el curso - simplemente no participaron en el pre-test y/o post-test. Obligar participación introduce sesgo de desmotivación: estudiantes forzados producen respuestas de baja calidad, contaminando la variabilidad real.

Las tasas de participación (Precálculo 70.6%, ESMA 77.8%) son razonables para estudios reales en aula.

XII. Síntesis y Conclusiones

A. Principales Hallazgos

Evidencia de Efectividad Pedagógica

1. Mejoras excepcionales en Precálculo I:

- Mejora promedio = 39.44 puntos (IC 95%: 30.22 a 48.66)
- Cohen's d = 1.45 (efecto muy grande, excepcional en educación)
- Hakes (g) = 0.577 (ganancia media, 57.7% del margen aprovechado)
- Supera el MEID de 5 puntos por un margen de 34.44 puntos
- t = 7.58, p < 0.001 (evidencia estadística muy fuerte)

2. Mejoras sustanciales en ESMA 3101:

- Mejora promedio = 32.14 puntos (IC 95%: 17.49 a 46.80)
- Cohen's d = 0.85 (efecto grande)
- Hakes (g) = 0.474 (ganancia media, 47.4% del margen aprovechado)
- Supera el MEID de 5 puntos por un margen de 27.14 puntos
- t = 3.80, p = 0.001 (evidencia estadística fuerte)

Principales Hallazgos (cont.)

3. Consistencia de resultados:

Ambos cursos superan ampliamente el criterio de efectividad mínima

4. Magnitud excepcional:

Los tamaños de efecto se encuentran en el rango superior de intervenciones educativas según meta-análisis internacionales

B. Contribución Metodológica

Este análisis demuestra la importancia de establecer criterios de efectividad mínima (MEID) en investigación educativa:

Ventajas del enfoque con threshold:

- **Relevancia pedagógica:** Distingue entre "estadísticamente significativo" y "pedagógicamente significativo"
- **Transparencia:** El criterio de 5 puntos es explícito y justificable
- **Replicabilidad:** Otros investigadores pueden usar el mismo criterio
- **Prevención de p-hacking:** El threshold se establece a priori

Recomendación: Los estudios de evaluación educativa deberían complementar las pruebas tradicionales (H_0: $\delta = 0$) con pruebas de superioridad usando thresholds pedagógicamente significativos.

C. Limitaciones Reconocidas

Interpretación Cautelosa Requerida

1. **No generalizable:** Muestras de conveniencia limitan las conclusiones a estos cursos específicos
2. **Heterogeneidad alta:** Grupos muy diversos con respuestas variables a la intervención
3. **Diseño pre-experimental:** Validez interna limitada por ausencia de aleatorización
4. **Datos faltantes:** Aproximadamente 25-30% de no-participación

D. Implicaciones para la Práctica

Para Profesores de Matemáticas y Estadística:

- **Medición baseline esencial:** El pre-test permite identificar necesidades y medir crecimiento real
- **Establecer criterios a priori:** Antes del semestre, defina qué constituye "mejora significativa"
- **Reconocer heterogeneidad:** Los grupos diversos requieren estrategias diferenciadas
- **Ir más allá del promedio:** Analice subgrupos para identificar quién se beneficia más/menos
- **Documentar sistemáticamente:** El avalúo riguroso requiere diseño, medición y análisis planificados

Para Investigadores Educativos

- **Adoptar MEID:** Establecer thresholds de efectividad mínima como estándar
- **Reportar tamaños de efecto:** Complementar p-values con medidas de magnitud (Cohen's d, Hakes)
- **Transparencia sobre limitaciones:** Reconocer muestras de conveniencia
- **Diseños más robustos:** Cuando sea posible, usar aleatorización y medidas repetidas
- **Intervalos de confianza:** Reportar ICs para todas las estimaciones

E. Futuras Direcciones

1. **Replicación con otros profesores:** ¿El mismo framework produce resultados similares en otros contextos?
2. **Análisis de subgrupos:** Estratificar por nivel de preparación inicial
3. **Medidas adicionales:** Incorporar otras variables (asistencia, tiempo de estudio, actitud)
4. **Seguimiento longitudinal:** ¿Las mejoras se mantienen en cursos posteriores?
5. **Análisis cualitativo:** Entrevistas con estudiantes sobre experiencias de aprendizaje
6. **Análisis de covarianza:** Controlar por variables como nivel inicial, edad, programa de estudios

XIII. Conclusión Final

Mensaje Central

Este estudio demuestra que en los cursos de Precálculo I y ESMA 3101 analizados durante el primer semestre 2025-26 en UPR Aguadilla, los estudiantes experimentaron mejoras promedio sustanciales y estadísticamente significativas:

- **Precálculo I (n=36):** Mejora de 39.44 puntos, d = 1.45, g = 0.577 - efectividad excepcional
- **ESMA 3101 (n=28):** Mejora de 32.14 puntos, d = 0.85, g = 0.474 - efectividad alta

Ambos cursos superaron ampliamente el criterio de efectividad mínima de 5 puntos, con evidencia estadística muy fuerte ($p < 0.001$).

Sin embargo, la verdadera contribución de este trabajo no es solo demostrar efectividad en contextos específicos, sino **proponer un marco metodológico más riguroso para evaluación educativa**: uno que vaya más allá de la mera significancia estadística para establecer criterios explícitos de significancia pedagógica.

La Pregunta que Debemos Hacernos

En educación, no basta con preguntar "¿funcionó?" Debemos preguntar:

- "¿Funcionó lo suficientemente bien para justificar la inversión de tiempo y recursos?"
- "¿La mejora observada tiene relevancia práctica en la vida académica de los estudiantes?"
- "¿Puedo replicar este resultado con confianza en otros contextos?"

> Las hipótesis con threshold de efectividad mínima nos obligan a confrontar estas preguntas incómodas pero esenciales.

Reflexión Final sobre los 5 Puntos

La controversia sobre los 5 puntos de compensación reveló una verdad más profunda sobre la evaluación educativa:

> *Cada nota es el resultado de decenas de decisiones arbitrarias del profesor - decisiones sobre las cuales el estudiante no tiene control.*

Los 5 puntos no son generosidad indebida. Son un reconocimiento humilde de que nuestros instrumentos de medición, por bien diseñados que estén, capturan solo una aproximación del conocimiento real del estudiante.

La pregunta no es: "¿Por qué compensar con 5 puntos?"

La pregunta es: "¿Por qué pretendemos que nuestras mediciones son tan

Referencias Bibliográficas

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates.

Etikan, I., Musa, S. A., & Alkassim, R. S. (2016). Comparison of Convenience Sampling and Purposive Sampling. *American Journal of Theoretical and Applied Statistics*, 5(1), 1-4.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data. *American Journal of Physics*, 66(1), 64-74.

Hattie, J. (2009). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. Routledge.

Jaeschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of health status: Ascertaining the minimal clinically important difference. *Controlled Clinical Trials*, 10(4), 407-415.

Kelley, K. (2007). Sample size planning for the coefficient of variation from the accuracy in parameter estimation approach. *Behavior Research Methods*, 39(4), 755-766.

Referencias Bibliográficas (cont.)

Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Medical Care*, 41(5), 582-592.

Pfister, R., Schwarz, K. A., Janczyk, M., Dale, R., & Freeman, J. B. (2013). Good things peak in pairs: A note on the bimodality coefficient. *Frontiers in Psychology*, 4, 700.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin.

Squire, P. (1988). Why the 1936 Literary Digest Poll Failed. *Public Opinion Quarterly*, 52(1), 125-133.

Walker, E., & Nowacki, A. S. (2011). Understanding equivalence and noninferiority testing. *Journal of General Internal Medicine*, 26(2), 192-196.

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, 57(1), 173-181.

Universidad de Puerto Rico en Aguadilla

Departamento de Matemáticas

Comprometidos con la excelencia en educación matemática

Prof. José Neville Díaz Caraballo

jose.neville@upr.edu

Fecha de elaboración: Octubre 2025