X Konferencja z cyklu Diagnostyka edukacyjna
20-22 IX 2004

Bolesław NIEMIERKO
Uniwersytet Gdański

# ZRÓWNYWANIE WYNIKÓW SPRAWDZIANU 2004 DO WYNIKÓW SPRAWDZIANU 2003

W trzecim roku pełnego stosowania egzaminu zewnętrznego po sześcioletniej szkole podstawowej Polsce, zwanego *sprawdzianem*, porównywanie wyników krajowych i lokalnych uzyskiwanych przez uczniów w kolejnych latach (2002-2003-2004), nabrało znaczenia. Są różne tego powody. Jednym z nich jest krzepnięcie systemu, wzrost zaufania do wyników sprawdzianu i chęć możliwie pełnego wykorzystania informacji, jaką przynosi. Innym powodem może być niepokój wywołany dużym wzrostem trudności zadań umieszczonych w arkuszu egzaminacyjnym w 2004 roku. Czy niższe wyniki są skutkiem trudniejszego sprawdzianu, czy obniżania się osiągnięć absolwentów szkoły podstawowej w skali kraju? Na to pytanie ma odpowiedzieć niniejszy raport.

## PROBLEM

Ze względu na konieczny *obiektywizm egzaminowania*, rozumiany jako dokładność, z jaką jego wyniki są wyznaczone właściwymi *standardami wymagań*, nie zaś po prostu uczeniem się rozwiązań wybranych zadań, treść sprawdzianu krajowego, szeroko i swobodnie upowszechniana natychmiast po jego zastosowaniu, musi być w całości odnawiana co roku. Zadania dla uczniów ulegają całkowitej wymianie, a mimo to chcielibyśmy, by sprawdzian przynosił wyniki nadające się do porównywania osiągnięć uczniów *w* kolejnych latach.

Chodzi nam o takie porównanie osiągnięć *populacji krajowej*, by układem odniesienia wyników danego roku były wyniki uzyskane w *roku bazowym* – poprzednim lub (gdy takie dane są dostępne) jeszcze wcześniejszym. Jest to porównanie *podłużne*, jedynie wartościowe porównanie osiągnięć populacji krajowej. Wyniki (pozycje) uczniów w okręgach, województwach, powiatach, gminach, szkołach i oddziałach są łatwo porównywalne w standardowej *skali staninowej*, precyzyjnie wykazującej *względny* wzrost lub spadek osiągnięć uczniów, ale jeżeli staniny obliczamy co roku na nowo, to ich średnia w populacji krajowej zawsze wyniesie dokładnie 5 punktów i nie dowiemy się niczego o postępie w całym kraju.

W sensie teoretycznym chodzi o coroczne *wersje równoważne (equivalent forms)* egzaminu, dające *wyniki równoważne* na podstawie procedury zamiany wyników. Wersje równoważne mierzą to samo (są skonstruowane według wspólnego planu testu) i tak samo dokładnie (rzetelnie), a różnią się tylko właściwościami poszczególnych zadań i statystycznymi rozkładami wyników.

Równoważność wersji osiągamy przez *zrównywanie wyników* (*equating*), czyli określanie wyników równoważnych innej wersji (*Y*) dla wyników danej wersji (*X*) testu, bezpośrednio (dwukrotne badanie tych samych uczniów) lub przez *kotwiczenie* (*anchoring*), to jest przez wykorzystanie innego testu (*K*) lub mniejszej grupy zadań, która może być nazwana *testką* (*testlet*), do *kalibrowania* (cechowania) skali wyników dwu lub więcej wersji testu.

Wersje równoważne są słabszą odmianą *wersji równoległych*, to jest testów mierzących „to samo w ten sam sposób", co oznacza narzędzia (1) budowane według jednego planu, (2) dające wyniki o jednakowych średnich, wariancjach i korelacjach z dowolną zmienną oraz (3) o równej rzetelności. By uzyskać wersje równoległe testu, trzeba je budować równocześnie, wybierając, według planu merytorycznego, zadania o znanych parametrach statystycznych. Takie działania są możliwe tylko w wysoko rozwiniętych ośrodkach pomiarowych, dysponujących odpowiednimi *bankami zadań*, czyli dużymi zasobami starannie sprawdzonych zadań testowych, ale i w tych przypadkach bywają zawodne, a więc proces zrównywania wersji bywa konieczny.

## PROCEDURA

Zastosowana tu procedura zrównywania sprawdzianów została wypróbowana w ubiegłym roku na przykładzie danych z lat 2002 i 2003 (Niemierko, 2003). Polega na zastosowaniu dwu *testek kotwiczących*, złożonych z zadań dwu zrównywanych sprawdzianów, tak dobranych parami, by ich rozwiązanie wymagało podobnych czynności w każdej testce.

W procedurze równoważenia wyników sprawdzianu 2003 do wyników sprawdzianu 2002 zastosowano w klasie V szkoły podstawowej, a więc w klasie programowo niższej, testki kotwiczące zawierające tylko po siedem zadań zamkniętych, a wiec mało rzetelne (ich współczynniki „alfa" wyniosły 0,47 i 0,55). W bieżącym roku zastosowano testki złożone z 28 i 27 zadań/czynności, a ich wewnętrzna zgodność w klasie V była wysoka (0,85 i 0,88). Te testki zostały połączone w jeden test w *dwu wersjach*: *A* (2003 + 2004) i *B* (2004 + 2003), dla uniknięcia efektów obycia testowego i zmęczenia. Nadto sprawdzono ankietą stopień uprzedniej znajomości zadań obu części każdej wersji i nie stwierdzono jego wpływu na wyniki.

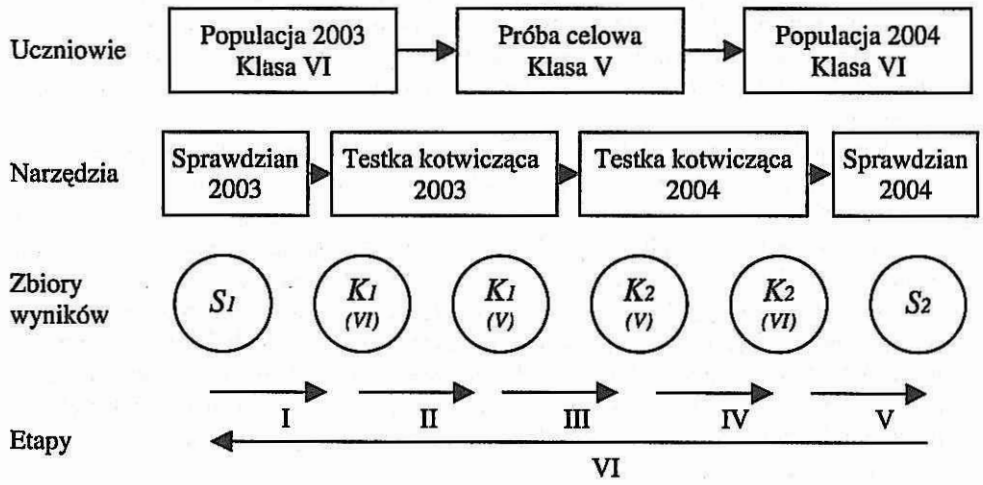
Ta procedura – optymalna, jak się zdaje, dla obecnego stanu naszych doświadczeń i możliwości organizacyjnych – obejmowała uzyskanie sześciu zbiorów wyników egzaminu:

S1 – sprawdzian w roku bazowym (cały kraj, 2003),
K1(VI) – wyniki testki kotwiczącej 2003 w klasie VI (cały kraj, 2003),
K1(V) – wyniki testki kotwiczącej 2003 w klasie V (próba celowa, 2004),
K2(V) – wyniki testki kotwiczącej 2004 w klasie V (próba celowa, 2004),
K2(VI) – wyniki testki kotwiczącej 2004 w klasie VI (cały kraj, 2004),
S2 – sprawdzian w nowym roku (cały kraj, 2004).

Liczba kolejnych porównań (etapów operacji zrównywania) wynosiła także sześć:

| | | |
|---|---|---|
| I: | *E1* i *K1(VI)* | (porównanie testu kotwiczącego ze sprawdzianem *2002*) |
| II: | *K1(VI)* i *K1(V)* | (porównanie klasy V z klasą VI przy zmianie warunków pomiaru) |
| III: | *K1(V)* i *K2(V)* | (porównanie dwu testów kotwiczących, *główny etap zrównywania*) |
| IV: | *K2(V)* i *K2(VI)* | (porównanie klasy VI z klasą V przy zmianie warunków pomiaru) |
| V: | *K2(VI)* i *E2* | (porównanie testu kotwiczącego ze sprawdzianem *2003*) |
| VI: | *E2* i *E1* | (porównanie dwu sprawdzianów, *synteza zrównywania*) |

Całość przyjętej procedury przedstawia Rysunek 1:




**Rysunek 1.** Procedura zrównywania wyników sprawdzianu 2004 do 2003

## METODA EKWICENTYLOWA

*Wyniki równoważne* (*equivalent scores*) to wyniki dwu lub więcej wersji równoważnych testu mające jednakowe rangi centylowe w dowolnej grupie badanych. Można je uzyskać za pomocą *metody ekwicentylowej* (*equipercentile method*).

*Metoda ekwicentylowa* jest zrównywaniem wyników poszczególnych wersji testu na podstawie ich *rang centylowych*. Prowadzi do tabeli zamiany surowych wyników jednej wersji testu (*Y*) na surowe wyniki innego testu (*X*). Pozwala na korektę średniej, wariancji i kształtu rozkładu wersji *Y* ze względu na wynik wersji *X*.

Zaawansowane podręczniki pomiaru dydaktycznego poświęcają metodzie ekwicentylowej wiele uwagi (Angoff, 1971, s. 508 – 600; Petersen i in., 1989, s. 221 – 262). Podkreślają jej wyższość nad *liniowym przekształceniem* wyników, które może prowadzić tylko do przesunięcia rozkładu wzdłuż skali i do zmiany wielkości jednostki (jej „wydłużenia" lub „skrócenia"), ale nie koryguje kształtu rozkładu wyników. Przestrzegają zarazem przed niedokładnościami, jakie są skutkiem założenia o ciągłości mierzonej właściwości, a w szczególności technik interpolacji i ekstrapolacji jej wartości.

Odręczne wygładzanie (*smoothing*) krzywych także powoduje pewne dowolności, zwłaszcza na krańcach skali, gdzie danych jest mało lub mogą się spiętrzać („efekt pułapu", rzadziej „efekt dna").

## ZRÓWNYWANIE WYNIKÓW SPRAWDZIANU

Zastosowanie metody ekwicentylowej w sześciu etapach zrównywania wyników *Sprawdzianu 2004* do wyników *Sprawdzianu 2003* jest przedstawione na rysunkach 2 — 7, załączonych do tego sprawozdania. A oto krótkie uwagi o kolejnych etapach:

(Rysunek 2) Etap I (*S1 - K1(VI)*). Testka kotwicząca *K1* okazuje się nieco łatwiejsza od sprawdzianu *S1* (różnica +0,03 w skali łatwości), ale całkiem dobrze różnicuje, co wyraża się stosunkowo większym rozrzutem wyników i prawie równą rzetelnością (*alfa* = 0,83 wobec 0,87 dla sprawdzianu), a krzywa ekwicentylowa przybiera kształt „esowaty".

(Rysunek 3) Etap II (*K1(VI) — K1(V)*). Jak było do przewidzenia, testka kotwicząca *K1* jest systematycznie trudniejszą dla piątoklasistów, ale różnica jest umiarkowana (-0,08 w skali łatwości). Wobec wzrostu rozrzutu wyników, rzetelność (*alfa* = 0,85) nieco wzrasta, co może świadczyć o pilnej pracy uczniów (przy mniejszej, jak wiemy, doniosłości wyników). Krzywa wygina się regularnym łukiem ku klasie VI.

(Rysunek 4) Etap III (*K1(V) — K2(V)*). Nowa testka, *K2*, jest silniejsza pomiarowo od dawnej, *K1*. Jest dużo trudniejsza (różnica -0,14 w skali łatwości), przynosi większy rozrzut i wzrost rzetelności (*alfa* = 0,88). Krzywa wygina się wielkim łukiem ku testce *K1*. Należy przypomnieć, że kolejność rozwiązywania dwu testek była zamieniona dla połowy próby (*counterbalancing*), by nie mogła wpływać na ich właściwości statystyczne.

(Rysunek 5) Etap IV (*K2(V) — K2(VI)*). Wyniki uczniów klasy VI w testce *K2* są, oczywiście, wyższe niż w klasie V (różnica +0,11 w skali łatwości) Krzywa odchyla się systematycznie ku klasie VI. Rozrzut wyników zmalał bardzo niewiele, a rzetelność (*alfa* = 0,88) pozostała na poprzednim, wysokim poziomie.

(Rysunek 6) Etap V (*K2(VI) — S2*). Wracamy do skali czterdziestojednopunktowej przy średniej niższej o aż o 3,06 punktu, znacznie większym rozrzucie wyników niż w poprzednim roku i bardzo wysokiej rzetelności (*alfa* = 0,90, pamiętać należy jednak o tym, że podział zadań złożonych na zerojedynkowe „czynności" powoduje sztuczne podwyższenie współczynnika). Ponieważ testka kotwicząca i sprawdzian w klasie VI są prawie równe pod względem łatwości (różnica +0,01), a kształt rozkładów jest podobny (z „garbem" po stronie wyników wysokich), krzywa ekwicentylowa niewiele odbiega od przekątnej wykresu.

(Rysunek 7) Etap VI (*S2—- S1*). Wracamy do roku bazowego *2003*. Dopiero teraz dokonuje się *równoważenie* wyników, bo sprawdziany są jednakowo długie i prawie równie rzetelne. Przenoszenie skal *0 — 40 punktów* na kolejne testki było tylko ich kalibrowaniem. Inaczej niż w badaniu pilotażowym (*2003 — 2002*), gdzie sprawdziany okazały się bardzo podobne co do trudności i krzywa ekwicentylowa „wiła się" wzdłuż głównej przekątnej wykresu, tym razem krzywa systematycznie (wyjątek stanowi stanin 9, najwyższy) i silnie wygina się ku wiele łatwiejszemu sprawdzianowi *2003*, co jednak wcale nie przesądza o obniżaniu się osiągnięć uczniów w tych latach. Dopiero

przekształcenie zbioru wyników 2004 na wyniki równoważne 2003 i obliczenie *średniej wyników równoważnych* pokaże nam, czy to nastąpiło.

**Tabela 1.** Wartości punktowe centyli C0,1 — C99 oszacowane metodą ekwicentylową

| Centyl | Sprawdzian 2003 | Testka 2003 Klasa VI | Testka 2003 Klasa V | Testka 2004 Klasa V | Testka 2004 Klasa VI | Sprawdzian 2004 |
|---|---|---|---|---|---|---|
| 0,1 | 2,6 | 1,4 | 0,6 | 0,2 | 0,3 | 0,7 |
| 0,2 | 6,7 | 4,1 | 2,2 | 0,8 | 1,5 | 3,5 |
| 0,5 | 8,8 | 5,6 | 3,2 | 1,3 | 2,1 | 4,5 |
| 1 | 10,9 | 7,2 | 4,6 | 2,1 | 3,4 | 6,5 |
| 2 | 13,0 | 9,0 | 6,3 | 3,0 | 4,8 | 8,6 |
| 5 | 15,8 | 11,3 | 8,4 | 4,3 | 6,7 | 11,3 |
| 10 | 18,9 | 13,7 | 10,7 | 6,0 | 8,9 | 14,5 |
| 20 | 22,8 | 16,5 | 13,6 | 8,6 | 11,7 | 18,4 |
| 30 | 25,6 | 18,5 | 15,7 | 10,6 | 13,8 | 21,3 |
| 40 | 27,9 | 20,2 | 17,4 | 12,4 | 15,6 | 23,8 |
| 50 | 29,8 | 21,5 | 18,9 | 14,0 | 17,2 | 26,0 |
| 60 | 31,5 | 22,8 | 20,4 | 15,8 | 19,0 | 28,4 |
| 70 | 32,7 | 23,8 | 21,6 | 17,2 | 20,4 | 30,1 |
| 80 | 34,8 | 25,2 | 23,5 | 19,6 | 22,4 | 32,7 |
| 90 | 36,6 | 26,5 | 15,2 | 21,9 | 24,5 | 34,7 |
| 95 | 38,0 | 27,3 | 26,6 | 24,1 | 25,9 | 38,0 |
| 99 | 39,4 | 27,9 | 27,8 | 26,3 | 26,8 | 39,6 |

Tabela 1 pokazuje, że wszystkie zastosowane testy są w pewnym stopniu dotknięte *efektem pułapu*, polegającym na skupianiu się wyników w górnej części skali (najsilniej w przedziale C60 – C70), co powoduje, że najdokładniej zrównoważone są wyniki bardzo niskie i niskie (w dolnej ćwiartce rozkładu). Ze względu wszakże na trafność pomiaru (motywację uczniów) nie należałoby w przyszłości zwiększać trudności tych testów.

## WYNIKI ZRÓWNANE

Tabelka *S2/S1*, zamieszczona na prawym krańcu rys. 7, zawiera końcowy wynik sześciu etapów przekształceń, bezpośrednio użyteczny do porównań wyników *Sprawdzianu 2004* z wynikami sprawdzianu z poprzedniego roku. Dla wygody czytelnika raportu jest ona tu powtórzona jako tabela 2.

**Tabela 2.** Wyniki Sprawdzianu 2004 zrównane do wyników Sprawdzianu 2003

| S2 | S1 | S2 | S1 | S2 | S1 | S2 | S1 |
|---|---|---|---|---|---|---|---|
| 0 | 0,0 | | | | | | |
| 1 | 3,4 | 11 | 15,7 | 21 | 25,5 | 31 | 33,8 |
| 2 | 5,1 | 12 | 16,8 | 22 | 26,4 | 32 | 34,5 |
| 3 | 6,7 | 13 | 17,8 | 23 | 27,2 | 33 | 35,1 |
| 4 | 7,9 | 14 | 18,7 | 24 | 28,1 | 34 | 35,9 |
| 5 | 9,1 | 15 | 19,7 | 25 | 29,0 | 35 | 36,5 |
| 6 | 10,3 | 16 | 20,7 | 26 | 29,9 | 36 | 37,1 |
| 7 | 11,5 | 17 | 21,7 | 27 | 30,7 | 37 | 37,7 |
| 8 | 12,5 | 18 | 22,6 | 28 | 31,5 | 38 | 38,3 |
| 9 | 13,5 | 18 | 23,6 | 29 | 32,1 | 39 | 39,1 |
| 10 | 14,6 | 20 | 24,6 | 30 | 32,9 | 40 | 40,0 |

Według tabeli 2, posługując się interpolacją, możemy „z grubsza" oszacować wartość średniego wyniku *Sprawdzianu 2004*, *S2*, w skali wyników *Sprawdzianu 2003*, *S1*. Średnia 25,6 z 2004 roku odpowiada średniej 29,5 w 2003 roku, o prawie cały punkt wyższej od średniej sprawdzianu *S1*, która wynosiła 28,6. Nie powinniśmy jednak dowierzać takiemu prostemu odczytowi, bo nie uwzględnia on kształtów rozkładów liczebności, co było przecież powodem stosowania metody ekwicentylowej. Należy średnią arytmetyczną (wariancję, korelacje) obliczyć na nowo, podstawiając wartości z kolumny *S1* do wszystkich wierszy tabeli rozkładu liczebności. Tak obliczona średnia wynosi 28,930 i ta właśnie wartość posłuży nam do porównania osiągnięć absolwentów szkoły podstawowej w latach 2003 – 2004.

## OSZACOWANIE BŁĘDU ZRÓWNYWANIA

Oszacowanie *błędu zrównywania*, to jest błędu standardowego dokonanego przekształcenia, może być dokonane według następującego wzoru (por. Angoff, 1971, s. 570n):

$$\sigma_{d_E} = \sqrt{\sigma_{E_1}^{\,2}[1+z_{E_2}^{\,2}(1+\bar{r})](\frac{2-r_{E_1K_{1(VI)}}}{n_1-1}+\frac{2-r_{K_{1(V)}K_{2(V)}}}{n_2-1}+\frac{2-r_{E_2K_{2(VI)}}}{n_3-1})}$$

gdzie poszczególne symbole mają następujące znaczenie:

$\sigma_{d_S}$ – błąd standardowy zrównywania wyników dwu sprawdzianów,

$z_{S_2} = \frac{x_{S_2} - \overline{x_{S_2}}}{\sigma_{S_2}}$ – wartość standardowa wyniku sprawdzianu *2003*

$$\bar{r} = \sqrt{\frac{r_{S_1K_{1(VI)}}^{\,2} + r_{K_{1(V)}K_{2(V)}}^{\,2} + r_{S_2K_{2(VI)}}^{\,2}}{3}} = \sqrt{\frac{0{,}941+0{,}548+0{,}949}{3}} = 0{,}901$$

Wzór pokazuje zależność błędu zrównywania od:

1. wariancji wyników sprawdzianu bazowego (*2003*),
2. wielkości trzech prób (populacji) uczniów,
3. korelacji wyników uzyskanych w poszczególnych próbach: gdy jest zupełna, to wariancja błędu transformacji jest sumą wariancji błędów trzech niezależnych prób, a gdy jest zerowa, wariancja błędu transformacji jest sumą wariancji błędów sześciu niezależnych prób,
4. wartości standardowej transformowanego wyniku: dla wyników oddalonych od średniej błąd transformacji rośnie.

Dla dokonanego przekształcenia średniej arytmetycznej wyników sprawdzianu *2004* na średnią arytmetyczną wyników sprawdzianu *2003* błąd standardowy wynosi:

$$\sigma_{d_E} = \sqrt{45{,}33[1+0(1+0{,}901)](\frac{2-0{,}970}{530577-1}+\frac{2-0{,}740}{1836-1}+\frac{2-0{,}974}{517081-1})} =$$

$$= \sqrt{45{,}33(0{,}00000191+0{,}000687+0{,}00000198)} = 0{,}177$$

Jak widzimy, błąd standardowy transformacji pochodzi prawie wyłącznie z wariancji wyników *Sprawdzianu 2002* i z wariancji błędu próby piątoklasistów, która jest o dwa rzędy wielkości większa niż wariancja błędu populacji ogólnopolskiej. Jego wartość stanowi niewiele ponad połowę wartości błędu szacowanego w poprzednim roku (w toku badań pilotażowych), a to dlatego, że testki są rzetelniejsze (ich korelacja wynosiła wtedy tylko 0,505) oraz dlatego, że próba piątoklasistów jest staranniej dobrana (Dolata, 2004) i liczniejsza (przed rokiem obejmowała tylko 604 uczniów).

## RÓŻNICA WYNIKÓW DWU SPRAWDZIANÓW

Różnica między średnimi wynikami sprawdzianu w latach 2004 (zrównanymi) i 2003 wynosi:

| 28,930 punktu – 28,613 punktu = 0,317 punktu |
|---|

Ta wielkość różnicy (0,317), porównana z poprzednio oszacowanym błędem standardowym zrównywania (0,177), wykazuje stosunek z = 1,79, co w dwustronnym teście istotności różnic między dwiema nieskorelowanymi średnimi pozwala na odrzucenie hipotezy zerowej (o braku różnic w osiągnięciach uczniów między zbiorami wyników S2 a S1) z prawdopodobieństwem błędu $p < 0{,}07$, co nie spełnia warunku powszechnie stosowanego kryterium istotności statystycznej ($p < 0{,}05$). Możemy jednak przyjąć, że krajowe wyniki uzyskane w sprawdzianie w 2004 roku nie są niższe niż wyniki uzyskane w 2003 roku, a prawdopodobieństwo, że są wyższe, jest duże (0,93).

Zrównywanie wyników *Sprawdzianu 2003* do wyników *Sprawdzianu 2002*, przeprowadzone w ramach badań pilotażowych, dało inny wynik. Różnica średnich (w skali wyników Sprawdzianu 2002) wyniosła -0,82 punktu, co – mimo mniejszej liczebności próby piątoklasistów, a więc większego błędu standardowego zrównywania i niższej mocy testu istotności – okazało się istotne statystycznie ($p < 0{,}02$). Zatem

tendencja spadkowa wyników sprawdzianu na pewno została powstrzymana i zapewne odwróciła się w bieżącym roku.

## INTERPRETACJA RÓŻNICY WYNIKÓW

Interpretacja *wzrostu wyników punktowych* sprawdzianu i prawdopodobnego *wzrostu osiągnięć* szóstoklasistów wykracza poza problematykę statystycznego zrównywania wyników. Potrzebne są na to dalsze badania, oparte na porównaniach wyników w dłuższym okresie i bardziej szczegółowych analizach obszarów programowych postępu i warstw populacji, w jakich występuje. Pewne znaczenie może mieć także badanie opinii egzaminatorów i działaczy oświatowych o tych zjawiskach.

Spadek wyników sprawdzianu w drugim roku jego powszechnego stosowania (w latach 2002 – 2003) mógł być spowodowany głównie zaostrzeniem kontroli jego przebiegu, to jest ograniczeniem *oszustwa egzaminacyjnego* ze strony niektórych uczniów (podpowiadanie i „ściąganie") oraz niektórych komisji egzaminacyjnych (tolerowanie porozumiewania się, a niekiedy nawet podpowiadanie rozwiązań i korygowanie błędów). Pewne znaczenie mogły mieć też takie czynniki, jak dekoniunktura ekonomiczna, wzrost bezrobocia i kryzys władzy politycznej w Polsce, to jest okoliczności, które tłumaczą przejściowy spadek motywacji uczenia się i wyników pomiaru dydaktycznego w innych krajach (Harnischfeger, 1975). Generalnie jednak, osiągnięcia uczniów rosną wraz z postępem ekonomicznym i cywilizacyjnym danego kraju, najwyraźniej na pierwszym szczeblu kształcenia (Scottish Council..., 1967).

Wzrost wyników sprawdzianu w trzecim roku jego stosowania (w latach 2003 – 2004) mógł być spowodowany *pozytywnym efektem zwrotnym* systemu egzaminacyjnego, to jest jego korzystnym wpływem na program i przebieg kształcenia, a także wzrostem *obycia egzaminacyjnego* (*test-taking skill*) uczniów, w tym zwłaszcza umiejętności rozwiązywania zadań zamkniętych. Zrobiono sporo, by upowszechnić wiedzę o sprawdzianie i by przygotować nauczycieli do odpowiedniej pracy z uczniami. Oddziaływanie systematycznych czynników rozwoju ekonomiczno-społecznego kraju na edukację jest prawdopodobne, ale niemożliwe do zaobserwowania w tak krótkim okresie, jak jeden rok.

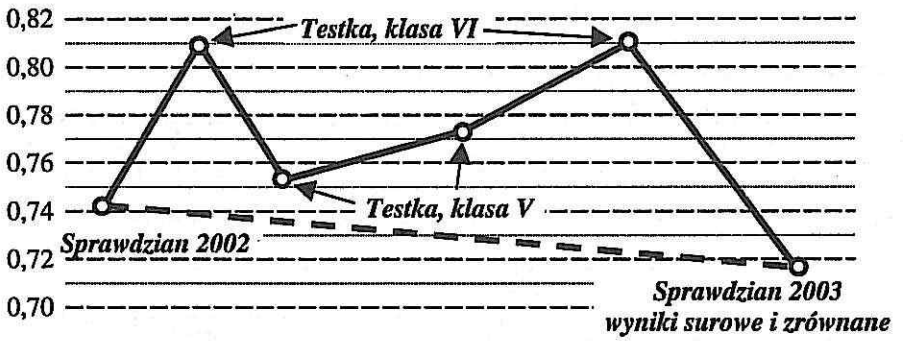
## OCENA PROCEDURY ZRÓWNYWANIA WYNIKÓW SPRAWDZIANÓW

Metoda ekwicentylowa ma nas uniezależnić od wahań łatwości testów i od kształtu rozkładów liczebności. Między *Sprawdzianem 2002* a *Sprawdzianem 2003* nie stanowiło to problemu, bo różnica łatwości była niewielka (wynosiła około 0,2 punktu), a rozkłady były podobne. Gdyby intuicyjnie przyjąć *równoległość* dwu wersji sprawdzianu, *2003* i *2002*, byłoby to założenie prawdziwe, choć nie udowodnione, bo we wcześniejszej praktyce egzaminacyjnej wersje testów uważane za równoległe okazywały się często mocno odmienne (Niemierko i in., 1975). Testki kotwiczące natomiast były bardzo łatwe, co pokazuje tabela 3.

**Tabela 3.** Porównanie średnich wskaźników łatwości testów w badaniach pilotażowych

| Symbol | Zbiór danych | Liczebność zbioru | Średnia łatwość | Różnica | Różnica skumulowana |
|---|---|---|---|---|---|
| S1 | Sprawdzian *2002* | 537 614 | 0,738 | | |
| K1(VI) | Testka kotwicząca 1 | 537 614 | 0,810 | ]+ 0,072 | + 0,072 |
| K1(V) | Testka kotwicząca 1 | 604 | 0,751 | ]– 0,059 | + 0,013 |
| K2(V) | Testka kotwicząca 2 | 604 | 0,774 | ]+ 0,023 | + 0,036 |
| K2(VI) | Testka kotwicząca 2 | 464 785 | 0,806 | ]+ 0,032 | + 0,068 |
| S2 | Sprawdzian *2003* | 464 785 | 0,718 | ]– 0,088 | – 0,020 |

Średnie wskaźniki łatwości sześciu pomiarów, zestawione w tabeli 3, są zilustrowane rysunkiem 8, ułatwiającym ich porównanie:




**Rysunek 8.** Wahania łatwości testów w pilotażowym zrównywaniu wyników sprawdzianów

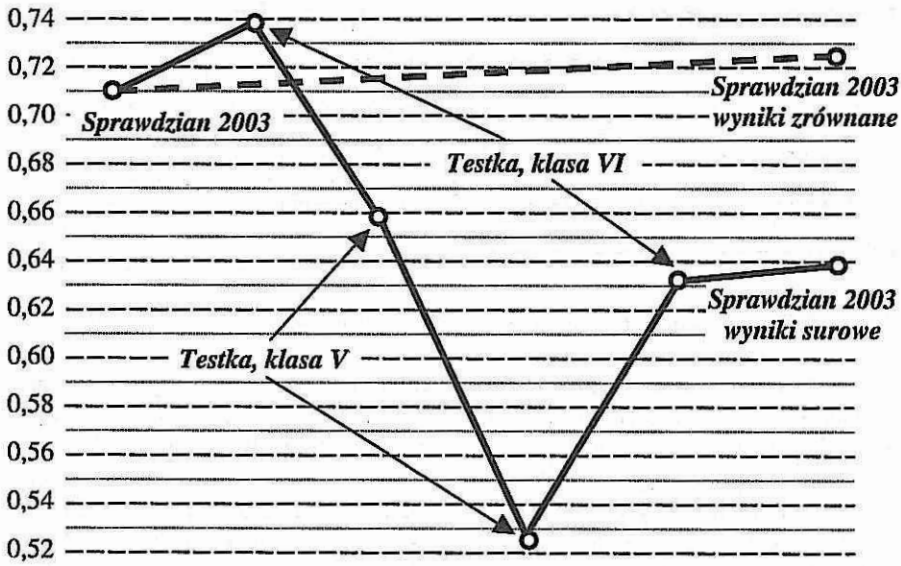
Największe różnice na rys. 8 (najbardziej strome odcinki krzywej zaznaczonej linią ciągłą) dotyczą przejścia od sprawdzianu do testki kotwiczącej (dane z 2002 roku) i od testki kotwiczącej do sprawdzianu (dane z 2003 roku) w klasie VI. Najmniejsze różnice występują między wynikami dwu sprawdzianów i dwu testek w tej samej klasie. Ta równoległość czyni zrównywanie wyników dwu sprawdzianów operacją poznawczo mało rewelacyjną, choć cenną jako doświadczenie metodologiczne. W istocie różnica między średnią arytmetyczną wyników zrównanych (*Sprawdzianu 2003* do *Sprawdzianu 2002*), wynoszącą 28,67, a średnią arytmetyczną wyników surowych, wynoszącą 28,71, pozostaje w granicach błędu dokonywanych zaokrągleń. Zrównywanie wyników, zaznaczone na rys. 8 linią kropkowaną, doprowadziło (w przybliżeniu) do tego samego punktu na wykresie, co obliczenie średniej arytmetycznej wyników *Sprawdzianu 2003* z surowych danych.

Sześć etapów zrównywania wyników *Sprawdzianu 2003* do *Sprawdzianu 2004* daje obraz całkowicie odmienny od przedstawionego w tabeli 3 i na rys. 8. Zbiór surowych danych z 2003 roku uzupełniono o około 15 procent, co minimalnie obniżyło ich wskaźnik łatwości (o 0,003). Zasadnicze znacznie miało: (a) wprowadzenie w 2004 roku zdecydowanie trudniejszej wersji sprawdzianu (o 0,077, a więc o blisko osiem punktów procentowych) i (b) zastosowanie testek kotwiczących dorównujących trudnością sprawdzianom. Pokazuje to tabela 4.

**Tabela 4.** Porównanie średnich wskaźników łatwości testów w raportowanych badaniach

| Symbol | Zbiór danych | Liczebność zbioru | Średnia łatwość | Różnica | Różnica skumulowana |
|---|---|---|---|---|---|
| S1 | Sprawdzian *2003* | 530 577 | 0,715 | | |
| K1(VI) | Testka kotwicząca 1 | 530 577 | 0,739 | ]+ 0,024 | + 0,024 |
| K1(V) | Testka kotwicząca 1 | 1 836 | 0,659 | ]– 0,080 | – 0,056 |
| K2(V) | Testka kotwicząca 2 | 1 836 | 0,523 | ]– 0,136 | – 0,192 |
| K2(VI) | Testka kotwicząca 2 | 517 081 | 0,634 | ]+ 0,111 | – 0,081 |
| S2 | Sprawdzian *2004* | 517 081 | 0,638 | ]+ 0,004 | – 0,077 |

Rys. 9 ilustruje zmiany wskaźnika łatwości w kolejnych etapach procedury zrównywania przedstawionej w tabeli 4:




**Rysunek 8.** Wahania łatwości testów w raportowanym zrównywaniu wyników sprawdzianów

Obszar zmienności wskaźnika łatwości na rys. 9 jest ponad dwukrotnie większy od przedstawionego na rys. 8. Na tak wielkie wahania składa się: (1) różnica trudności sprawdzianów i reprezentujących je testek w obu klasach i (2) różnica osiągnięć uczniów między klasą V a klasą VI. Inaczej natomiast niż w badaniu pilotażowym, testki kotwiczące są w klasie VI dość dokładnie wyrównane pod względem trudności z macierzystymi sprawdzianami. Na tle dużej i mocno spadającej łatwości sprawdzianów oraz zygzaków łatwości testek w kolejnych porównaniach, lekko wznoszący się w górę trend wyników zrównanych jawi się jako dowód stałości zjawisk masowych: „spokojnej", powolnej zmiany wskaźników skuteczności edukacji.

## PERSPEKTYWY ZRÓWNYWANIA WYNIKÓW SPRAWDZIANU

Dokonane analizy były pionierskie w skali kraju, ale od co najmniej półwiecza są rutynowym działaniem wielkich ośrodków pomiarowych w USA i w kilku innych krajach, głównie anglosaskich. Z tych doświadczeń przyjdzie nam zapewne coraz szerzej korzystać.

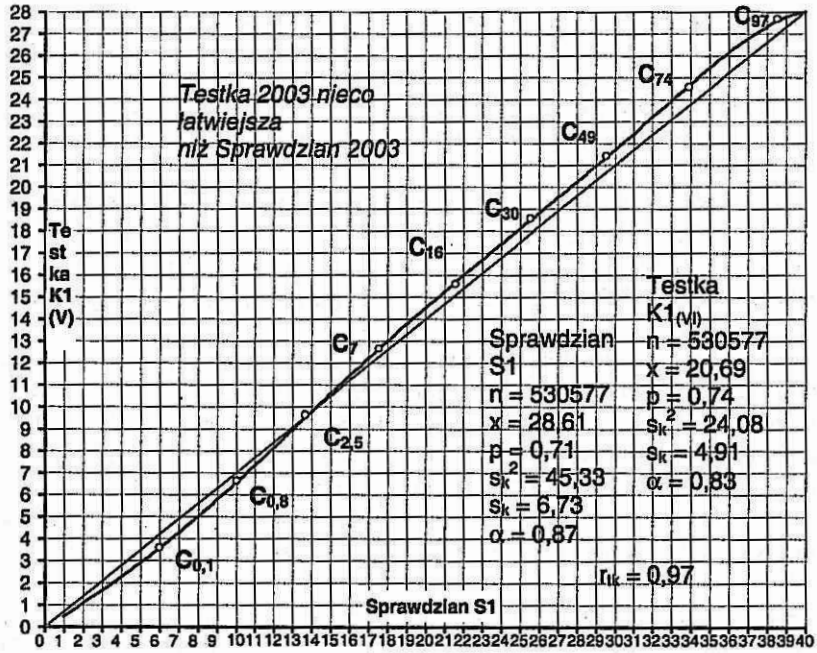
Pod względem *merytorycznym* znaczenie corocznego zrównywania wyników egzaminów zewnętrznych, czyli ich przenoszenia na skalę wyników *roku bazowego* (wyjściowego), nie budzi większych wątpliwości. Władze oświatowe i polityczne oraz opinia publiczna wykazują duże zainteresowanie rozwojem osiągnięć szkolnych młodzieży w każdym wieku.

Pod względem *pomiarowym* analizy podłużne wyników egzaminu, jako swego rodzaju *monitorowanie egzaminów*, wydają się niezbędne. Trzeba jasno odróżniać optymalizację trudności zadań od poziomu osiągnięć uczniów i długofalowych trendów skuteczności edukacyjnej. Wskazane byłoby zatem objęcie analizami obu części egzaminu gimnazjalnego, a w przyszłości – także i nowej matury.

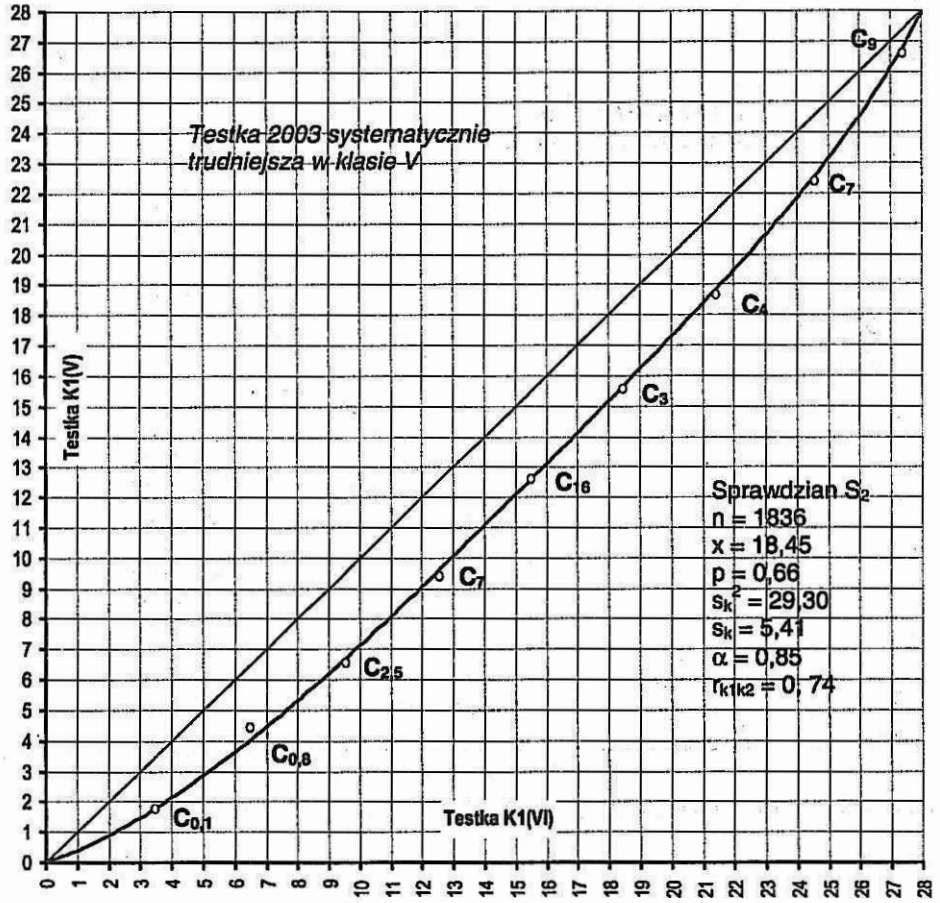
Pod względem *organizacyjnym* wiele można ulepszyć, przede wszystkim w zakresie doboru próby uczniów stosowanej do *standaryzacji arkuszy egzaminacyjnych*. Reprezentatywne (odpowiednio liczne i starannie dobrane) próby mogą być wykorzystane do zrównywania wyników kolejnych sprawdzianów już „na starcie", przed egzaminem. Doświadczenia kolejnych lat egzaminowania zewnętrznego pokazują, że *pretestowanie* (badania wstępne) dużej liczby pojedynczych zadań testowych na terenie kraju, a zatem ujawnienie ich treści niektórym uczniom, nie podnosi znacząco wyników tych uczniów. Stopniowo będziemy zdobywać się na odwagę budowania kolejnych testów z banków zadań.

Pod względem *technicznym* jesteśmy w przededniu rewolucji. Pojawiają się możliwości wykorzystania *probabilistycznych modeli wyniku zadania testowego* (*IRT*), pozwalających na optymalizowanie informacji uzyskiwanej w kolejnych egzaminach, na różnych poziomach osiągnięć, a bez stosowania testek kotwiczących (Hambleton i Swaminathan, 1985; Szaleniec, 2002). Przy silniejszych założeniach o jednorodności mierzonej właściwości rosną możliwości interpretacji osiągnięć uczniów.

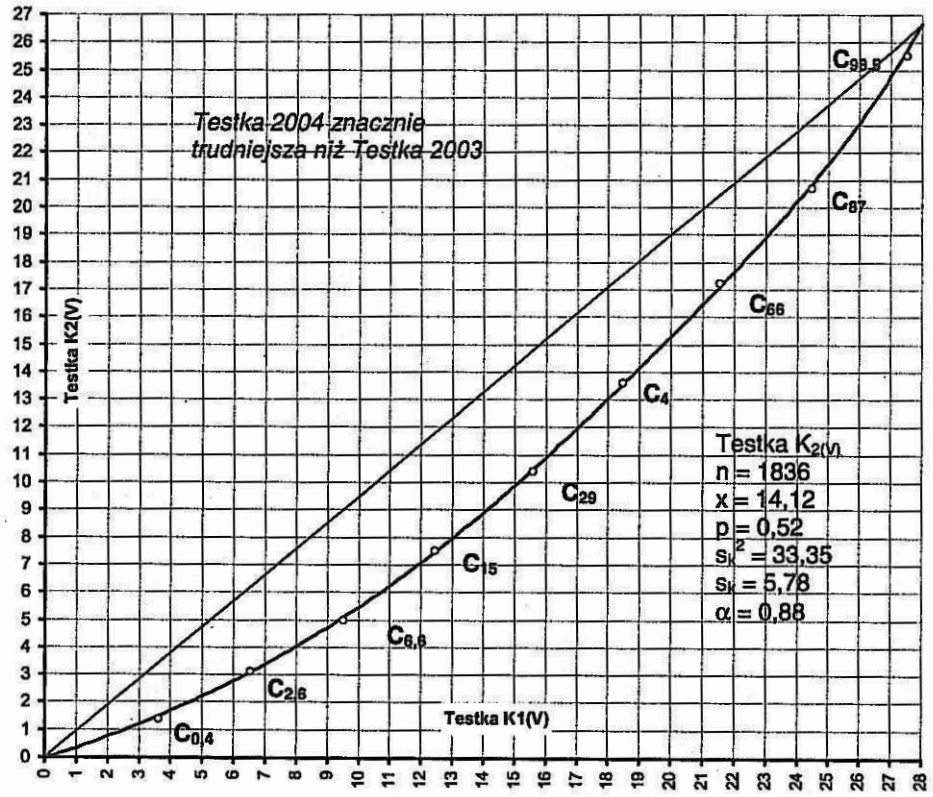
Pozostaje wzgląd *polityczny*. O ile w ubiegłym roku upowszechnianie spostrzeżeń dokonanych przez zrównywania wyników sprawdzianów (*2003 – 2002*) można było uznać za przedwczesne (charakter pilotażowy, mała próba kotwicząca, niepewność interpretacji), o tyle w obecnym roku (*2004 – 2003*) obraz porównań podłużnych zasługuje już na prezentację publiczną. Sprawdzono i ulepszono metodologię porównań, a przełamanie tendencji spadkowej wyników sprawdzianu stanowi istotną wartość. Nie mamy, rzecz jasna, pewności, że wzrost wyników sprawdzianu będzie utrzymany w następnych latach, ale dobra wiadomość bieżąca o naszej edukacji jest sama przez się cenna.



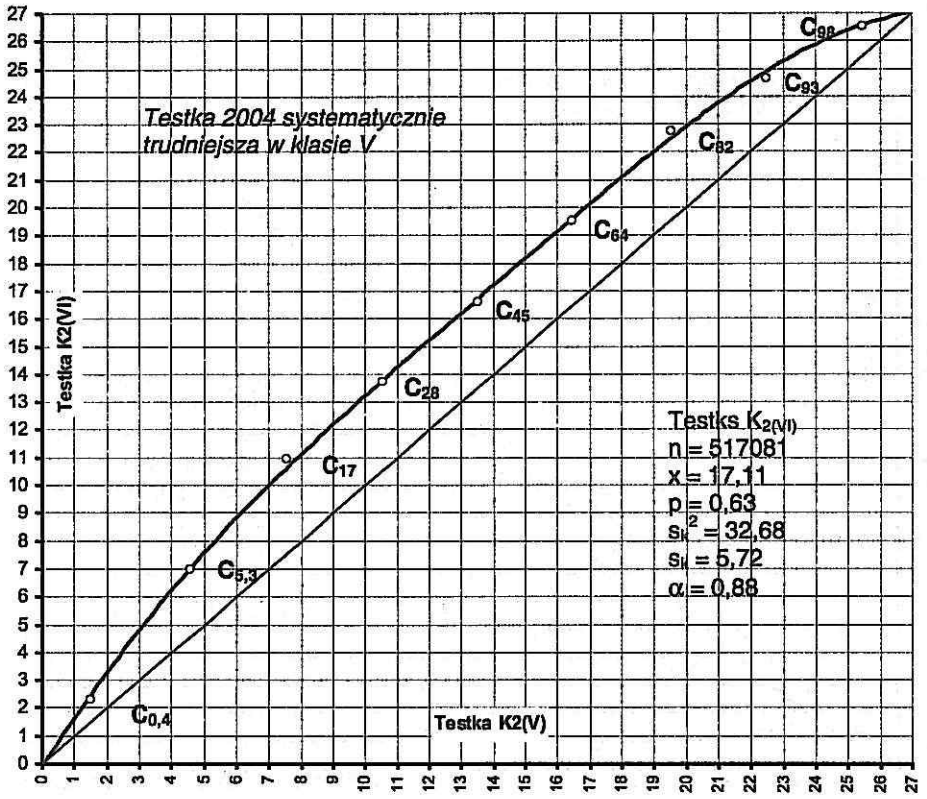

**Rysunek 2.** SPRAWDZIAN 2003, S1, a TESTKA KOTWICZĄCA 2003, $K1_{(VI)}$



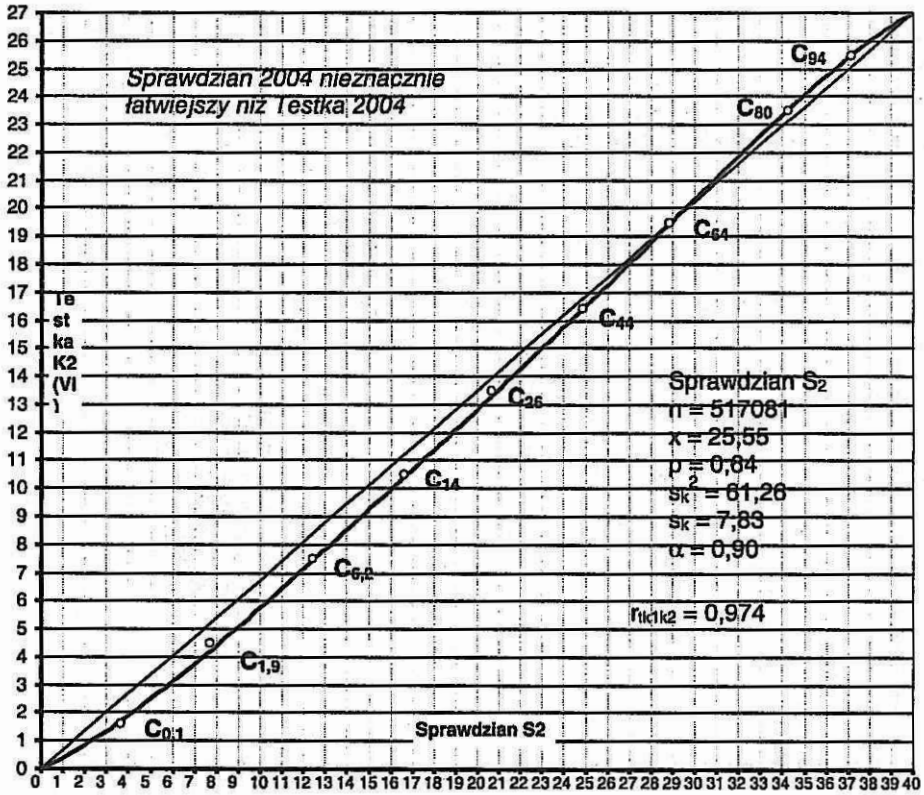

**Rysunek 3.** TESTKA KOTWICZĄCA 2003 w klasie VI, $K_{1(VI)}$ i w klasie V, $K_{1(V)}$



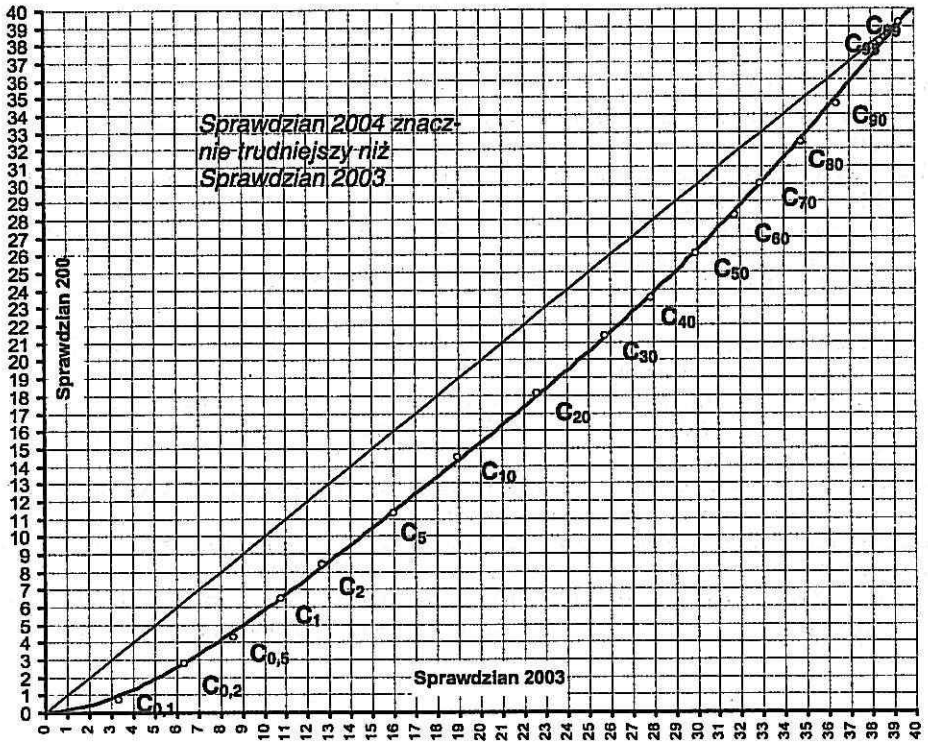

**Rysunek 4.** TESTKI KOTWICZĄCE: 2003, $K_{1(V)}$ I 2004, $K_{2(V)}$ w klasie V




**Rysunek 5.** TESTKA KOTWICZĄCA 2004 w klasie VI, $K_{2(VI)}$ i w klasie V, $K_{2(V)}$




**Rysunek 6.** TESTKA KOTWICZĄCA 2004, $K_{2(VI)}$, a SPRAWDZIAN 2004, $S_2$




**Rysunek 7.** SPRAWDZIAN 2003, $S_1$, a SPRAWDZIAN 2004, $S_2$

# LITERATURA

Angoff W.H., (1971), *Scales, norms, and equivalent scores.* (w:) R.L. Thorndike (red.) *Educational Measurement. Second edition*, American Council on Education, Washington.

Dolata R., (2004), *Konstruowanie próbki do badań kalibracyjnych*, maszynopis.

Hambleton R.H., H. Swaminathan, (1985) *Item Response Theory: Principles and Applications*, Kluver-Nijhoff, Hingham.

Harnischfeger A., D.E. Wiley, (1975), *Achievement test score decline: Do we need to worry?* „Journal of Educational Measurement", nr 3.

Niemierko B., (2003), *Zrównywanie wyników sprawdzianu 2002 do wyników sprawdzianu 2003*, maszynopis.

Niemierko B., Balińska H. i Zarębska J., (1975), *Wersje równoległe testu osiągnięć szkolnych.* „Kwartalnik Pedagopgiczny", nr 4.

Petersen, N.S. Kolen M.J., Hoover H.D., (1989) *Scaling, norming, and equating* (w:) R.L. Linn (red.) *Educational Measurement. Third edition*, American Council on Education, New York.

Scottish Council for Research in Education, (1967), *Rising Standards in Scottish Primary Schools*, University of London Press, London.

Szaleniec H., (2002) *Probabilistyczne modele wyniku zadania testowego (Item Response Theory)* (w:) B. Niemierko (red.) *Ocenianie szkolne, ekonomika i polityka oświatowa, probabilistyczne modele pomiaru.* Skrypt nr 3 dla uczestników III Podyplomowego Studium Ewaluacji Dydaktycznej na Uniwersytecie Gdańskim, Międzywydziałowe Studium Pedagogiczne UG, Gdańsk.