Trafność testów egzaminacyjnych

Artur Pokropek, Tomasz Żółtak

IFiS PAN

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

IBE EDUKACYJNA WARTOŚĆ DODANA

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Plan prezentacji

- EWD i trafność testów egzaminacyjnych
- Pięć postulatów trafności dla skal pomiarowych
- Wskaźniki egzaminacyjne a wyniki testów PISA
- Zewnętrzne kryteria trafności
- Dyskusja

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
IBE EDUKACYJNA WARTOŚĆ DODANA
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

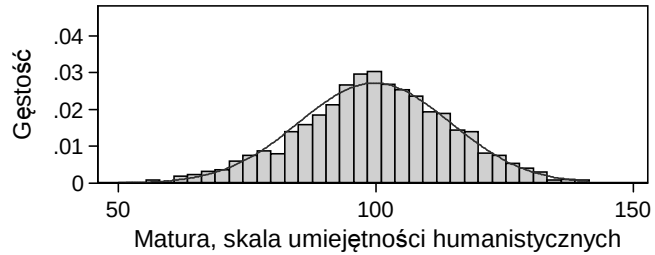
EWD i trafność testów egzaminacyjnych

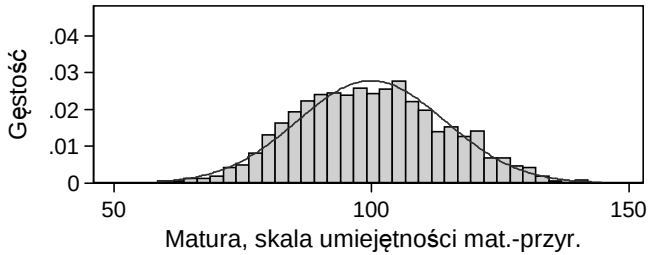
- Aby oszacować relatywny przyrost wiedzy uczniów w szkołach, niezbędne są pomiary umiejętności, w momencie gdy rozpoczynają i kończą oni naukę.
- Warunkiem koniecznym dla trafności EWD jest trafność tych pomiarów, czyli stopień, w jakim nadają się one do szacowania relatywnego przyrostu wiedzy.
- Nawet najbardziej wyrafinowane metody statystycznie nic nie pomogą, jeżeli podstawa szacowania EWD okaże się niedostosowana do celów, jakie są przed tym wskaźnikiem stawiane.

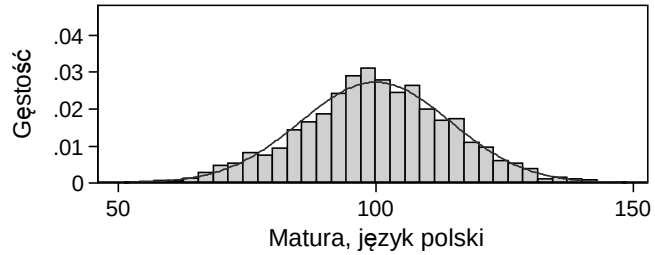
Pięć postulatów trafności dla skal pomiarowych

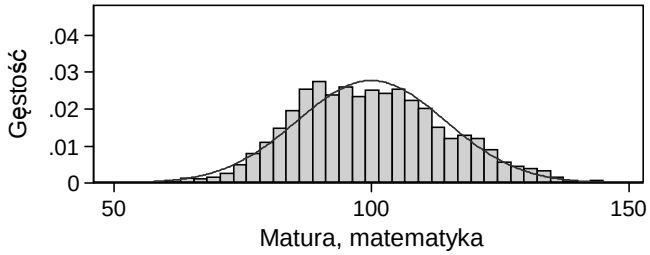
- W literaturze poświęconej trafności wskaźnika EWD pojawia się, pięć postulatów odnoszących się do skal pomiarowych wykorzystywanych w konstrukcji wskaźnika EWD (Linn 2008, Reckase 2008). Skale wykorzystywane w modelowaniu EWD powinny być zatem:
  - (1) mierzone na skali interwałowej;
  - (2) wysoce rzetelne ;
  - (3) nieobciążone błędami systematycznymi;
  - (4) mierzyć ten sam konstrukt (dla jednej miary EWD) ;
  - (5) skonstruowane z zachowaniem odpowiedniej reprezentacji treści nauczania.

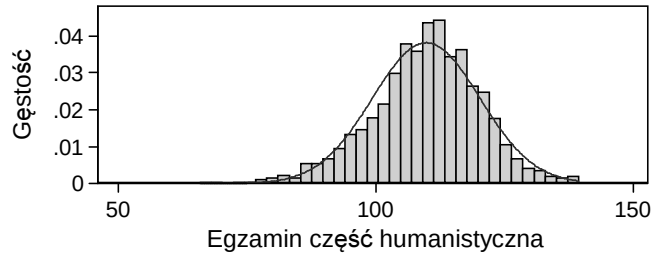
Wymóg interwałowego charakteru skal

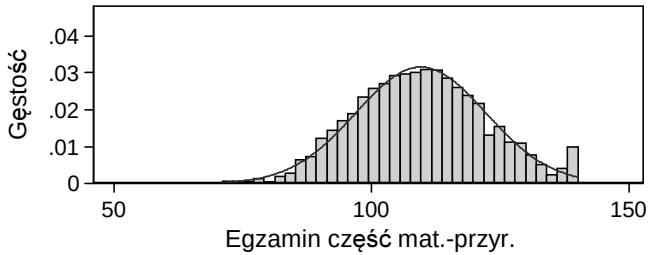












KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
IBE EDUKACYJNA WARTOŚĆ DODANA
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Wymóg zadowalającej rzetelności

- Rzetelność jest fundamentalną własnością testu, która decyduje o precyzji pomiaru umiejętności uczniów. Jeżeli test nie jest rzetelny, precyzja oceny umiejętności ucznia na jego podstawie jest niewielka. Niska rzetelność testów łamie założenia modelowania wykorzystywanego w konstrukcji EWD i może prowadzić do poważnych błędów w wyliczaniu wskaźnika (por. Pokropek 2010, 2013).
- Istnieje kilka miar rzetelności testu, jednak zdecydowanie najpopularniejszą jest Alfa Cronbacha, stosowana w Klasycznej Teorii Testów (KTT).
- Jako alternatywną miarę rzetelności, posłużyliśmy się więc oszacowaniem na podstawie wyników dwuparametrycznego modelu IRT, estymowanego w odniesieniu do danych z egzaminu gimnazjalnego.

Wymóg zadowalającej rzetelności

Egzamin	Wskaźnik rzetelności	
	Alfa	IRT
Matura wskaźnik humanistyczny	-	0,814
Matura wskaźnik mat.-przyr.	-	0,944
Matura język polski	-	0,811
Matura matematyka	-	0,934
EG wskaźnik humanistyczny	0,813	0,807
EG wskaźnik mat.-przyr.	0,899	0,911

Wymóg braku błędów systematycznych (zróżnicowane funkcjonowanie zadań)

- Test nie będzie trafny wtedy, gdy będzie mierzył coś, czego mierzyć nie powinien. Test oprócz zakładanej umiejętności szkolnej może na przykład „mierzyć płeć” ucznia. Dzieje się tak, gdy zadania w teście są stronnicze płciowo, tj. prawdopodobieństwo poprawnej odpowiedzi zależy nie tylko od poziomu umiejętności ucznia, ale również od płci osoby rozwiązującej zadanie.
- Do innych należą także: pochodzenie ucznia, status społeczno-ekonomiczny rodziny, wielkość miejscowości etc.
- W trafnym pomiarze cechy te nie powinny mieć znaczenia dla trudności zadania przy kontroli poziomu umiejętności ucznia.

Wymóg braku błędów systematycznych

- Do zbadania zróżnicowanego funkcjonowania zadań wykorzystywanych w pomiarach umiejętności, wykorzystywanych w modelowaniu EWD wybrana została prosta, ale efektywna metoda oparta na modelu regresji dla zmiennych porządkowych.
- Najpierw należy stworzyć wskaźnik umiejętności uczniów.
- Następnie estymowane są modele regresji dla zmiennych o charakterze porządkowym – osobno dla każdego zadania – w których zmienną wyjaśnianą jest poziom rozwiązania zadania (ile punktów zostało zdobytych przez ucznia), a zmiennymi wyjaśniającymi są: poziom umiejętności ucznia i zmienna, co do której mamy przypuszczenie, iż może być odpowiedzialna za stronniczy charakter zadania.

$$\text{ologit}(\pi_{ki}) = \beta_0 + \beta_1\theta' + \beta_2 G$$

Zróżnicowane funkcjonowanie zadań ze względu **<u>płeć uczniów</u>** (zadania faworyzujące dziewczęta zostały oznaczone symbolem „+”, natomiast zadania faworyzujące chłopców symbolem „-”.). B oznacza mały istotny dif, C duży istotny dif.

Egzamin	liczba zadań ogółem	liczba zadań z dif				% zadań			
		B+	C+	B-	C-	B+	C+	B-	C-
Matura wskaźnik humanistyczny	124	3	5	0	13	2%	4%	0%	10%
Matura wskaźnik mat.-przyr.	275	6	15	14	20	2%	5%	5%	7%
Matura język polski	24	0	1	0	0	0%	4%	0%	0%
Matura matematyka	45	2	0	3	1	4%	0%	7%	2%
EG wskaźnik humanistyczny	47	0	0	2	5	0%	0%	4%	11%
EG wskaźnik mat.-przyr.	49	5	1	3	6	10%	2%	6%	12%

Zróżnicowane funkcjonowanie zadań ze względu na **lokalizację szkoły** (zadania faworyzujące uczniów szkół wiejskich zostały oznaczone symbolem „+”, natomiast zadania faworyzujące uczniów szkół miejskich symbolem „-”.). B oznacza mały istotny dif, C duży istotny dif.

Egzamin	liczba zadań ogółem	liczba zadań z dif				% zadań			
		B+	C+	B-	C-	B+	C+	B-	C-
Matura wskaźnik humanistyczny	124	0	8	4	5	0%	6%	3%	4%
Matura wskaźnik mat.-przyr.	275	3	13	0	7	1%	5%	0%	3%
Matura język polski	24	1	0	2	1	4%	0%	8%	4%
Matura matematyka	45	1	0	0	1	2%	0%	0%	2%
EG wskaźnik humanistyczny	47	2	0	6	3	4%	0%	13%	6%
EG wskaźnik mat.-przyr.	49	0	0	0	0	0%	0%	0%	0%

Zróżnicowane funkcjonowanie zadań ze względu na **status społeczny rodziny** (zadania faworyzujące grupę uczniów o niskim SES zostały oznaczone symbolem „+”, natomiast zadania faworyzujące grupę uczniów o wysokim SES symbolem „-”.). B oznacza mały istotny dif, C duży istotny dif.

Egzamin	liczba zadań ogółem	liczba zadań z dif				% zadań			
		B+	C+	B-	C-	B+	C+	B-	C-
Matura wskaźnik humanistyczny	124	2	0	2	0	2%	0%	2%	0%
Matura wskaźnik mat.-przyr.	275	0	0	4	4	0%	0%	1%	1%
Matura język polski	24	0	0	0	0	0%	0%	0%	0%
Matura matematyka	45	0	0	0	0	0%	0%	0%	0%
EG wskaźnik humanistyczny	47	0	0	0	0	0%	0%	0%	0%
EG wskaźnik mat.-przyr.	49	0	0	0	0	0%	0%	0%	0%

Wymóg pomiaru tego samego konstruktu

Egzamin gimnazjalny	Matura	Wskaźnik EWD
część humanistyczna	Wskaźnik humanistyczny	EWD humanistyczne
	Język polski	EWD język polski
część matematyczno przyrodnicza	Wskaźnik matematyczno-przyrodniczy	EWD matematyczno-przyrodnicze
	Matematyka	EWD matematyka

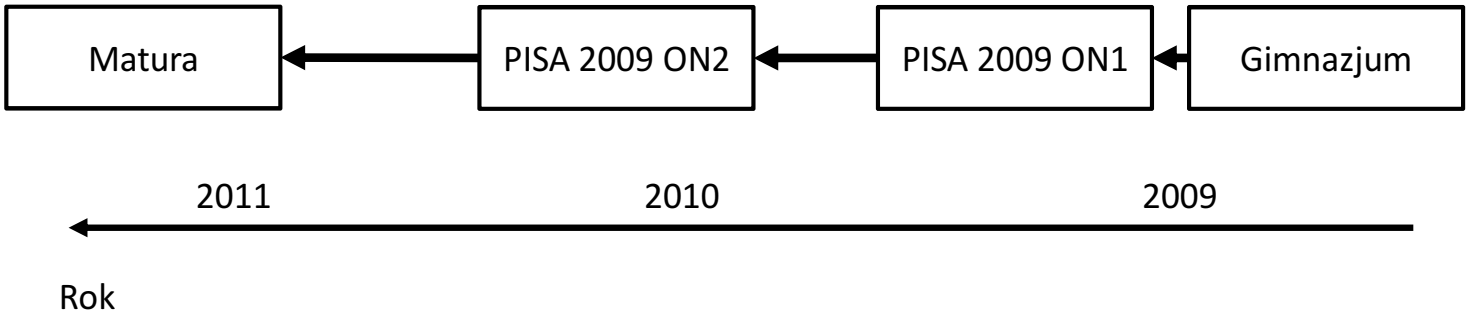
Wymóg pomiaru tego samego konstruktu

Egzamin	Egzamin					
	(1)	(2)	(3)	(4)	(5)	(6)
Matura wskaźnik humanistyczny (1)	1					
Matura wskaźnik mat.-przyr. (2)	0,816 (0,007)	1				
Matura język polski (3)	X	0,487 (0,018)	1			
Matura matematyka (4)	0,488 (0,018)	X	0,453 (0,018)	1		
EG wskaźnik humanistyczny (5)	0,591 (0,017)	0,487 (0,018)	0,547 (0,018)	0,447 (0,019)	1	
EG wskaźnik mat.-przyr (6)	0,575 (0,017)	0,842 (0,007)	0,528 (0,017)	0,822 (0,008)	0,573 (0,016)	1

Wymóg odpowiedniej reprezentacji zadań

- Postulat dużej liczby odpowiedniej jakości zadań w zderzeniu z ograniczeniami w zakresie czasu testowania jednego ucznia jest źródłem poważnego utrudnienia. Dlatego w pomiarach testowych nie będących egzaminami używa się schematów, w którym duża liczba zdań rozwiązywana jest przez ograniczoną populację uczniów tak, że nie wszyscy uczniowie rozwiązują wszystkie zadania (por. OECD 2009).
- Z podobną sytuacją mamy do czynienia w przypadku złożonych wskaźników maturalnych, poprzez które wiedza uczniów oceniana jest na podstawie kilkuset różnych zadań z różnych przedmiotów.
- Taka sytuacja niestety nie jest możliwa w przypadku egzaminu gimnazjalnego, w którym mamy do czynienia z dużo mniejszą liczbą zadań. Jednak korelacje między miarami maturalnymi (opartymi na dużej liczbie zadania) a miarami gimnazjalnymi (opartymi na stosunkowo niewielkiej liczbie zdań) są wysokie, to i reprezentacja treści na egzaminie gimnazjalnym powinna być przynajmniej zadowalająca.

Wskaźniki egzaminacyjne a wyniki testów PISA



Struktura egzaminu gimnazjalnego (poza językiem obcym) oraz testowania w PISA

PISA	Egzamin gimnazjalny	Matura
rozumienie czytanych tekstów (reading)	część humanistyczna	Wskaźnik humanistyczny
		Język polski
biegłość w zakresie przedmiotów przyrodniczych (science)	część matematyczno przyrodnicza	Wskaźnik matematyczno-przyrodniczy
biegłość matematyczna (math)		Matematyka

Macierz korelacji między skalowanymi umiejętnościami PISA 2009 ON1 i egzaminem gimnazjalnym

Test	(1)	(2)	(3)	(4)	(5)
PISA 2009 ON1: math (1)	1				
PISA 2009 ON1: reading (2)	0,889 (0,010)	1			
PISA 2009 ON1: science (3)	0,955 (0,009)	0,925 (0,012)	1		
EG wskaźnik humanistyczny (4)	0,541 (0,012)	0.659 (0,015)	0,529 (0,014)	1	
EG wskaźnik mat.-przyr (5)	0,864 (0,012)	0,641 (0,017)	0.704 (0,011)	0,573 (0,016)	1

Macierz korelacji między skalowanymi umiejętnościami PISA 2009 ON2 i egzaminem maturalnym

Test	(1)	(2)	(3)	(4)	(5)	(6)	(7)
PISA 2009 ON2: math (1)	1						
PISA 2009 ON2: reading (2)	0,806 (0,013)	1					
PISA 2009 ON2: science (3)	0,946 (0,010)	0,865 (0,014)	1				
Matura wskaźnik humanistyczny (4)	0,535 (0,015)	0,645 (0,019)	0,606 (0,025)	1			
Matura wskaźnik mat.-przyr. (5)	0,778 (0,015)	0,577 (0,010)	0,715 (0,016)	0,816 (0,007)	1		
Matura język polski (6)	0,497 (0,027)	0,621 (0,019)	0,558 (0,026)	x	0,487 (0,018)	1	
Matura matematyka (7)	0,791 (0,015)	0,525 (0,020)	0,578 (0,023)	0,488 (0,018)	x	0,453 (0,018)	1

Zewnętrzne kryteria trafności

- Dopełnieniem analizy trafności narzędzi pomiarowych jest analiza relacji między badanym narzędziem a zewnętrznymi kryteriami, które zgodnie z teorią powinny być z nim związane.
- Test Ravena (klasyczny test inteligencji)
- Skala odnosząca się do czytania skonstruowania została na podstawie baterii dziewięciu stwierdzeń:

(1) czytam tylko wtedy, kiedy muszę; (2) czytanie to jedno z moich ulubionych zajęć; (3) lubię rozmawiać z innymi na temat książek; (3) trudno mi jest doczytać książkę do końca; (4) cieszę się, gdy dostaję w prezencie książkę; (4) czytanie to strata czasu; (5) lubię chodzić do księgarni lub biblioteki; (6) czytam tylko po to, aby uzyskać potrzebne informacje; (7) nie mogę czytać dłużej niż parę minut; (8) lubię wyrażać opinie na temat książek; (9) lubię wymieniać się książkami z innymi. Wszystkie pozycje zadawane na 4-punktowej skali: *(a) zdecydowanie się nie zgadzam, (b) nie zgadzam się, (c) zgadzam się, (d) zdecydowanie się zgadzam*.

Trafność egzaminów CKE w świetle innych kryteriów

Egzamin	Korelacje latentne	
	IQ	czytanie
Matura wskaźnik humanistyczny	0,391 (0,023)	0,414 (0,020)
Matura wskaźnik mat.-przyr.	0,614 (0,015)	0,069 (0,021)
Matura język polski	0,389 (0,024)	0,398 (0,021)
Matura matematyka	0,675 (0,018)	0,097 (0,022)
EG wskaźnik humanistyczny	0,323 (0,026)	0,325 (0,022)
EG wskaźnik mat.-przyr.	0,618 (0,021)	0,164 (0,022)

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
IBE EDUKACYJNA WARTOŚĆ DODANA
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Podsumowanie

- Analiza testów wykorzystywanych do konstrukcji maturalnego wskaźnika EWD potwierdza jej trafność.
- Skale wykorzystywane do konstrukcji EWD spełniają wszystkie podstawowe kryteria trafności, wykazując jedynie drobne odstępstwa od wyznaczonych wzorców.
  - Nie ma żadnych silnych dowodów na to, że skale wykorzystywane w EWD maturalnym mają charakter inny niż interwałowy.
  - Każdy z badanych testów odznacza się przynajmniej przeciętną rzetelnością.
  - Analiza zadań egzaminacyjnych sugeruje co prawda możliwość występowania stronniczości zadań, lecz tylko dla płci, co w przypadku modelowania EWD ma marginalny charakter.
  - Wysokie korelacje między miarami na wejściu i na wyjściu świadczą przynajmniej o podobieństwie mierzonych konstruktów
- Korelacje z pomiarem PISA i zewnętrznymi kryteriami pozwalają wierzyć, iż skale, które powstały na podstawie zadań egzaminacyjnych, mierzą to, co było celem pomiaru.

Dziękujemy za uwagę!

Artur Pokropek, Tomasz Żółtak

IFiS PAN

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

IBE EDUKACYJNA WARTOŚĆ DODANA

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY