Liczba zadań a rzetelność testu na przykładzie testów biegłości językowej z języka angielskiego

Tomasz Żółtak
Instytut Badań Edukacyjnych
oraz Uniwersytet Warszawski

Plan wystąpienia

- Testy z punktu widzenia statystyki - psychometria
- Podstawowe założenia Klasycznej Teorii Testu
- Rzetelność w Klasycznej Teorii Testu
- Związek pomiędzy rzetelnością a długością (liczbą zadań) testu
  - W teorii
  - W praktyce
- Podsumowanie

Testy z punktu widzenia statystyki

- Psychometria – dziedzina statystyki
- Statystyka nie mówi nam nic o tym, co mierzy test – aby to stwierdzić trzeba odwołać się do analizy treści zadań.
- Może nam jednak powiedzieć, czy możemy uważać, że to, co test mierzy (czymkolwiek by to nie było), mierzy dobrze.
- Jeśli test *dobrze mierzy* to, w danych opisujących wyniki testu powinniśmy spodziewać się występowania pewnych prawidłowości o charakterze statystycznym.
- Jeśli takich prawidłowości nie znajdujemy, lub są one bardzo słabe, jest to znak, że być może dałoby się test poprawić, aby lepiej mierzył on interesującą nas cechę.

Podstawowe założenia 1

- Przy pomocy testu chcemy badać pewną **cechę, której nie daje się zmierzyć bezpośrednio**:
  - To, co chcemy zmierzyć nie daje się sprowadzić do *sumy* pytań zawartych w teście.
  - Np. *umiejętność posługiwania się językiem angielskim* to nie tylko zdolność do poprawnego poradzenia sobie z konkretnymi zadaniami, ale również:
    - Poradzenia sobie z innymi zadaniami, które potencjalnie mogłyby znaleźć się w teście.
    - Poradzenia sobie w realnych sytuacjach wymagających wykorzystania języka (przybliżanych przez użyte w teście zadania).

Podstawowe założenia 2

- **Poprawne (lub nie) rozwiązanie zadań zawartych w teście nie jest całkowicie pewnym wskaźnikiem posiadania (natężenia) interesującej nas cechy:**
  - Związki pomiędzy mierzoną cechą a rozwiązaniami zadań mają charakter statystyczny (probabilistyczny).
  - Oczywiście chcemy, by te związki były dosyć silne.
  - Poradzenie sobie w konkretnej sytuacji wynika po części również ze zdarzeń, które możemy traktować jako losowe:
    - lepszej lub gorszej predyspozycji danej osoby w danym dniu,
    - lepszego lub gorszego opanowania przez daną osobę akurat tej części materiału, która znalazła się w teście, itp.

Podstawy Klasycznej Teorii Testu

wynik testu = wartość prawdziwa + błąd pomiaru

- Błąd pomiaru ma charakter losowy.
- Dobry test charakteryzuje się niewielkimi błędami pomiaru.
- **Rzetelność:**

$$r = D^2(\textit{wartości prawdziwe}) \,/\, D^2(\textit{wyniki testu})$$

$$1 \geq r \geq 0$$

gdzie $D^2()$ oznacza wariancję (miara zróżnicowania).

Rzetelność mówi, jaka część zróżnicowania wyników testu związana jest z *prawdziwym* zróżnicowaniem badanej cechy a więc, **jak dokładny jest pomiar danym testem.**

Im większa (bliższa 1) rzetelność, tym lepiej.

Szacowanie rzetelności

- Alfa Cronbacha:

$$\alpha_C = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} \sigma_i^2}{\sigma_\Sigma^2}\right)$$

Odpowiednia, gdy dla wszystkich zadań siła związku z badaną cechą jest taka sama.

- Alfa Feldt-Raju:

$$\alpha_{FR} = \frac{1}{1 - \sum_{i=1}^{k}\left(\frac{\sigma_{i\Sigma}}{\sigma_\Sigma^2}\right)^2}\left(1 - \frac{\sum_{i=1}^{k} \sigma_i^2}{\sigma_\Sigma^2}\right)$$

gdzie:

k – liczba zadań

σ^2_i - wariancja wyników i-tego zadania

σ^2_Σ - wariancja wyników całkowitych testu

$\sigma_{i\Sigma}$ - kowariancja wyników i-tego zadania i wyników całkowitych testu

Rzetelność w Klasycznej Teorii Testu

Rzetelność zależy od:

- Liczby zadań w teście – im więcej zadań, tym dokładniejszy powinien być pomiar.
  - *Wzór proroczy* Spearmana-Browna:

    r_k – rzetelność dla k zadań

    r_1 - rzetelność pojedynczego zadania

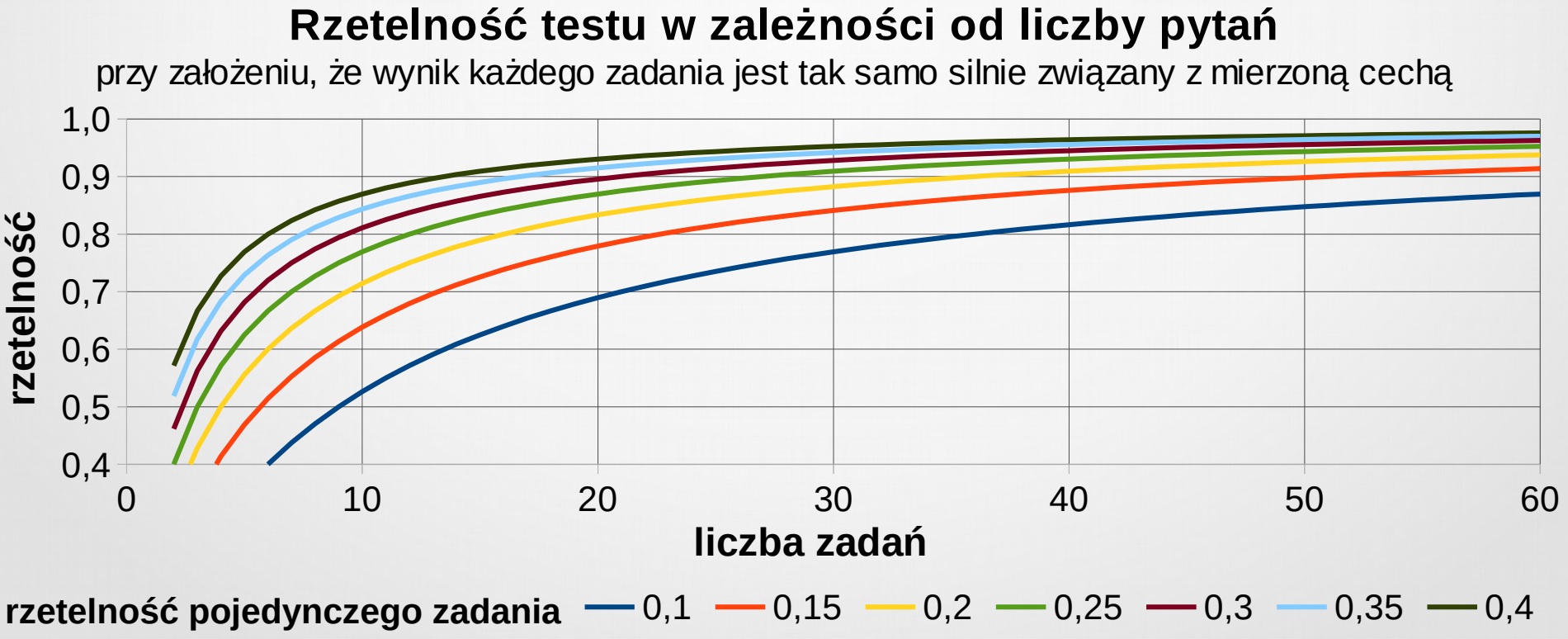
    $$r_k = \frac{k\, r_1}{1 + (k-1)\, r_1}$$

    Odpowiedni, gdy dla wszystkich zadań siła związku z badaną cechą jest taka sama.
- Siły związku pomiędzy wynikiem pojedynczego zadania a natężeniem mierzonej cechy – im większa, tym pomiar dokładniejszy.

Rzetelność w Klasycznej Teorii Testu

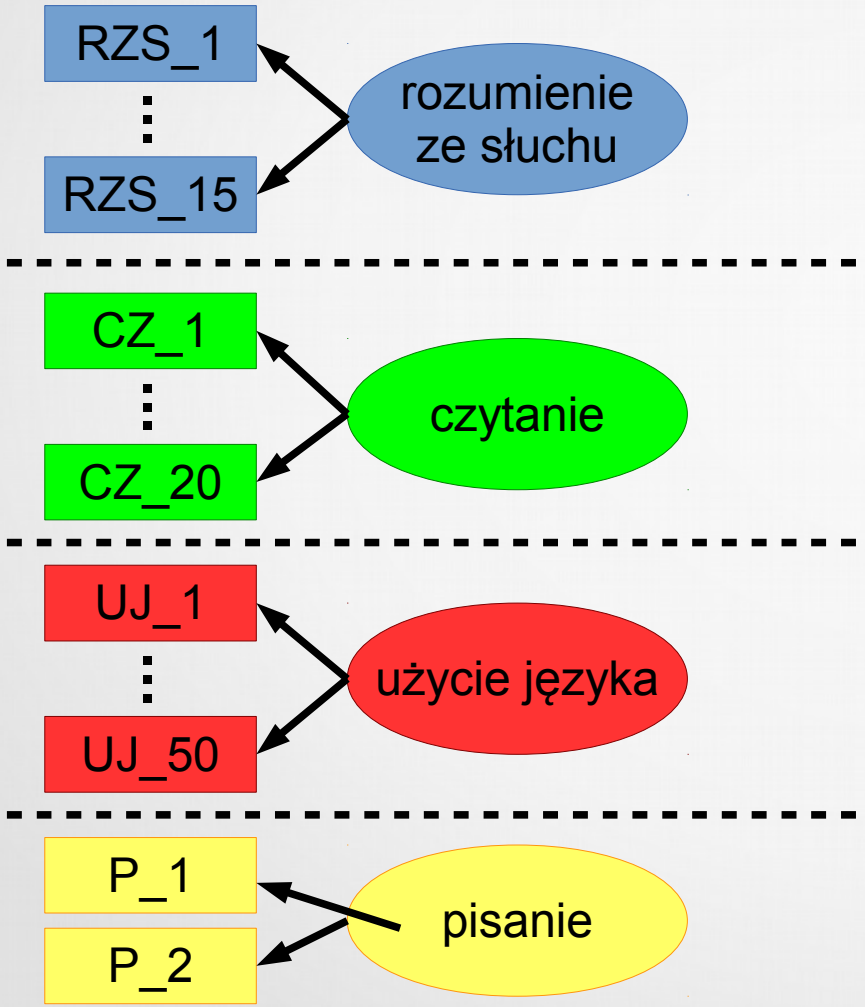
Rzetelność zależy od:

- Liczby zadań w teście – im więcej zadań, tym generalnie dokładniejszy powinien być pomiar.
- Siły związku pomiędzy wynikiem pojedynczego zadania a natężeniem mierzonej cechy – im większa, tym pomiar dokładniejszy.

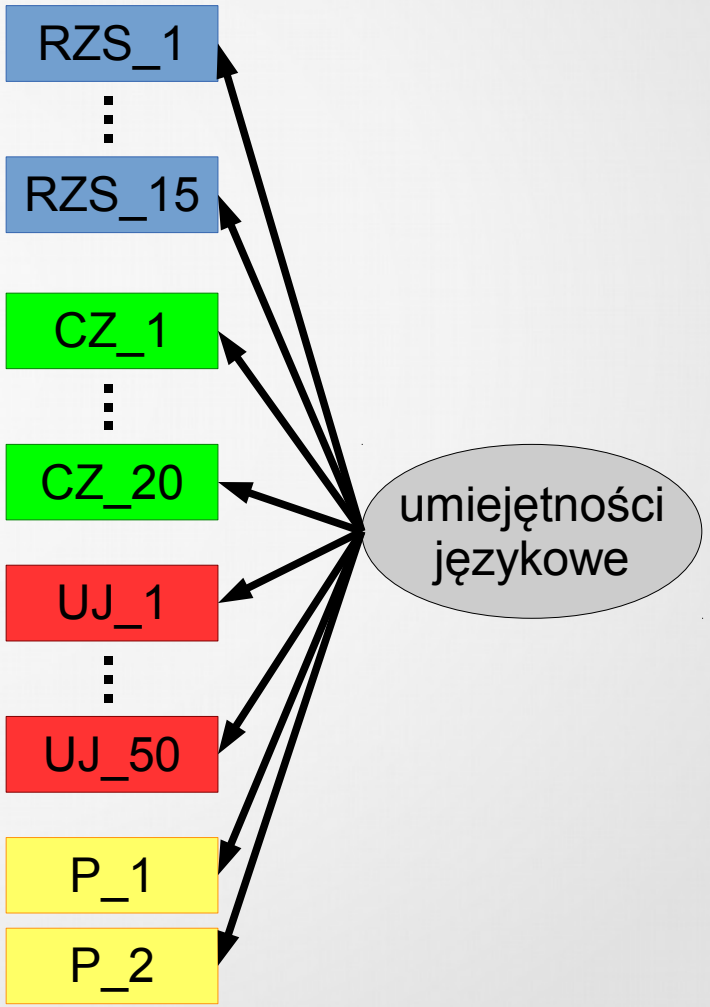


Co jest celem testowania?

Analiza każdej skali oddzielnie?



Analiza wyników całego testu?



Rzetelność a spójność treściowa

Rzetelność zależy od siły związku pomiędzy wynikiem pojedynczego zadania a natężeniem mierzonej cechy – im większa, tym pomiar dokładniejszy.

Przy tym jeśli przyjmujemy, że związki pomiędzy wynikami rozwiązywania poszczególnych zadań są wynikiem tego, że są one wskaźnikami tej samej cechy, będącej przedmiotem badania, to:

- Wyróżnione powyżej zdanie jest równoważne stwierdzeniu, że **rzetelność jest tym większa, im silniejsze są wzajemne związki pomiędzy wynikami rozwiązywania poszczególnych zadań w teście**.
- Z tego powodu mówi się czasem, że rzetelność jest miarą *wewnętrznej spójności* testu – bada, czy zadania mierzą *to samo*.

Rzetelność a spójność treściowa

Jak interpretować niską rzetelność testu?

- Test jest zbyt krótki (zawiera zbyt mało zadań).

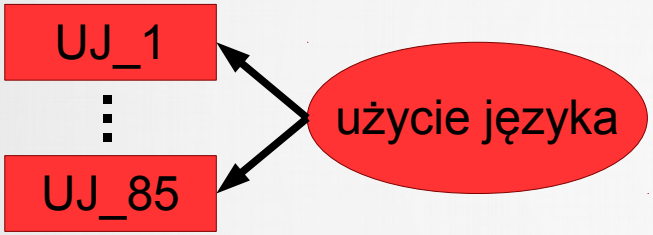
  i/lub

- Różne zadania (być może grupy zadań), które tworzą test, są ze sobą słabo powiązane. To, że pewne z nich rozwiązuje się poprawnie (lub nie) ma niewielki związek z tym, czy poprawnie (lub nie) rozwiązuje się inne.
  - Testy pokrywające szeroki zakres treściowy charakteryzują się więc co do zasady (dla zadanej liczby zadań w teście) niższą rzetelnością niż testy mierzące wąsko zdefiniowane umiejętności.

Rzetelność a spójność treściowa

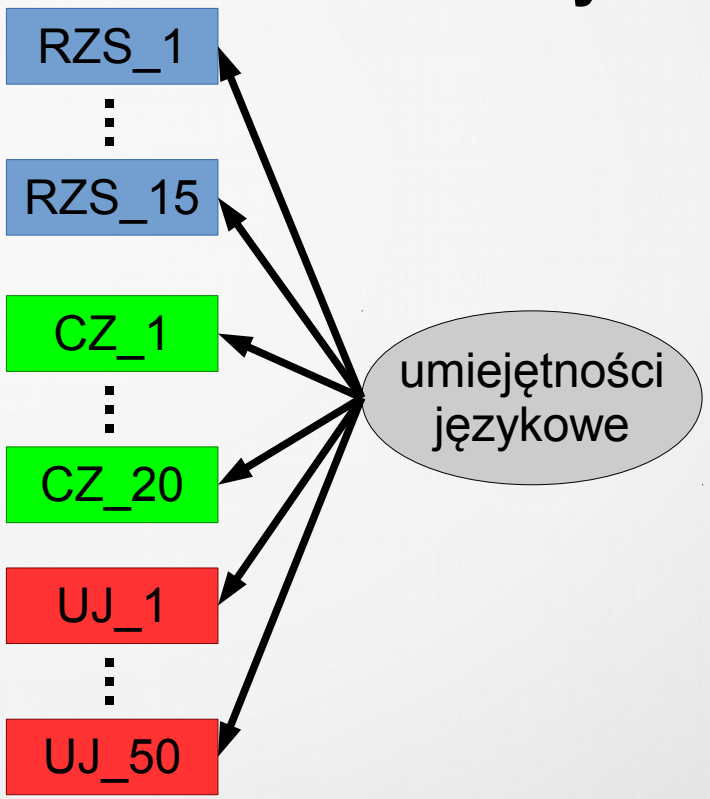
Test A:

85 zadań z jednej dziedziny



Test B:

85 zadań z trzech różnych dziedzin



Możemy się spodziewać, że:

rzetelność A > rzetelność B

Rzetelność a długość testu

Jeśli za punkt wyjścia przyjmiemy długi test pokrywający szeroki zakres treści:

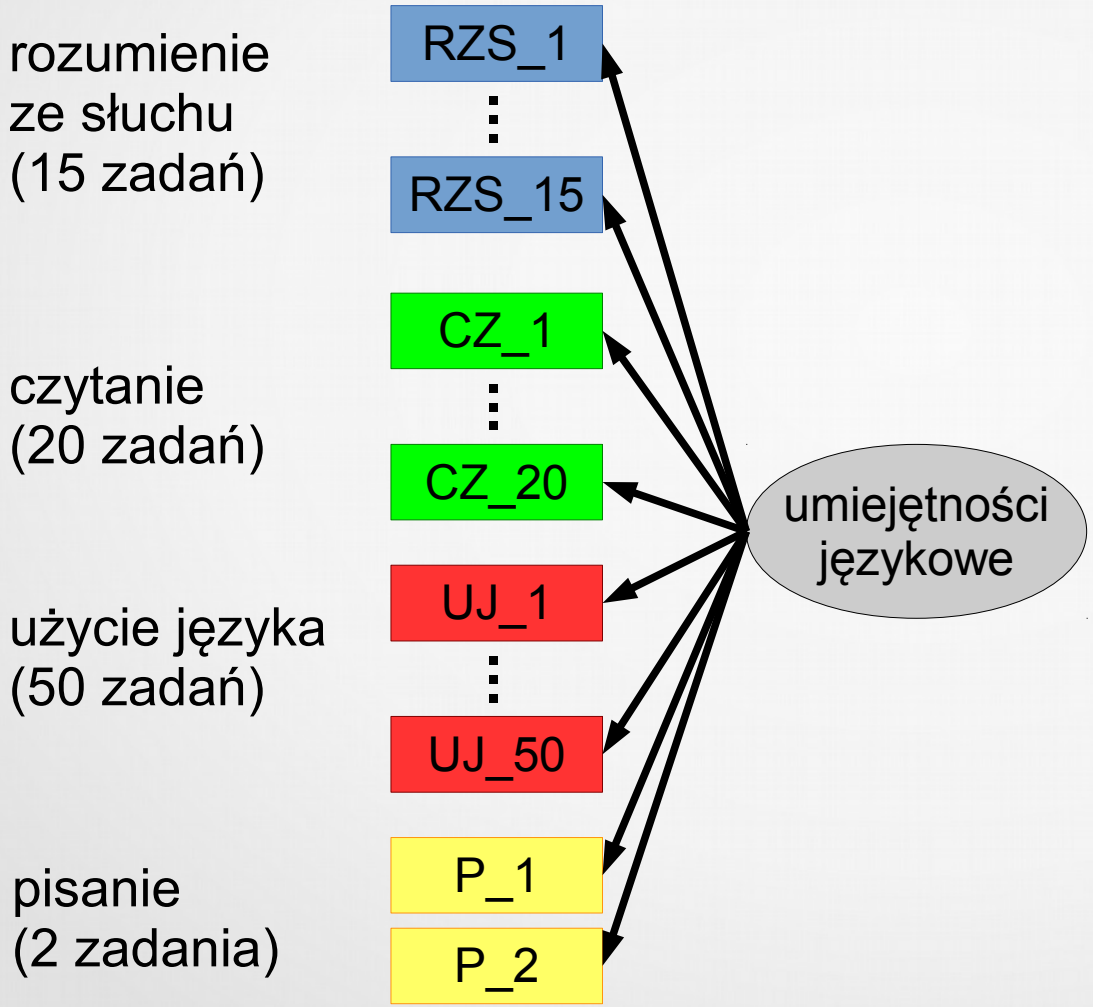
- Spadek rzetelności związany ze skróceniem testu w praktyce będzie typowo mniejszy, niż wynikałoby to ze wzoru Spearmana-Browna.
  - Efekt mniejszej długości testu może być w pewnym stopniu rekompensowany zwiększeniem spójności treściowej (zawężeniem zakresu treściowego).
- **Czy warto zwiększać rzetelność *kosztem* zakresu treściowego testu?**

Rzetelność a długość testu

- W praktyce *wzór proroczy* Spearmana-Browna ma ograniczone zastosowanie dla przewidywania wpływu zmniany liczby zadań na rzetelność testu, gdyż zakłada, że wyniki rozwiązania każdego zadania są równie silnie powiązane z badaną cechą.
  - Założenie to bardzo rzadko bywa spełniane w praktyce.
- Jeśli dysponujemy wynikami rozwiązywania długiego testu, możemy wykorzystać je do symulacyjnego badania wpływu zmniejszenia liczby zadań na rzetelność:
  - Analizujemy rzetelność testów, usuwając z danych losowo wybrane zadania.

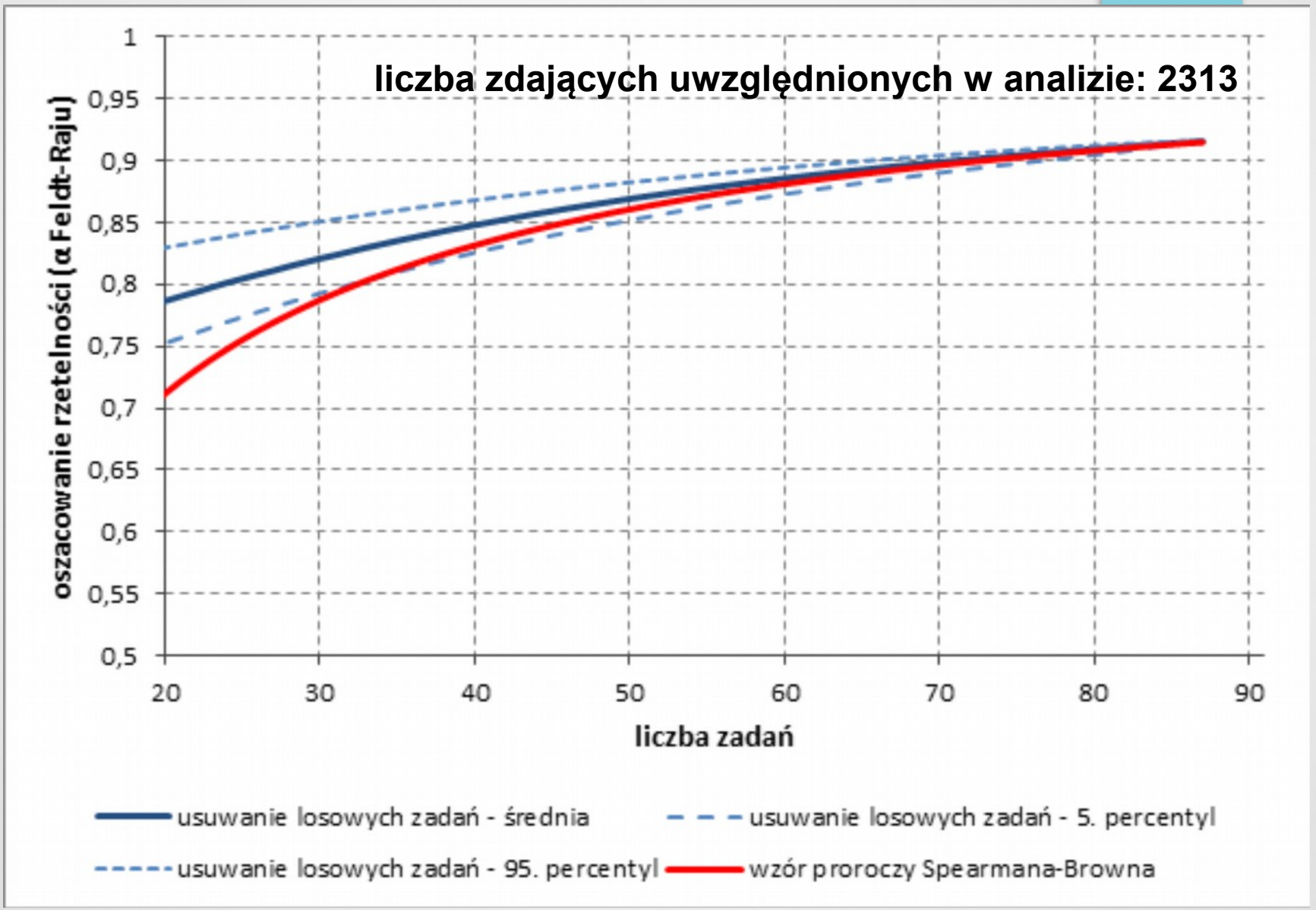
Zależność rzetelności od długości testu - egzamin z j. angielskiego na poziomie B2 w czerwcu 2012 r.

Struktura testu



Dla każdej badanej długości testu analizowano do 100 tys. losowo wybranych sposobów usunięcia z testu zadanej liczby zadań z tym, że **zadania z pisania nigdy nie były usuwane**, aby nie doprowadzić do drastycznej zmiany kompozycji treściowej testu.

Zależność rzetelności od długości testu - egzamin z j. angielskiego na poziomie B2 w czerwcu 2012 r.

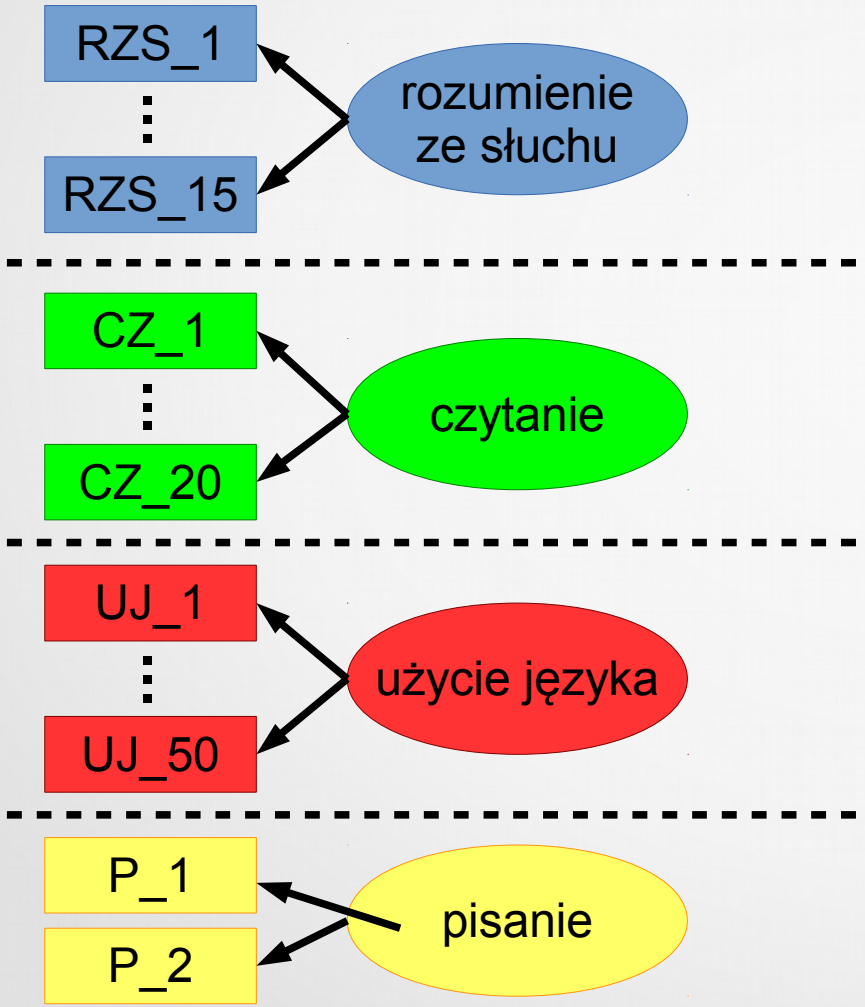


Podsumowanie

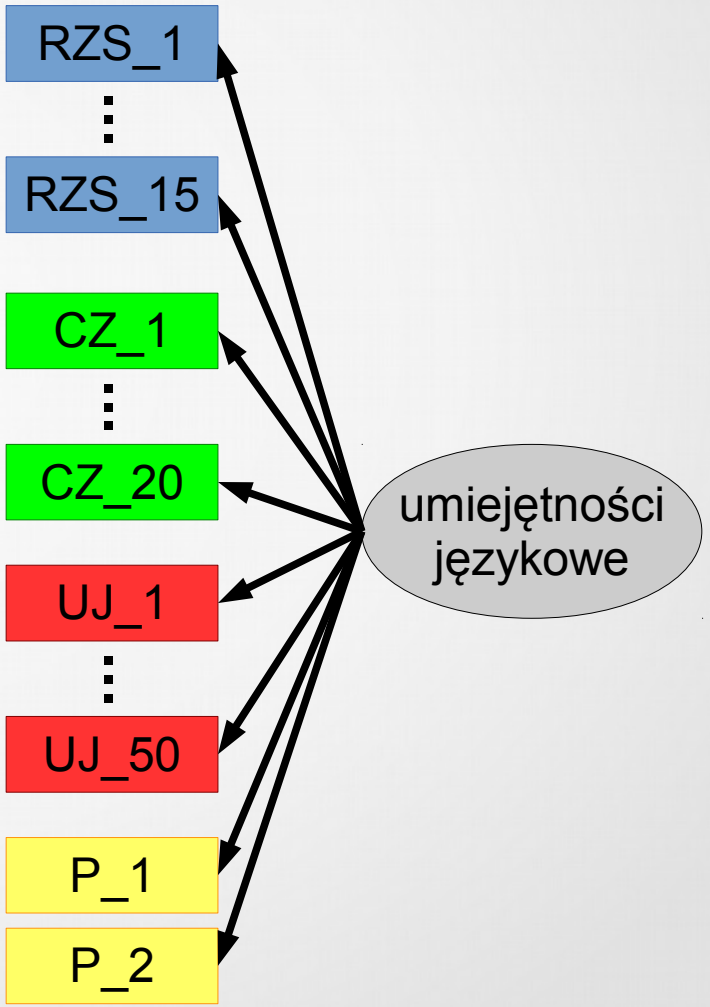
- Rzetelność jest miarą dokładności pomiaru testu.
  - Warto jednak pamiętać, że nie mówi nam, co właściwie test mierzy.
- Rzetelność rośnie wraz ze zwiększaniem się liczby zadań w teście, ale:
  - Jest to wzrost nieliniowy – najpierw znaczny, z dodawaniem kolejnych zadań staje się coraz mniejszy (np. dużo większy wzrost rzetelności uzyskamy wydłużając test z 10 zadań do 20 niż z 20 do 30).
  - Zachodzi interakcja z zawartością treściową testu: jeśli wzrost liczby zadań wiąże się z rozszerzaniem zakresu treściowego, wzrost rzetelności będzie mniejszy.

Co jest celem testowania?

Analiza każdej skali oddzielnie?



Analiza wyników całego testu?



Podsumowanie

- Jeśli celem pomiaru jest przede wszystkim ocena ogólnych umiejętności, wystarczy posługiwanie się względnie krótkimi testami.
  - Wydaje się, że z punktu widzenia tego celu liczba zadań w analizowanym egzaminie mogłaby być zmniejszona nawet o połowę, bez dużych strat dla dokładności pomiaru.
- Jeśli celem pomiaru jest dostarczenie wyczerpującej diagnozy umiejętności zdającego w ramach różnych dziedzin (skal), konieczne jest wykorzystywanie dłuższych testów, zapewniających odpowiednią dokładność pomiaru w ramach poszczególnych skal.

Dziękuję za uwagę!

Tomasz Żółtak
t.zoltak@ibe.edu.pl