Modelos de Combinación de Medidas : Confiabilidad, Validez y Consecuencias para la Evaluación Docente

José Felipe Martínez
Jonathan Schweig

University of California, Los Angeles
Graduate School of Education

IV Seminario Internacional sobre Calidad de la Educación. Bogotá, Colombia, Noviembre 7, 2013

Indice

- Evaluación Docente: Contexto
  - Porqué Evaluar, Qué Evaluar, Cómo Evaluar
- Evaluación Docente con Indicadores Múltiples
  - Consideraciones Conceptuales y Metodológicas
  - Modelos de Combinación de Indicadores
- Estudio Empírico
  - Métodos: Simulación y base de datos MET
  - Resultados: Confiabilidad y Precisión
  - Resultados: Consistencia de juicio entre modelos
  - Discusión y Consideraciones Finales

Evaluación Docente: Contexto

Evaluación Docente y Política Educativa

- Gran interés en mejorar la evaluación docente a nivel internacional. Asociado a varios factores:
  - Interpretación pesimista de resultados en pruebas nacionales e internacionales (e.g. Feuer, 2012)
  - Evidencia (teórica y empírica) del "efecto docente" (e.g. Baker et. al. 2010)
  - Sistemas de evaluación perfunctorios, superficiales, e inefectivos, sin valor (in)formativo para maestros, directores, o autoridades
  - Evidencia (teórica y empírica) del efecto de sistemas comprensivos de supervisión y desarrollo profesional (e.g. Taylor and Tyler, 2012)

Ejemplos Notables

- Estados Unidos
  - Race to the Top (2010)
    - Denver (2010)
    - DCPS (2011)
    - New York, Los Ángeles, Chicago (2012)
  - Toledo, Cincinnati (1990s)
- A nivel internacional
  - Singapur (2006)
  - Chile (2003)
  - México (1993,2009,2014)
  - Colombia (2002…)

Porqué Evaluar?

- Distintas motivaciones, inferencias y usos:
  - Orientar el desarrollo profesional de los docentes
  - Incentivar a los "*mejores*" docentes
  - Ayudar a mejorar aspectos de la practica docente
  - Desvincular a docentes con problemas serios persistentes
  - Informar las políticas educativas
  - Identificar y propagar las practicas efectivas
  - Otros…

  - Y cualquier combinación de las anteriores.

Qué Evaluar?

- *Competencias Docentes* (Reynolds, 1999):
  - Conocimiento: Sujeto, Pedagógico
  - Habilidad: Conocimiento *aplicado*
  - Disposición: Actitudes, Percepciones, Creencias
  - Practicas: Procesos de Aula (e.g. instrucción, evaluación en aula, manejo)
- Y ..
  - Antigüedad, Preparación
  - Ciudadanía, contribuciones a la comunidad
  - *Efectividad:* Habilidad de mejorar el aprendizaje de sus alumnos

Cómo Evaluar?

Constructos del Docente *(Qué?)*	Medidas *(Cómo?)*
Conocimientos y Habilidades *(Sustantivos, pedagógicos, aplicados)*	Pruebas Estandarizadas Pruebas de Rendimiento Viñetas
Prácticas, Procesos de Aula *(instrucción, evaluación, manejo)*	Encuestas, Bitácoras Observación en Aula, Video Artefactos, Portafolios
Disposición (Creencias, *actitudes)*	Encuestas, Entrevistas
Ciudadanía (contribución a la comunidad)	Encuestas, Entrevistas, Autoevaluación
Efectividad (contribución al aprendizaje de los alumnos)	Pruebas Estandarizadas; "Valor Agregado"

Cual de los métodos es *mejor*?

- Ninguno es inherentemente preferible
- Todos ofrecen una imagen parcial del constructo
  - *Desempeño, Calidad, Competencia, o Efectividad Docente*
- Y (des)ventajas metodológicas y prácticas
  - Cobertura
  - Confiabilidad
  - Validez
  - Costo
  - Implicaciones logísticas
  - Incentivos y presiones
  - Etc… (Correnti and Martinez, 2012)

Evaluación Docente e Indicadores Múltiples

- Existe un acuerdo prácticamente universal acerca de la necesidad de usar indicadores múltiples en la evaluación docente

En el contexto educativo, ninguna decisión descripción que tendrá un impacto significativo de deberá hacer con base en un solo puntaje o medida. Otra información relevante deberá tomarse en cuenta si esta mejora la validez de la decisión.

Standards for Educational and Psychological Testing, Standard 13.7

(AERA, APA, & NCME, 1999)

Evaluación Docente e Indicadores Múltiples

"Entendemos bien que las pruebas estandarizadas no capturan todos los aspectos y cualidades importantes de la enseñanza. Por ello llamamos a utilizar indicadores múltiples para evaluar a los docentes. De preferencia (in an ideal world), esos mismos datos también deberían guiar la practica docente y el desarrollo profesional."

Duncan (2011)

Indicadores Múltiples: Confiabilidad y Validez

Indicadores Multiples : Supuestos

- Supuesto General:
  - Combinar indicadores múltiples conduce a decisiones *mejor informadas (más validas)* sobre los docentes

Precisión	-Clasificar a los docentes en categorías más finas y estables (De Pascale, 2012; Steele et. al. 2010)
Validez	-Imagen más completa de calidad docente (Goe, 2011) -Menor incentivo para el fraude (Steele et. al. 2010)
Retroalimentación	- Ayudar al maestro a ajustar y mejorar sus estrategias y practicas en el aula (Duncan, 2011)
Relevancia	- Mayor confianza en los resultados de la evaluación (Glazerman et. al. 2011)

Indicadores Múltiples y Validez

"Un juicio evaluativo integrado que refleja el grado en que la evidencia empírica y teórica sugiere que las inferencias y acciones basadas en los puntajes son apropiadas."

Messick (1989)

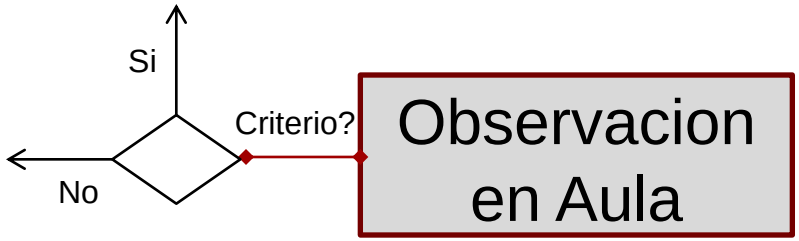
- Como evaluar validez al combinar indicadores?
  - Depende del constructo: y el modo en que este se refleja en los indicadores disponibles
  - Depende del contexto: La validez se refiere a inferencias y usos especificos
  - Depende de que se entiende por *combinar*
    - No es obvio, no se explica por si mismo
    - Varios modelos posibles *(Brookhart, 2009)*
    - Empiezan a aplicarse a la evaluación docente

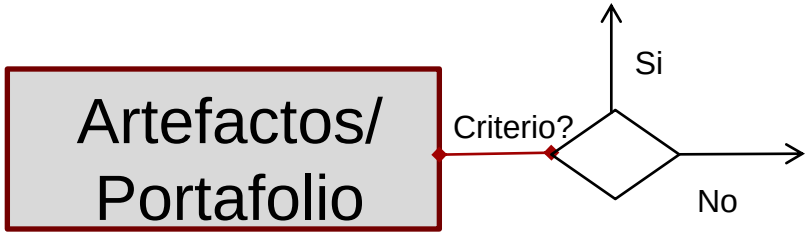
Modelos básicos para combinación de Indicadores Múltiples

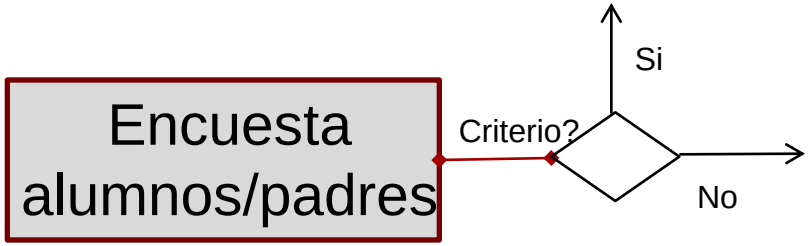
Modelo	Descripcion
Conjuntivo	Se debe satisfacer el criterio (*pasar*) con cada uno de los indicadores
Disyuntivo (Complementario)	Se debe satisfacer el criterio (*pasar*) con uno o mas de los indicadores
Compensatorio	Medida compuesta ponderada. (Un puntaje alto puede *compensar* puntajes menores)
Hibrido(*Complejo*)	e.g. Compensatorio-conjuntivo, Secuencial

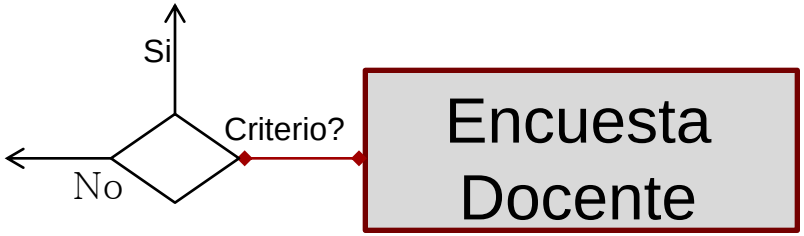
(Mehrens, 1989; Chester, 2003)

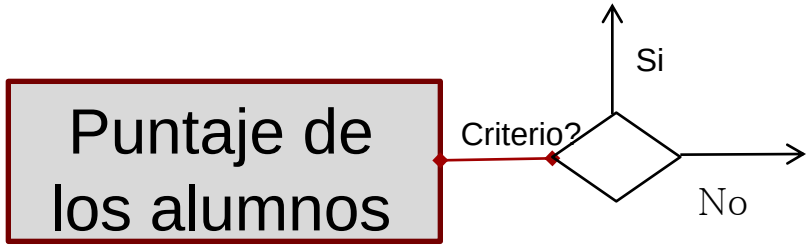
Modelo de Combinación 1: Conjuntivo, Disyuntivo









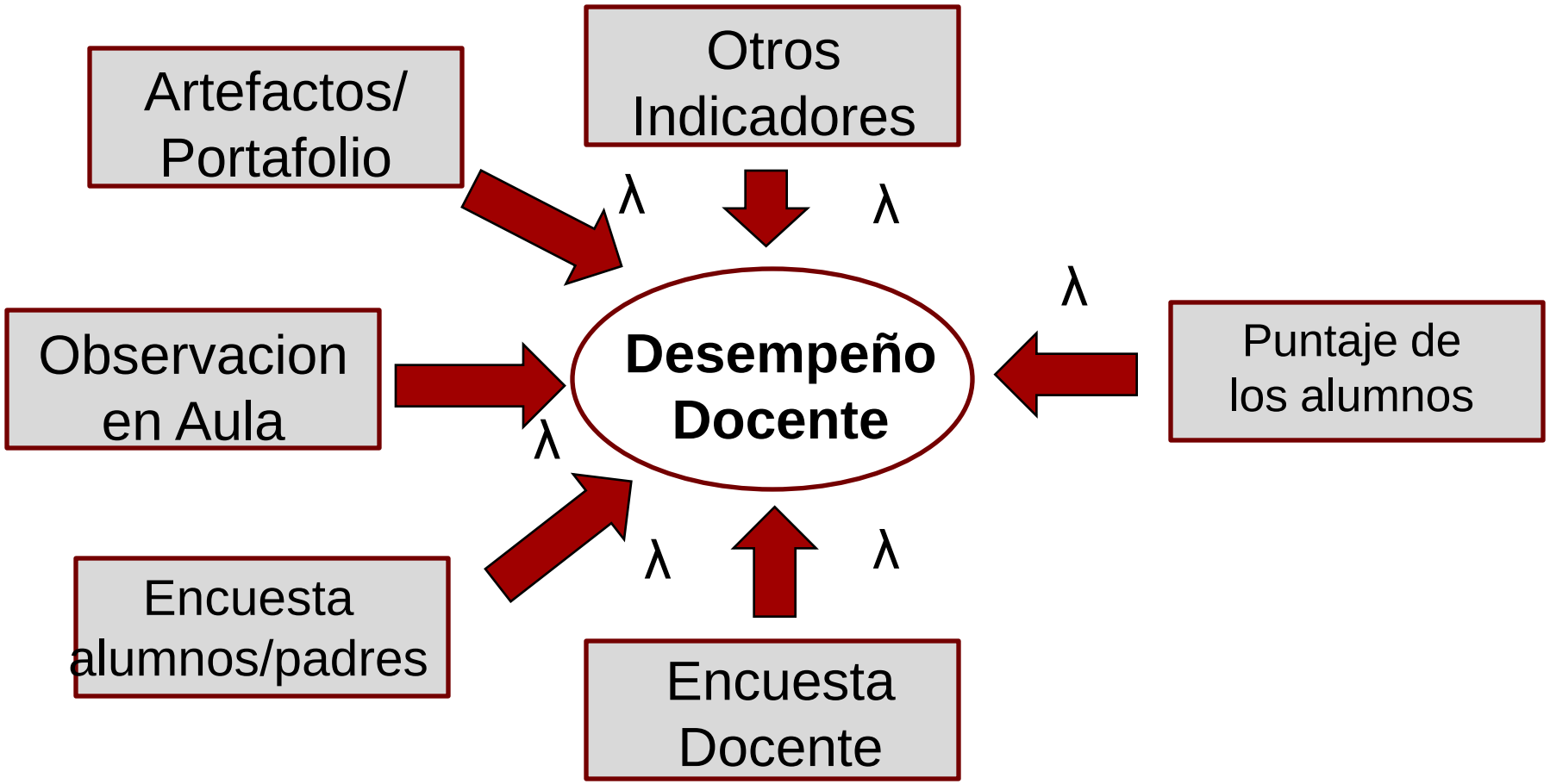


Reglas de Decisión y Confiabilidad

- El modelo de combinación es mas importante para la precisión de las inferencias resultantes que la propia confiabilidad de las medidas (Chester, 2003)

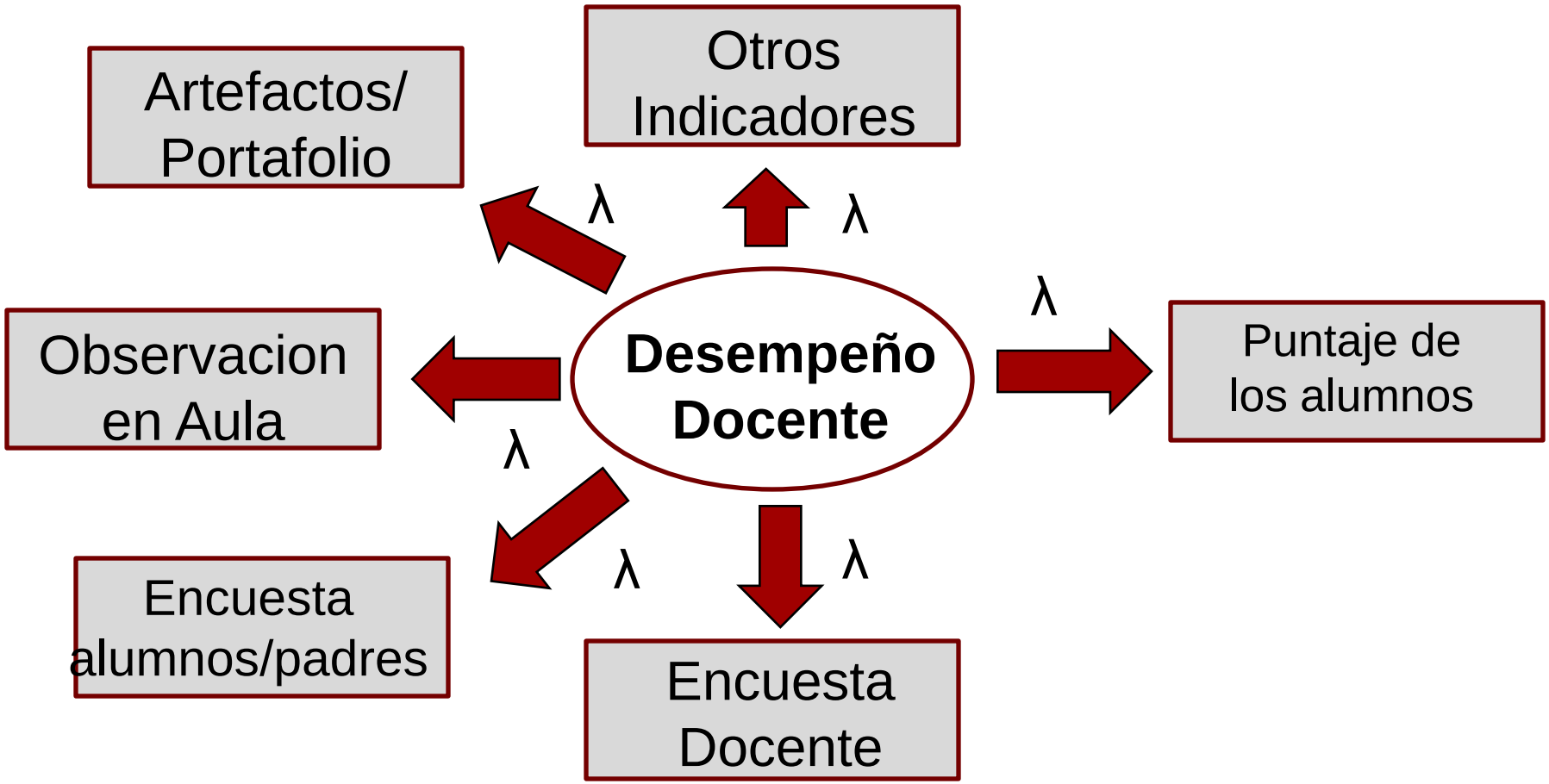
- Cada modelo incluye elementos de juicio
  - Porqué se requiere satisfacer k criterios y no k-1? Porqué esos criterios en particular?

Modelo de Combinación 2 (Compensatorio): "Componentes Principales/Confiabilidad"

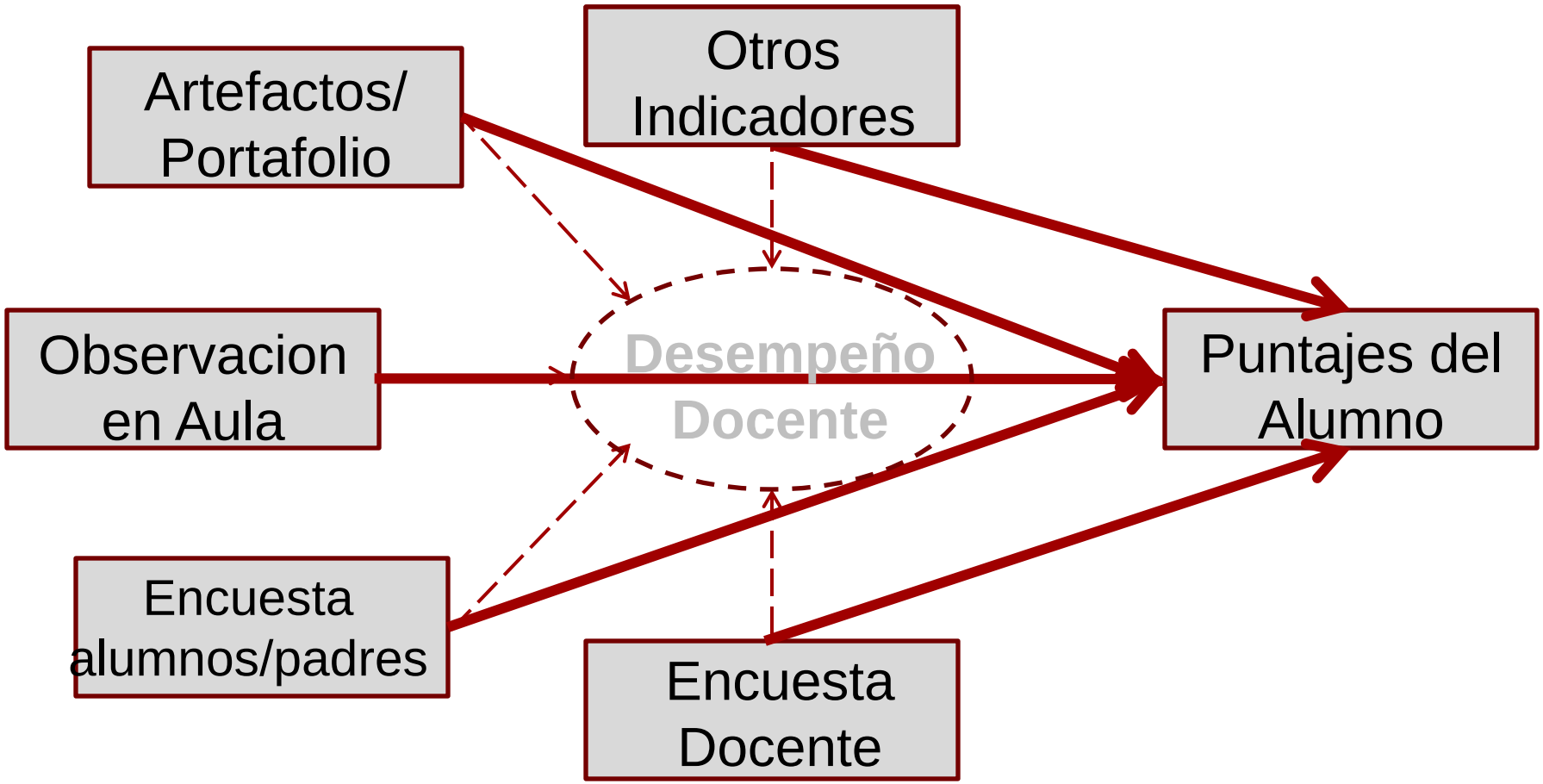


Modelo de Combinación 3 (Compensatorio): "Análisis Factorial"

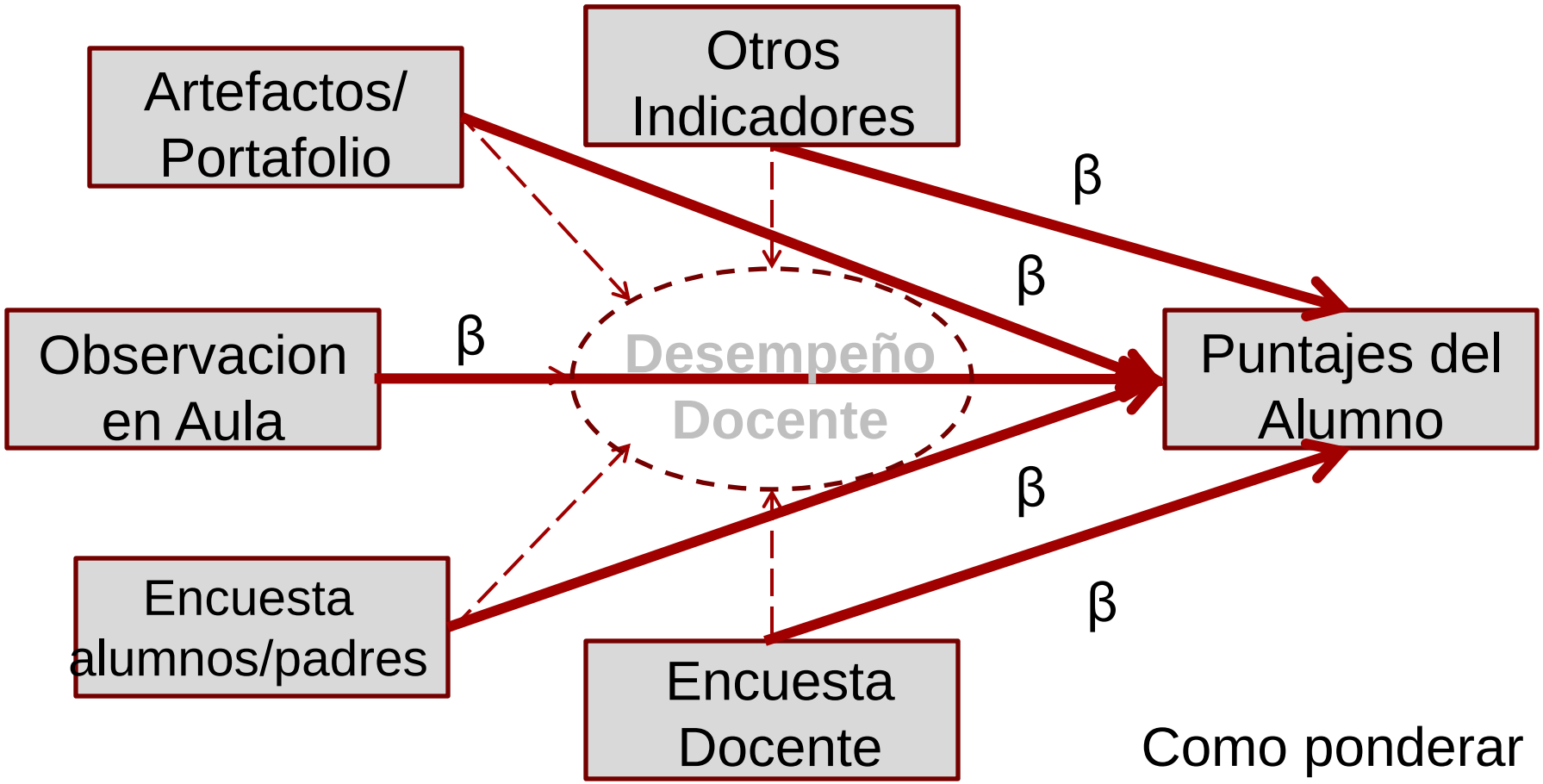
19



Modelo de Combinación 4 (Compensatorio): Peso Optimo (Puntajes como Criterio)

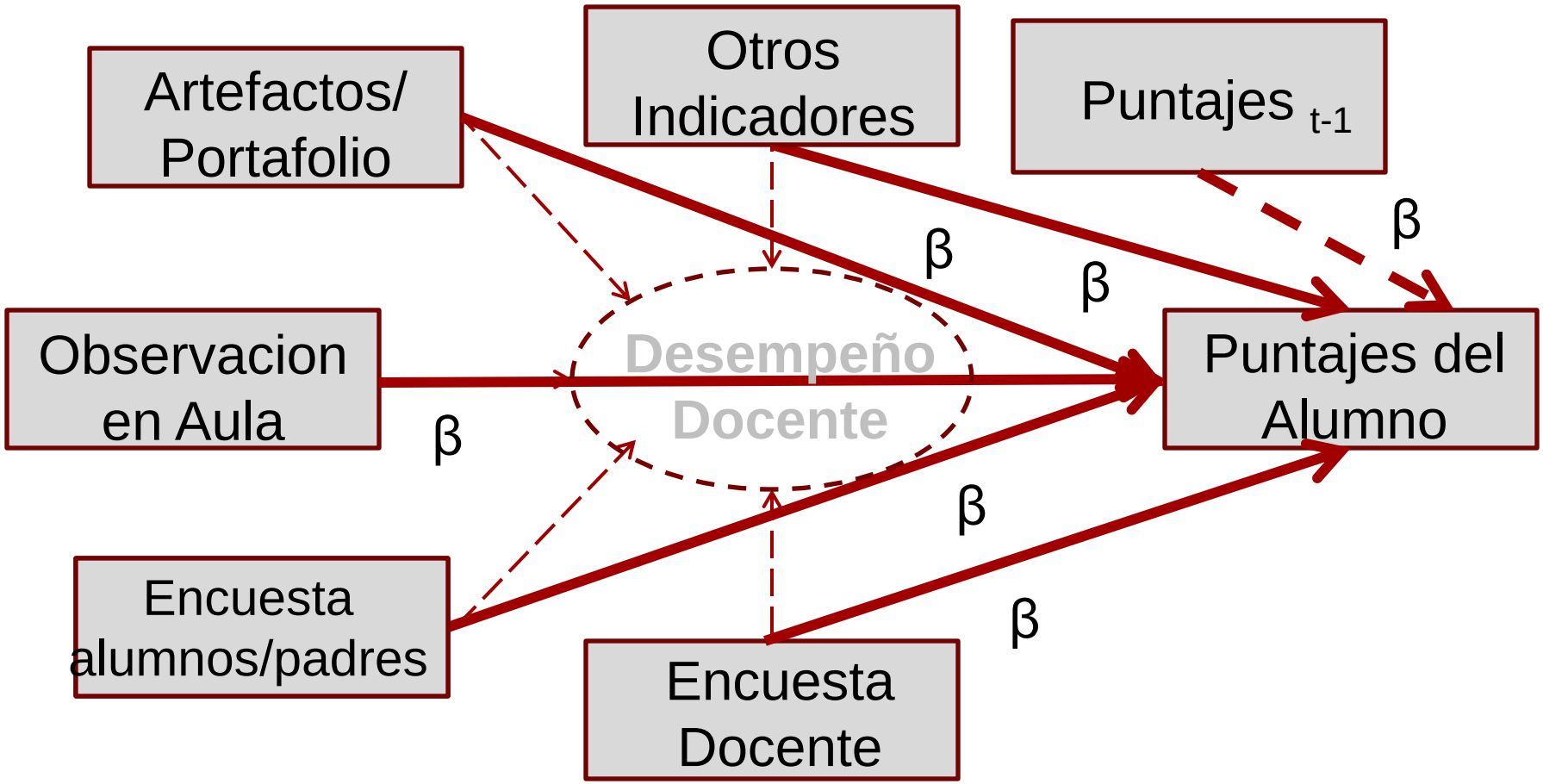


Modelo de Combinación 4 (Compensatorio): Peso Óptimo (Puntajes como Criterio)

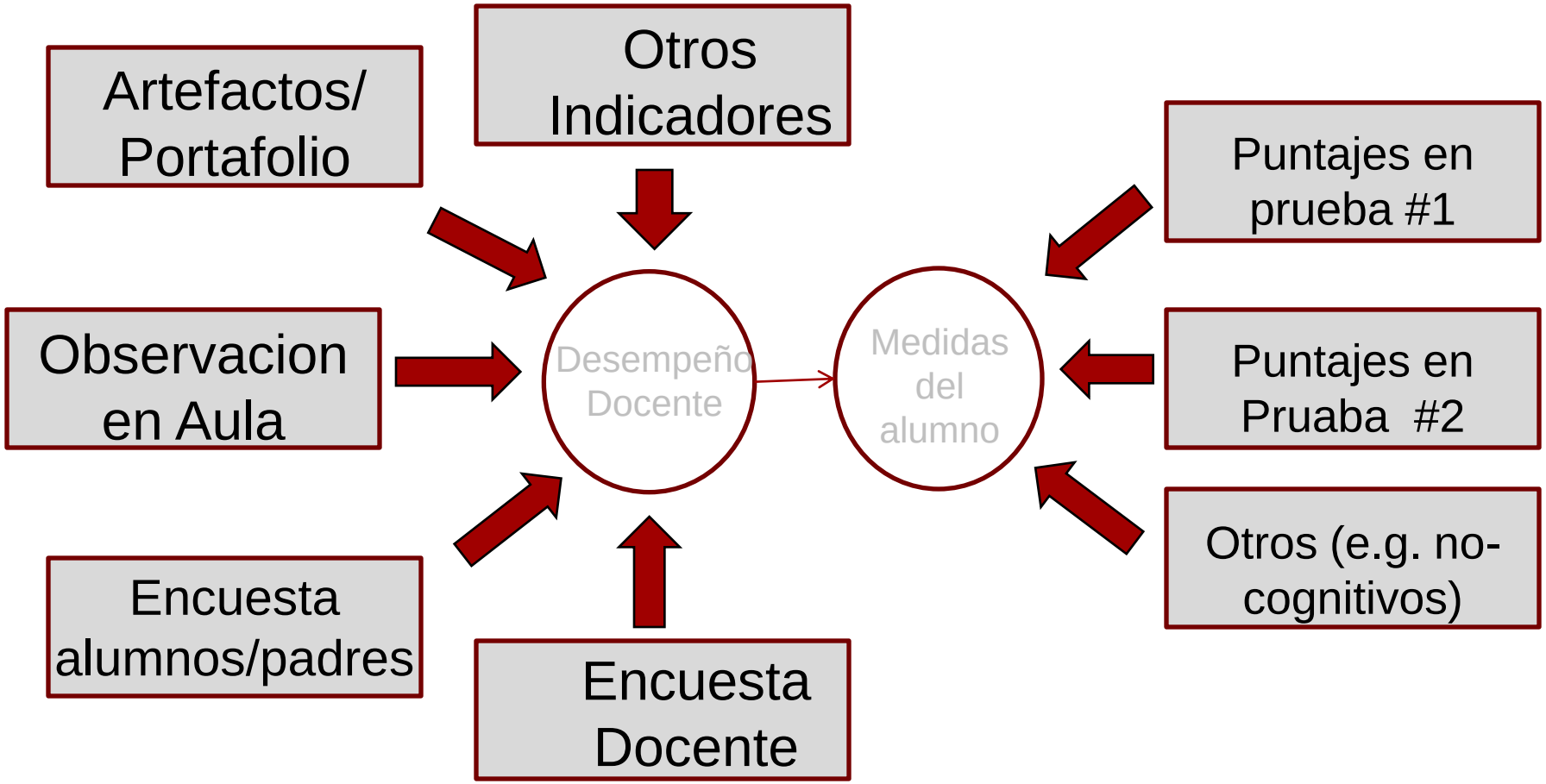


Como ponderar el criterio?

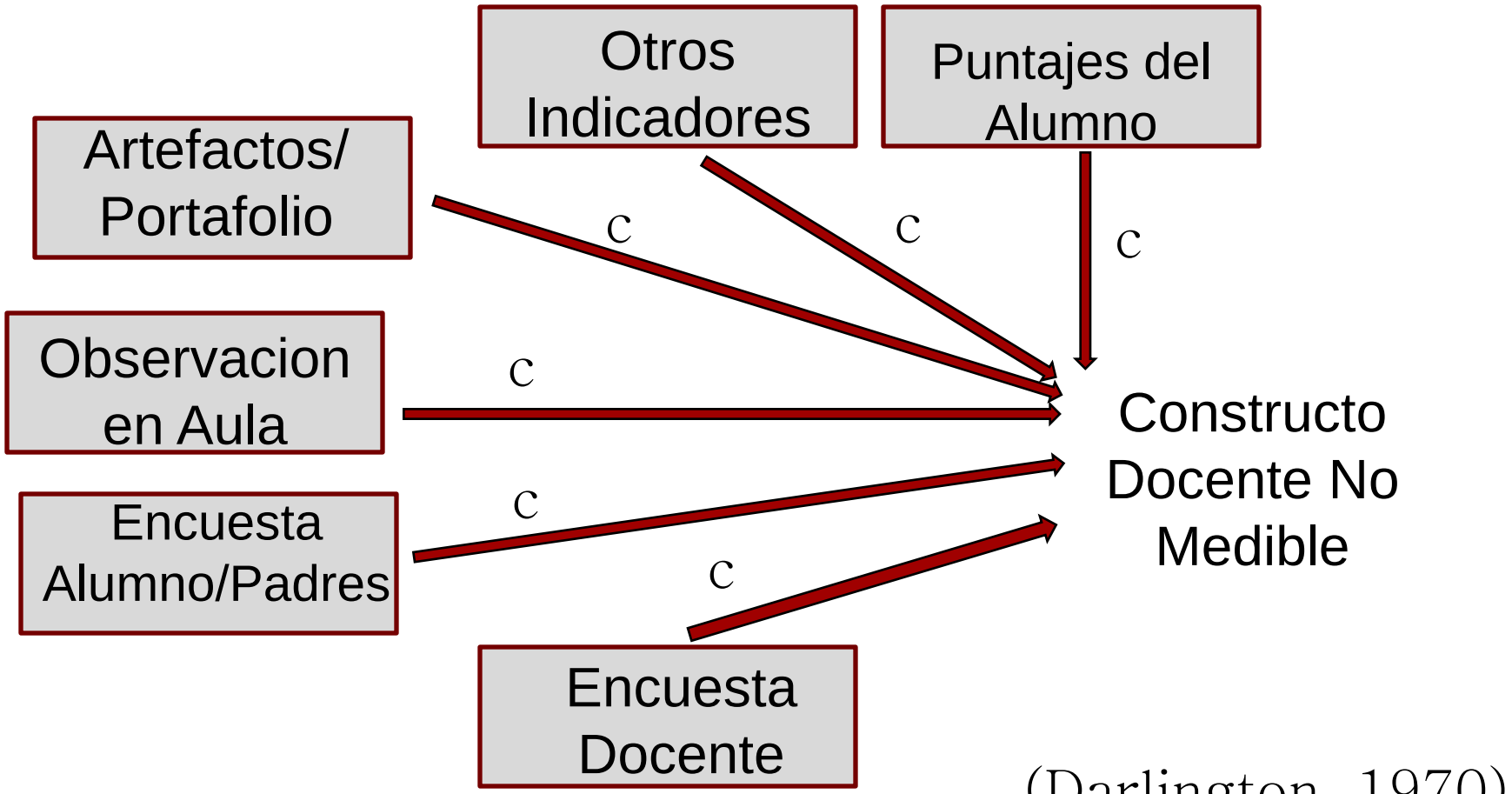
Modelo de Combinación 4b (Compensatorio): Puntajes como indicador y criterio (MET)



Modelo de Combinación 5 (Compensatorio): SEM/Correlaciones *Canónicas*



Modelo de Combinación 6 (Compensatorio): Criterio no medible, pesos teóricos



(Darlington, 1970)

Estudio Empírico:

Comparación de Modelos de Combinación de Indicadores en Evaluación Docente

El estudio: Indicadores Múltiples y Validez

- *Validez* concierne inferencias y usos específicos.
- En la evaluación docente estos asumen:
  - Los indicadores representan adecuadamente el constructo de interés (*desempeño docente*)
  - Es posible clasificar con precisión a los docentes con referencia a ese constructo

- Preguntas de Investigación:
  - Se producen inferencias similares sobre los docentes al usar diferentes modelos de combinación de indicadores?
    - Precisión en la clasificación de docentes
    - Clasificaciones y ordenamientos de los docentes

El estudio: Metodología

- Datos de MET (Measures of Effective Teaching)
  - N=389 docentes de 4to y 5to grados (2010)
  - Seis puntajes *reales* (T) 3 en Ingles y 3 en Matemáticas:
    - VAM (Valor Agregado); FFT (Observación, Danielson);

    SSC (Encuesta de Alumnos Tripod, 7c's)
- Simulación
  - 5000 réplicas (puntajes observados, O)
  - Distribución Normal Multivariada;
    - Media=T; S.D. de MET (Mihaly et. al; 2013, p.24)
      - FFT: 3 observadores, 4 observaciones
      - VAM, SSC: Tamaño promedio de aula

El estudio: Metodología

- Estimar indicadores promedio (Ingles-Matem.)
  - Tres puntajes T por docente/aula
- Reglas de decisión aplicadas a puntajes T y O
  - Criterio: Percentil 10, 50, 75,90 para cada indicador…
  - O el compuesto ponderado en modelos compensatorios
- *Precisión*
  - Acuerdo entre puntajes T-O para cada modelo
  - Kappa, Falsos positivos y negativos, Acuerdo Marginal
- *Consistencia*
  - *Acuerdo* entre decisiones con distintos modelos
  - Correlaciones Pearson

El estudio: Indicadores Múltiples y Validez

- Comparamos 8 modelos de combinación :
  - Complementario 1 (*Pasar* 1 de 3 indicadores)
  - Complementario 2 (*Pasar* 2 de 3 indicadores)
  - Conjuntivo (*Pasar* los 3 indicadores)
  - Seis tipos de modelo compensatorio (ponderado):
    - Pesos unitarios o uniformes (1/1/1 o 33%/33%/33%)
    - Análisis Factorial (29%/33%/36%)
    - Predicción Optima (81%/17%/2% → MET)
    - Confiabilidad Optima (19%/29%/50%)
    - Teórico/Consenso 1 (50%/25%/25% → MET)
    - Teórico/Consenso 2 (35%/25%/40% → DC, LAUSD)

Resultados: Correlación entre Medidas

	Valor Agregado (SVA)		Observacion (FFT)		Encuesta de Alumnos (SSC)	
	Math	English	Math	English	Math	English
SVA_Math	**1**					
SVA_ELA	**0.576**	**1**				
FFT_Math	**0.161**	**0.047**	**1**			
FFT_ELA	**0.153**	**0.106**	**0.468**	**1**		
SSC_Math	**0.128**	**0.201**	**0.303**	**0.204**	**1**	
SSC_ELA	**0.178**	**0.215**	**0.238**	**0.145**	**0.608**	**1**

	Average over Subjects					
	Value Added	Observation	Student Survey			
SVA	**1**					
FFT	**0.160**	**1**				
SSC	**0.222**	**0.292**	**1**			

Resultados: Precisión (Top 10%)

	Complementario		Conjuntivo	Compensatorio (Compuesto Ponderado)					
								Pesos Teóricos (de Consenso)	
	Pasar uno	Pasar dos	Pasar tres	Unidad	FA	Conf. Optima	Pred. Optima	50/25/25	35/40/25
Proporción Pases	**25.4%**	**3.85%**	**0.5%**	**10%**					
Acuerdo (Precisión)	**67.8**	**88.5**	**98.52**	**85.88**	**85.64**	**83.81**	**86.48**	**86.78**	**86.71**
% Falsos Neg.	**5.03**	**1.60**	**0.32**	**3.19**	**3.30**	**3.48**	**3.05**	**3.05**	**3.28**
% Falsos Pos.	**27.1**	**9.85**	**1.15**	**10.92**	**11.05**	**12.69**	**10.36**	**10.16**	**9.99**
% Acuerdo Marginal (M)	**80.2**	**58.6**	**37.30**	**68.17**	**67.05**	**65.19**	**68.60**	**69.55**	**67.20**
% Acuerdo Marginal (NM)	**63.5**	**89.7**	**98.84**	**87.86**	**87.71**	**85.88**	**88.47**	**88.70**	**88.89**

Resultados: Precisión (Top 25%)

	Complementario		Conjuntivo	Compensatorio (Compuesto Ponderado)					
								Pesos Teóricos (de Consenso)	
	Pasar uno	Pasar dos	Pasar tres	Unidad	FA	Conf. Optima	Pred. Optima	50/25/25	35/40/25
Proporción Pases	**52.1%**	**18.5%**	**2.3%**	**25%**					
Acuerdo (Precisión)	**69.4**	**77.6**	**94.42**	**78.9**	**78.7**	**76.83**	**78.18**	**78.4**	**80.0**
% Falsos Neg.	**7.48**	**6.63**	**1.38**	**7.17**	**7.30**	**7.54**	**6.63**	**6.52**	**7.37**
% Falsos Pos.	**23.1**	**15.6**	**4.15**	**13.8**	**13.9**	**15.61**	**15.18**	**15.0**	**12.5**
% Acuerdo Marginal (M)	**85.66**	**64.0**	**40.14**	**71.2**	**70.6**	**69.72**	**73.39**	**73.8**	**70.1**
% Acuerdo Marginal (NM)	**51.6**	**80.8**	**95.75**	**81.5**	**81.4**	**79.19**	**79.77**	**79.9**	**83.2**

Resultados: Precisión (Top 50%)

	Complementario		Conjuntivo	Compensatorio (Compuesto Ponderado)					
								Pesos Teóricos (de Consenso)	
	Pasar uno	Pasar dos	Pasar tres	Unidad	FA	Conf. Optima	Pred. Optima	50/25/25	35/40/25
Proporción Pases	**79.9%**	**50.1%**	**18.2%**	**50%**					
Acuerdo (Precisión)	**82.3**	**71.5**	**82.4**	**76.2**	**76**	**74.4**	**74.8**	**76.4**	**76.7**
% Falsos Neg.	**39.9**	**43.1**	**35.8**	**52.5**	**52**	**47.7**	**48.8**	**52.8**	**53.6**
% Falsos Pos.	**7.19**	**13.8**	**11.05**	**13.0**	**12.9**	**13.20**	**12.51**	**12.7**	**12.3**
% Acuerdo Marginal (M)	**11.01**	**13.8**	**6.88**	**10.7**	**11.0**	**12.38**	**12.68**	**10.7**	**10.9**
% Acuerdo Marginal (NM)	**91.61**	**70.8**	**41.21**	**73.8**	**74.0**	**74.90**	**74.91**	**74.3**	**75.2**

Resultados: Precisión x puntos de corte

	Complementario		Conjuntivo	Compensatorio (Compuesto Ponderado)					
								Pesos Teóricos (de Consenso)	
Punto de Corte	Pasar uno	Pasar dos	Pasar tres	Unidad	FA	Conf. Optima	Pred. Optima	50/25 /25	35/40 /25
10%	**67.8**	**88.5**	**98.52**	**85.88**	**85.6**	**83.81**	**86.48**	**86.78**	**86.71**
25%	**69.4**	**77.6**	**94.42**	**78.9**	**78.7**	**76.83**	**78.18**	**78.4**	**80.0**
50%	**82.3**	**71.5**	**82.4**	**76.2**	**76**	**74.4**	**74.8**	**76.4**	**76.7**
90%	**98.7**	**89.6**	**70.31**	**87.4**	**87.3**	**85.1**	**85.2**	**87.2**	**88.7**

Resultados. Acuerdo entre modelos(Top 10%)

Observed

Complem. 1+	Complem. 2+	Conjuntivo	Unidad	FA	Conf. Optima	Pred. Optima	50/25/25	35/40/25
1	0.646	0.538	0.699	0.699	0.713	0.695	0.692	0.690
0.784	1	0.892	0.901	0.899	0.879	0.868	0.83	0.907
0.751	0.967	1	0.836	0.836	0.821	0.841	0.801	0.846
0.846	0.933	0.905	1	0.986	0.929	0.854	0.824	0.964
0.846	0.933	0.905	0.990	1	0.942	0.843	0.811	0.953
0.845	0.928	0.904	0.964	0.974	1	0.805	0.878	0.896
0.846	0.918	0.905	0.990	0.928	0.917	1	0.888	0.858
0.846	0.923	0.905	0.974	0.964	0.943	0.959	1	0.842
0.846	0.933	0.905	0.990	0.979	0.953	0.928	0.969	1

True

Resultados. Acuerdo entre modelos (Top 25%)

Observed

Complem. 1+	Complem. 2+	Conjuntivo	Unidad	FA	Conf. Optima	Pred. Optima	50/25/25	35/40/25
1	0.597	0.372	0.637	0.636	0.651	0.654	0.639	0.620
0.663	1	0.776	0.866	0.864	0.839	0.797	0.766	0.873
0.501	0.838	1	0.735	0.735	0.721	0.715	0.695	0.752
0.717	0.900	0.774	1	0.981	0.906	0.788	0.760	0.950
0.717	0.900	0.774	0.990	1	0.924	0.772	0.743	0.936
0.722	0.889	0.773	0.917	0.922	1	0.724	0.835	0.861
0.707	0.828	0.774	0.835	0.830	0.778	1	0.850	0.790
0.712	0.884	0.774	0.943	0.933	0.866	0.882	1	0.784
0.715	0.902	0.776	0.972	0.967	0.889	0.828	0.941	1

True

Resultados. Acuerdo entre modelos (Top 50%)

Observed

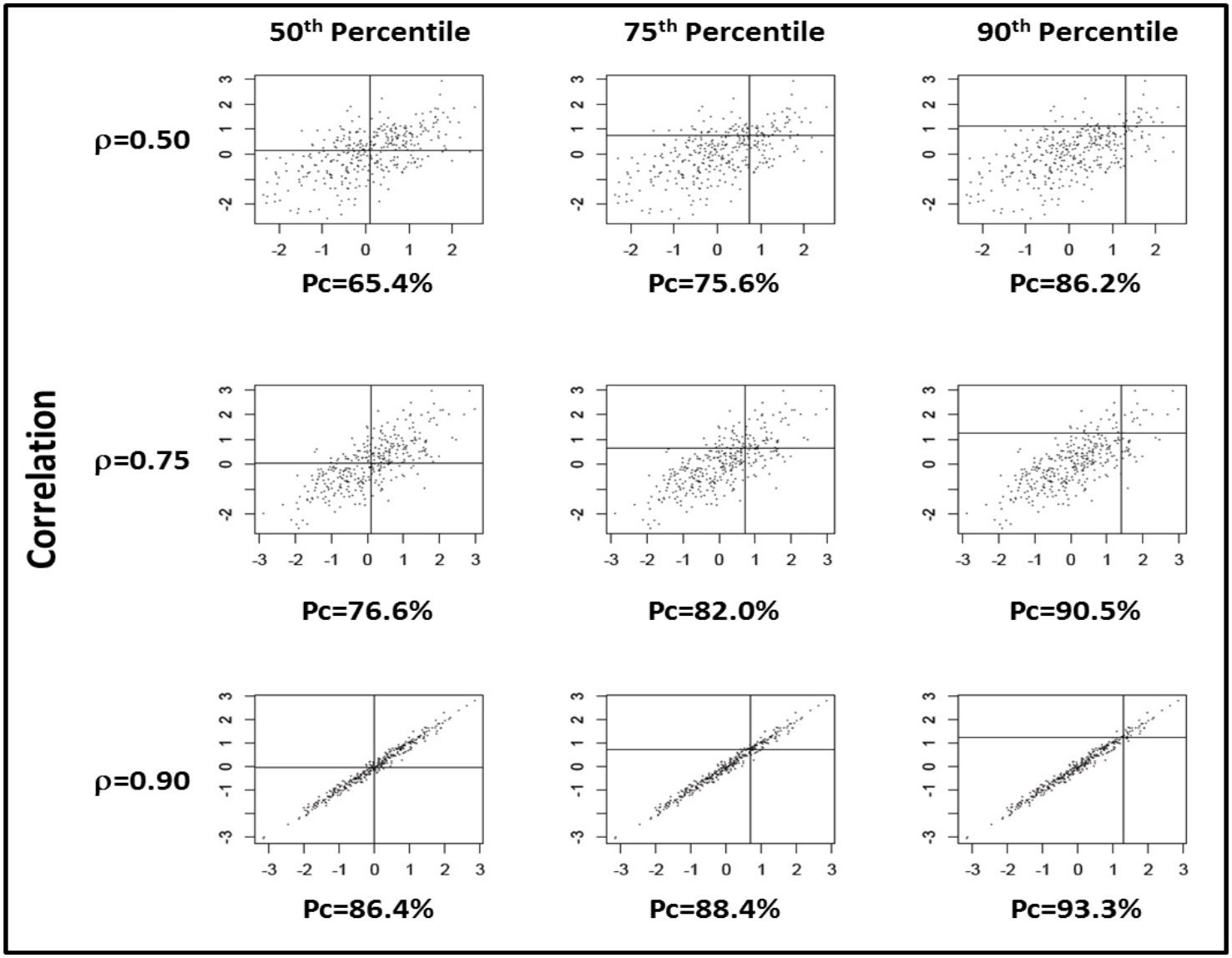
Complem. 1+	Complem. 2+	Conjuntivo	Unidad	FA	Conf. Optima	Pred. Optima	50/25 /25	35/40 /25
1	0.650	0.302	0.634	0.638	0.662	0.658	0.617	0.642
0.502	1	0.651	0.853	0.850	0.755	0.777	0.744	0.855
0.236	0.471	1	0.668	0.664	0.622	0.644	0.657	0.659
0.499	0.717	0.473	1	0.978	0.744	0.762	0.737	0.945
0.498	0.727	0.473	0.979	1	0.724	0.744	0.717	0.930
0.689	0.706	0.473	0.856	0.779	1	0.849	0.920	0.772
0.499	0.562	0.473	0.568	0.547	0.434	1	0.835	0.763
0.499	0.727	0.473	0.825	0.804	0.681	0.732	1	0.763
0.499	0.727	0.473	0.886	0.876	0.753	0.578	0.835	1

True

Resultados.
Correlaciones entre Modelos Compensatorios

	Compensatorio (Compuestos Ponderados)					
True\Observed	Unidad	FA	Conf. Optima	Pred. Optima	Teoria/Consenso 50/25/25	Teoria/Consenso 35/40/25
Unidad	1	0.997	0.947	0.734	0.956	0.985
FA	0.998	1	0.965	0.693	0.935	0.975
Confiabilidad Optima	0.960	0.974	1	0.568	0.838	0.885
Prediccion Optima	0.775	0.741	0.639	1	0.896	0.735
Consenso (50 25 25)	0.965	0.949	0.874	0.910	1	0.957
Consenso (35 40 25)	0.991	0.984	0.921	0.771	0.963	1

Consistencia en clasificación con distintos puntos de corte y correlaciones



Discusión: Indicadores Múltiples y Precisión

- La precisión depende del modelo utilizado para combinar indicadores…
  - Mayor con modelos Conjuntivos o Compensatorios
  - Menor con modelos Complementarios
- De la confiabilidad de los indicadores
  - Un indicador de baja calidad puede influenciar resultados
- Y del punto de corte…
  - Mayor en los extremos de la distribución
- Este es un escenario optimista
  - Datos completos, tamaño de aula medio, observaciones *confiables*, punto de corte único
  - **La precisión con certeza será menor en la práctica**

Discusión: Indicadores Múltiples y Validez

- Acuerdo exacto entre modelos (clasificación)
  - 30% - 90% según el par de modelos de que se trate
- Correlación entre modelos compensatorios
  - 0.50-0.99 Es alta esta correlación?
    - Las inferencias sobre maestros individuales aun pueden diferir significativamente entre modelos (!)
    - 25%-98% de varianza compartida
    - 65% a 95% de acuerdo exacto
- Este es un escenario optimista:
  - Sólo se usa un punto de corte

Discusión:
Indicadores Múltiples y Validez

- Inferencias inestables entre- e intra-modelos
- Cual modelo es preferible?
  - Que tipo de error se quiere minimizar?
  - Con respecto a que constructo? (multi/unidimensional)
- Como se debe crear un índice ponderado?
  - No se pueden determinar ponderaciones empíricas sin un criterio ultimo o de alta calidad
  - La precisión o poder predictivo de un modelo no son evidencia suficiente de validez
  - La ponderación de lógica circular es sólo eso (Rothstein, 2013)
  - La *ponderación teórica o de consenso* será la norma

Modelo de Combinación 0: No Combinar (!)

- No es obligatorio usar índices sintéticos:
  - Se pierde información para usos formativos
  - Propiedades psicométricas desconocidas
- No "combinar indicadores" , si no usar "una combinación de indicadores"
  - Usados individualmente se conoce mejor su precisión (y calidad) para propósitos específicos
    - *e.g desarrollo profesional, incentivos, mejora*
  - Y en conjunto para juicios sumativos si se considera adecuado (Mehrens, 1989; *Brookhart, 2009)*

Discusión: Indicadores Múltiples y Validez

- Se debe examinar la validez con dos enfoques:
  - A. Desde el punto de vista de la medición.
    - Argumento de validez para inferencias/usos especificos
    - Basado en fuentes de evidencia (AERA, APA, NCME, 1999)
      - Soporte Teórico y representación de constructo
      - Consistencia (Confiabilidad)
      - Patrones esperados de correlación (indicadores/criterios)
      - Efectos y consecuencias de uso (previstas y no).
  - B. Desde el punto de vista de política pública. Evaluar la evaluación docente como programa (Robin, Stuart, Zanutto, 2004)

- Consideraciones Finales

Conclusión:
Porque Evaluar a los Docentes?

- Hay buenas razones para hacerlo
  - Cultura de responsabilidad, reflexión y mejora
  - Informar el desarrollo profesional y mejorar la práctica docente
  - Hacer prioritario el aprendizaje de los alumnos
- Sin embargo:
  - La simple combinación de indicadores falibles no resulta en inferencias *mejores* o menos falibles.
    - Pero si en inferencias mas complejas
  - El uso conjunto de indicadores presenta retos técnicos pero también prácticos y políticos

Conclusión:
Porque Evaluar a los Docentes?

- *"El futuro de los niños de (inserte país) esta en juego"*, si, pero esto nos llama a proceder con cautela
  - Desarrollar buenos indicadores requiere tiempo,
  - Mas aun implementar sistemas de evaluación solidos basados en indicadores múltiples,
  - Y evaluar los efectos de uso en resultados relevantes
- Esta complejidad es de difícil comunicación a autoridades, docentes, y prensa
  - *Cualquier evaluación* no es necesariamente "mejor" que ninguna (o que la que tenemos)
  - El riesgo de decisiones injustas, y resultados no deseados y contraproducentes es real

Muchas Gracias

jfmtz@ucla.edu