**dr Maciej Jakubowski**
Wydział Nauk Ekonomicznych, Uniwersytet Warszawski

# Wpływ rzetelności pomiaru umiejętności uczniów na edukacyjną wartość dodaną

## Wstęp

Rzetelność pomiaru umiejętności uczniów ma kluczowy wpływ na rzetelność ocen edukacyjnej wartości dodanej. Niska rzetelność pomiaru nie tylko zmniejsza precyzję ocen EWD, ale i może prowadzić do ich systematycznego zawyżenia lub zaniżenia, szczególnie w metodach wykorzystujących tylko dwukrotny pomiar umiejętności, takich jak stosowane w naszym kraju do oceny efektywności gimnazjów. Artykuł pokazuje, jak uwzględnienie błędu pomiaru wpływa na oszacowania EWD oraz w jaki sposób możliwe jest skorygowanie o błąd pomiaru wyników sprawdzianu w modelach EWD proponowanych do oceny gimnazjów w Polsce.

## Błąd pomiaru umiejętności uczniów

Testy edukacyjne określają poziom umiejętności uczniów jedynie z pewną dokładnością. Stopień powtarzalności wyników pomiaru osiągnięć uczniów na równoległych testach nazywamy rzetelnością. Im większa rzetelność (bliższa 1), tym bardziej podobne będą wyniki ucznia na powtórzonym teście. Wysoka rzetelność testów jest cechą jak najbardziej pożądaną. Oznacza ona, że pomiar umiejętności uczniów nie jest przypadkowy i odpowiada tej samej ukrytej własności, którą chcemy ocenić.

Błąd pomiaru jest ściśle związany z rzetelnością testu. O ile inne kwestie pozostają bez zmian, to wzrost błędu pomiaru obniża rzetelność testu i odwrotnie. Błąd pomiaru powodowany jest wieloma czynnikami – zarówno konstrukcją samego testu, jak i przypadkowymi zdarzeniami. W niniejszym artykule zakładamy, że nie jest on związany z prawdziwym, ukrytym poziomem umiejętności, posiada stałą wariancję, a jego średnia wynosi zero. W praktyce założenie o stałej wariancji jest często nieuzasadnione. Wiadomo, że błąd pomiaru rośnie wraz z oddalaniem się od środka skali pomiarowej, o ile do skalowania wykorzystywane są powszechne dziś techniki w rodzaju IRT, a testy przygotowywane są tak, aby jak najlepiej mierzyć zdolności jak największej liczby uczniów, a więc są najbardziej precyzyjne dla uczniów o przeciętnym poziomie umiejętności. W przypadku egzaminów zewnętrznych w Polsce nie jest łatwo określić zarówno ogólny poziom błędu pomiaru, jak i to dla jakich przedziałów ma on zmienne wartości. Można też przypuszczać, że błąd pomiaru jest niesystematyczny, chociażby dlatego, że związany jest z tzw. „efektem egzaminatora". Gdyby testy były przypisywane losowo egzaminatorom, to błąd pomiaru wynikający z różnic w ocenianiu można

by uznać za losowy, o średniej zero nawet dla dowolnych podgrup uczniów. Jednak gdy testy nie są losowo przydzielane egzaminatorom, tak jak w Polsce, błąd pomiaru w sposób systematyczny może wpływać na wyniki testów dla grup uczniów, w tym przypadku przede wszystkim dla uczniów z różnych OKE lub przypisanych różnym egzaminatorom. Kwestie te jednak pomijamy w niniejszym badaniu zakładając, że błąd pomiaru ma średnią zero i jest homoskedastyczny (ma stałą wariancję).

Głównym problemem, jaki rodzi niska rzetelność lub też wysoki błąd pomiaru, jest nieprecyzyjność, a przez to niewielka wiarygodność i mała przydatność testów umiejętności. W typowych zastosowaniach będzie to powodować, że diagnozowanie rozwoju uczniów będzie obarczone bardzo dużym błędem, być może niedopuszczalnym w danych zastosowaniach lub np. oceny poziomu umiejętności uczniów między szkołami będą bardzo zmienne, głównie ze względu na spory komponent losowy powodowany niską rzetelnością testów. Stąd zmniejszanie błędów pomiaru, ale i podnoszenie rzetelności testów przez zwiększanie ich spójności i stosowanie odpowiednich narzędzi testowych, są głównym celem każdej instytucji takie testy organizującej. Rzetelność testu ma też jednak znaczenie dla dalszego wykorzystania wyników egzaminów, np. do oceny efektywności szkół przez metody edukacyjnej wartości dodanej. W dalszej części artykułu skupimy się na tym zagadnieniu w kontekście modeli EWD dla polskich gimnazjów.

## Klasyczne podejście do błędu pomiaru zmiennych w regresji liniowej i wpływ błędu pomiaru na EWD

Jak wspomniano powyżej, przyjmujemy, że błąd pomiaru nie jest skorelowany z mierzoną cechą i ma w populacji średnią zero oraz stałą wariancję. Błąd pomiaru to po prostu losowy czynnik, którym obarczona jest obserwowalna cecha odzwierciedlająca interesującą nas zmienną ukrytą. To klasyczne podejście możemy zapisać w poniższy sposób:

$$x_i = x_i^* + m$$
$$m \sim N(0, \sigma_m 2)$$

gdzie $x_i$ to obserwowalny wynik ucznia na teście umiejętności, $x_i^*$ to ukryta cecha, którą test mierzy, a $m$ to błąd pomiaru o zerowej średniej E[m]=0 i stałej wariancji $\sigma_m^2$. Zakładamy przy tym, że błąd pomiaru nie jest skorelowany z poziomem prawdziwych, ukrytych umiejętności, a więc $\mathrm{Cov}[x_i^*, m] = 0$, z czego wynika, że jest skorelowany z obserwowalnym wynikiem, co odpowiada najczęściej rzeczywistości[1].

Szacując edukacyjną wartość dodaną dla gimnazjów w Polsce, wykorzystujemy model regresji liniowej, gdzie zmienną zależną jest wynik egzaminu gimnazjalnego, a wśród zmiennych niezależnych wynik sprawdzianu. W ten sposób objaśniając

[1] Jeśli błąd pomiaru nie jest skorelowany z obserwowalnym wynikiem, to musi być skorelowany z ukrytą cechą. Jednak w tym przypadku błąd pomiaru nie miałby wpływu na oszacowania regresji, jedynie powiększając niewyjaśnioną wariancję, podobnie jak przy błędzie pomiaru w zmiennej zależnej (por. Baum, 2006, s. 217.).

wyniki egzaminu gimnazjalnego, bierzemy pod uwagę poziom umiejętności ucznia po szkole podstawowej, a więc pozostałą wariancję możemy traktować jako zależną od przyrostu poziomu umiejętności w gimnazjum. EWD gimnazjów szacowane jest jako średnia z reszt (różnic między wynikami rzeczywistymi a przewidywanymi przez regresję dla całej populacji) dla danej szkoły lub bezpośrednio jako efekt stały lub losowy gimnazjum (por. Jakubowski, 2008). Metoda EWD zdominowała inne podejścia, jeśli chodzi o ilościową ocenę efektywności pracy szkół lub nauczycieli, jednak wykorzystywane w praktyce modele nie biorą pod uwagę błędu pomiaru, o którym wiadomo, że ma niepomijalny wpływ nawet w przypadku najbardziej wyrafinowanych narzędzi testowych (por. McCaffrey i in., 2005). Dla nas kwestia wpływu błędu pomiaru jest o tyle istotna, że w modelu regresji liniowej z dwoma pomiarami (egzamin gimnazjalny i sprawdzian lub egzamin gimnazjalny i matura) błąd pomiaru może zaburzać uzyskiwane oceny szkół znacznie silniej niż w modelach uwzględniających pomiar wielokrotny, najczęściej coroczny, tak jak w większości modeli wykorzystywanych w USA. Inaczej mówiąc, o ile błąd pomiaru może być pomijany w krajach, gdzie podstawą modeli EWD są wyniki corocznych lub częstszych testów z kilku przedmiotów, to już w Polsce może mieć on kluczowe znaczenie dla oceny EWD, bowiem dwukrotny pomiar, na dodatek ze zmienną strukturą mierzonych umiejętności, stanowi całą informację, jaką posiadamy o umiejętnościach ucznia.

Wpływ błędu pomiaru na EWD pokażemy przez prosty model regresji liniowej z jedną zmienną objaśniającą. Przyjmijmy, że $y_{ij}$ to wynik egzaminu gimnazjalnego *i*-tego ucznia *j*-tego gimnazjum, a $x_{ij}$ to wynik tego ucznia na sprawdzianie trzy lata wcześniej. Załóżmy, że egzamin gimnazjalny mierzony jest z błędem $m_y$ a sprawdzian z błędem $m_x$. Gwiazdką oznaczymy prawdziwe, ukryte poziomy umiejętności uczniów, które mierzą obydwa egzaminy. Wtedy równanie regresji liniowej EWD można zapisać jako:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}$$

a po uwzględnieniu błędu pomiaru jako:

$$y_{ij}^{*} + m_y = \beta_0 + \beta_1 (x_{ij}^{*} + m_x) + \varepsilon_{ij}$$

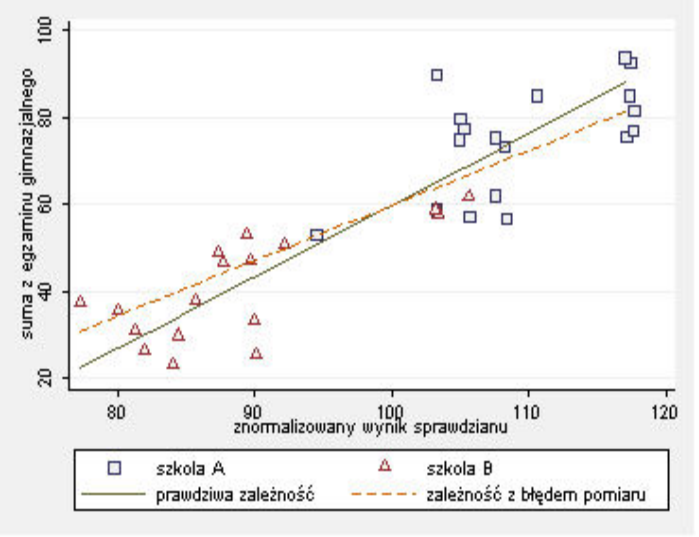
co po przekształceniu daje:

$$y_{ij}^{*} = \beta_0 + \beta_1 x_{ij}^{*} + (\beta_1 m_x - m_y + \varepsilon_{ij})$$

z czego wynika, że błąd pomiaru wyników egzaminu gimnazjalnego nie wpływa na oszacowania parametrów (pamiętajmy, że z założenia wartość oczekiwana błędu pomiaru wynosi 0), a jedynie zwiększa niewyjaśnioną wariancję, a przez to obniża precyzję pomiaru EWD. Jednak błąd pomiaru wyników sprawdzianu wpływa na EWD, bowiem powoduje, że składnik resztowy $(\beta_1 m_x - m_y + \varepsilon_{ij})$ jest skorelowany z regresorem $x_{ij}$, bo zgodnie z założeniem $\mathrm{Cov}[x_i^{*}, m] \neq 0$.

Kluczowym założeniem regresji liniowej jest brak korelacji między losowym błędem a zmiennymi zależnymi. Występowanie takiej korelacji powoduje, że oceny parametrów będą obciążone, a więc będą odbiegać od prawdziwych wartości niezależnie od wielkości próby. W przypadku, gdy $x_{ij}$ jest jedynym regresorem, to ocena parametru $\beta_1$ będzie zaniżona (por. Baum, 2006, s. 217.). W ten sposób uczniowie z niższymi wynikami sprawdzianu będą mieli wyższe przewidywane wyniki egzaminu gimnazjalnego, a przez to szkoły do których uczęszczają będą uzyskiwały niższą EWD, niż gdybyśmy znali (potrafili zmierzyć bez błędu) prawdziwą relację między poziomem umiejętności w szkole podstawowej i gimnazjum. EWD gimnazjów z uczniami o wyższych wynikach sprawdzianu będzie zawyżane.

Mechanizm ten pokazuje wykres poniżej, gdzie linią ciągłą zaznaczono „prawdziwą", nieobserwowalną bezpośrednio relację, a linią przerywaną relację między rzeczywistym wynikiem sprawdzianu a łącznym wynikiem egzaminu gimnazjalnego z roku 2007 (są to rzeczywiste dane, które analizujemy w dalszej części artykułu). Wybrano dwa przykładowe gimnazja o różnym poziomie uczniów „na wejściu". Ten rocznik w szkole A to uczniowie o wysokich wynikach sprawdzianu (mediana = 36), podczas gdy w szkole B to uczniowie o stosukowo niskich wynikach sprawdzianu (mediana = 22). EWD szkoły A mierzone bez korekty na błąd pomiaru wyniosło +3.7, a szkoły B -3.5. Po korekcie z założoną rzetelnością testu równą 0.77, EWD szkoły A spadło do +0.1, a EWD szkoły B wzrosło do +0.5. W jednym i drugim przypadku różnica jest bardzo wysoka i wyniosła ok. 4 punktów egzaminacyjnych (ok. 1 odchylenia standardowego EWD). Patrząc na rozrzut wyników uczniów i dopasowane linie regresji bez i z korektą na błąd pomiaru, powinno być jasne, jak wpływa to na pomiar EWD. Po korekcie linia staje się bardziej stroma, przez co dla wyższych wyników sprawdzianu przewidywane wyniki egzaminu gimnazjalnego są wyższe i odwrotnie w przypadku niższych wyników. O ile dla prawdziwej zależności linia regresji „przecina" grupę uczniów szkoły A, to dla zależności mierzonej z błędem większość tych uczniów leży powyżej linii. Dla szkoły B sytuacja jest dokładnie odwrotna. Dla zależności mierzonej z błędem pomiaru leżą oni poniżej prostej, a po korekcie prosta przecina ich grupę. Stąd EWD szkoły A maleje po korekcie, a EWD szkoły B rośnie, co będzie prawdą dla każdego gimnazjum o podobnym poziomie wyników sprawdzianu. Oczywiście szkoły te zostały dobrane tak, aby pokazać skrajne przypadki wpływu błędu pomiaru na EWD. Dla szkół z przeciętnymi wynikami zmiana będzie niewielka, a trzeba podkreślić, że będzie to większość gimnazjów.




**Wykres 1. Wpływ błędu pomiaru na oszacowanie zależności między sprawdzianem i egzaminem gimnazjalnym**

Ogólnie rzecz biorąc, błąd pomiaru w zmiennej zależnej zwiększa jedynie niewyjaśnioną wariancję, o ile nie mamy do czynienia z modelem z efektami losowymi. Natomiast błąd pomiaru w zmiennej zależnej zaniża oszacowany współczynnik przy tej zmiennej, a jeśli w modelu jest więcej zmiennych, to ich współczynniki mogą być zaniżone lub zawyżone. W modelach z efektami losowymi sprawa się dodatkowo komplikuje, bo błąd pomiaru w zmiennej zależnej wpływa na oszacowanie wariancji reszt, a przez to na oszacowanie rozkładu zmiennej ukrytej (w EWD jest to domniemana jakość nauczania szkół). W każdym przypadku korekta modelu jest tym bardziej potrzebna, im większy błąd pomiaru i im mniej pomiarów.

## Baza danych i wyniki badania empirycznego dla Polski

Istnieje stosunkowo niewiele prac empirycznych podejmujących problem błędu pomiaru w kontekście modeli edukacyjnej wartości dodanej[2]. Jak już zaznaczono, w modelach z wielokrotnym pomiarem losowy błąd pomiaru powinien się „znosić", o ile model został odpowiednio skonstruowany, jednak w każdym przypadku rzetelność testu wpływa na oceny edukacyjnej wartości dodanej szkół czy nauczycieli (por. McCaffrey i in., 2005). W modelach wielopoziomowych, z efektami losowymi, błąd pomiaru może być uwzględniony w rozmaity sposób, także dla zmiennych zależnych, jednak metody te są często wrażliwe na założenia

[2] Ogólny opis modelowania EWD można znaleźć w: Raudenbush, Bryk, 2002; McCaffrey i in., 2005. Polska metoda została opisana w: Dolata (red.), 2007; Jakubowski, 2008.

przyjmowane przez badacza. Z tego zapewne powodu błąd pomiaru jest w praktyce rzadko uwzględniany w innych niż naukowe zastosowaniach EWD.
Ladd i Walsh (2002) pokazali, w jaki sposób błąd pomiaru wpływa na oceny EWD w jednym ze stanów USA, gdzie używano metody zbliżonej do modelu EWD dla polskich gimnazjów. Poniżej powtarzamy podejście Ladd i Walsh, opierające się na zmiennych instrumentalnych. W naszym przypadku wynik sprawdzianu będzie korygowany przez wynik egzaminu próbnego[3]. Regresja ze zmiennymi instrumentalnymi to typowy sposób korekty błędu pomiaru, o ile posiadamy kilka pomiarów tej samej zmiennej (por. Greene, 2003). O ile takich danych nie posiadamy, a znana jest za to rzetelność testu, to możemy ją bezpośrednio wykorzystać do skorygowania modelu regresji i np. uzyskania oszacowania nachylenia wyników sprawdzianu po uwzględnieniu błędu pomiaru. Odpowiednie modele dla regresji ze zmienną instrumentalną lub bezpośrednią korektą na błąd pomiaru można znaleźć w większości podręczników do ekonometrii czy statystyki (por. Greene, 2003; Wooldridge, 2002). W badaniu wykorzystaliśmy procedury -ivreg- (dla regresji ze zmiennymi instrumentalnymi, tzw. „instrumental variables regression") oraz -eivreg- (dla korekty na błąd pomiaru przez podanie rzetelności, tzw. „errors-in-variables regression") w programie Stata. Czytelników odsyłamy do dokumentacji (StataCorp, 2007) oraz pracy Bauma (2006) po dokumentację statystyczną i zastosowanie praktyczne tych procedur.

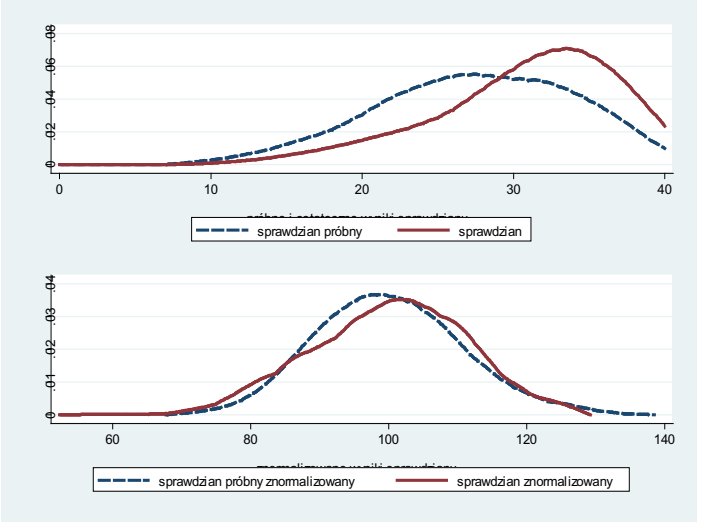
Dane wykorzystane do badania pochodzą z badania pilotażowego przeprowadzonego w powiecie oświęcimskim przez pracowników OKE Kraków (por. Szmigel, Rappe, 2005)[4]. W badaniu tym wyniki próbne i ostateczne sprawdzianu zostały połączone z wynikami egzaminu gimnazjalnego dla 839 uczniów, którzy początkowo uczęszczali do 14 szkół podstawowych, a potem do 83 gimnazjów. Ze względu na powtórny pomiar powstaje możliwość oszacowania błędu pomiaru sprawdzianu, a dołączenie wyników uczniów z egzaminu gimnazjalnego daje możliwość oszacowania modeli EWD. W analizach EWD z bazy odrzucono gimnazja z mniej niż 5 uczniami, ze względu na zbyt małą ilość informacji, pozostawiając 734 uczniów w 25 gimnazjach. Jest to próba niewielka, lecz wystarczająca do oszacowania modeli EWD i określenia wpływu błędu pomiaru na uzyskiwane oceny szkół.

Mimo że sprawdzian w obydwu wersjach powinien stanowić równoważny test, to uzyskane rozkłady są znacząco odmienne. Prawdziwy egzamin został napisany przez większość uczniów znacznie lepiej (średnia wyniosła 27.5 w egzaminie próbnym i 30.9 w ostatecznym), zmniejszyło się także zróżnicowanie wyników (z 6.3 do 5.8). Wyniki testów mają różne rozkłady, co ogranicza ich bezpośrednią

[3] W oryginalnym artykule Ladd i Walsh wykorzystali wyniki w czwartej klasie jako instrument dla pomiaru w klasie piątej w modelu objaśniającym wyniki w klasie szóstej. Kluczowym wymogiem jest, aby instrument był silnie skorelowany ze zmienną obarczoną błędem pomiaru, ale nie z błędem losowym (por. Wooldridge, 2002).

[4] Za udostępnienie danych Autor chciałby podziękować Krystynie Szmigel i Annie Rappe.

porównywalność, szczególnie jeśli wykorzystywane są metody liniowe, wrażliwe na kształt rozkładów i skale pomiaru. Aby uzyskać porównywalność dwóch wersji sprawdzianu oraz liniową zależność wyników sprawdzianu z wynikiem egzaminu gimnazjalnego, przeniesiono wyniki obydwu testów na skalę normalną o średniej 100 i odchyleniu standardowym 10[5]. Dokonano tego przez posortowanie wg rosnących wartości dwóch równolicznych zbiorów, pierwszego z wynikiem jednego z egzaminów, drugiego z losowym rozkładem zmiennej normalnej, a następnie ich połączenie, zachowując rosnący porządek w każdym z nich. W kolejnym kroku dla każdego wyniku sprawdzianu przypisano średnią wartość zmiennej normalnej między tym wynikiem a wynikiem niższym o jeden. W ten sam sposób przełożono na skalę normalną egzamin próbny jak i jego ostateczną wersję. Dla egzaminu gimnazjalnego nie było to niezbędne, ponieważ suma wyników z obydwu części ma rozkład zbliżony do normalnego i nie potrzebna była tu żadna korekta, szczególnie biorąc pod uwagę większą precyzję oceny (od 15 do 100 punktów w naszym zbiorze)[6]. Rozkłady wyników sprawdzianu przed i po znormalizowaniu przedstawia wykres poniżej.




**Wykres 2. Rozkład wyników sprawdzianu próbnego i ostatecznego przed i po przełożeniu na skalę normalną**

[5] Dla kilku uczniów przekodowano wyniki testu próbnego na minimalną wartość wyników testu ostatecznego (12), a dla wszystkich dodano 1, tak aby minimum i maksimum w obydwu testach było takie samo. Trzeba podkreślić, że jakiekolwiek przekształcenia liniowe tego rodzaju nie mają wpływu na współczynniki korelacji czy regresji liniowej, a przez to na EWD. Normalizacja także nie zmienia uporządkowania szkół, a jedynie ułatwia specyfikację równania regresji.

[6] W badaniu wykorzystano sumę wyników egzaminu gimnazjalnego z obydwu części. Analizy przeprowadzono także osobno dla części humanistycznej i matematyczno-przyrodniczej, uzyskując jakościowo podobne rezultaty.

Dla znormalizowanych rozkładów wyników obydwu wersji sprawdzianu obliczono korelację między nimi, która wyniosła ok. 0.77 (podobnie zresztą jak dla wyników nieprzetworzonych). W ten sposób można w najprostszy sposób określić rzetelność testu. Innym popularnym sposobem jest obliczanie statystyki Alfa Cronbacha, choć dla tego rodzaju testów wydaje się ona mało wiarygodna. W istocie, Alfa Cronbacha wynosi w tej próbie ok. 0.85, co można z góry przyjąć za wartość zawyżoną. Niemniej jednak te dwie wartości zostaną wykorzystane w dalszych analizach jako potencjalne wskaźniki rzetelności sprawdzianu.

Pierwsza tabela zawiera porównanie dwóch podstawowych modeli EWD: z efektami stałymi oraz losowymi, oszacowanymi początkowo bez korekty na błąd pomiaru, a następnie z korektą przez wykorzystanie wyników egzaminu próbnego jako zmiennej instrumentalnej (IV) dla ostatecznych wyników sprawdzianu. W modelach uwzględniono zmienne zerojedynkowe obecne w modelu EWD dla gimnazjów: płeć i dysleksję. Jak widać w obydwu modelach, zarówno z efektami stałymi, jak i losowymi, korekta na błąd pomiaru sprawdzianu zmienia zasadniczo oszacowania współczynników regresji. Nachylenie wyników sprawdzianu w obydwu przypadkach zmienia się z ok. 1.2 do 1.6, a więc bardzo znacznie. Zmianie ulegają także oceny parametrów dla płci oraz dysleksji.

**Tabela 1. Modele EWD z efektami stałymi i losowymi, bez korekty i z korektą błędu pomiaru wyników sprawdzianu przez zmienne instrumentalne (IV)**

| Model EWD | Efekty stałe | | Efekty losowe | |
|---|---|---|---|---|
| | bez korekty | z korekt IV | bez korekty | z korekt IV |
| Płe | 1.28<br>(0.75) | 0.57<br>(0.81) | 1.39<br>(0.75) | 0.62<br>(0.80) |
| Dysleksja w gimnazjum | 5.37***<br>(1.42) | 5.01***<br>(1.51) | 5.73***<br>(1.42) | 5.19***<br>(1.50) |
| Dysleksja w szkole podstawowej | -5.27***<br>(1.51) | -4.00*<br>(1.61) | -5.12***<br>(1.50) | -3.99*<br>(1.59) |
| Wynik sprawdzianu | 1.22***<br>(0.04) | 1.60***<br>(0.06) | 1.24***<br>(0.04) | 1.61***<br>(0.06) |
| Stała | -62.71 | -100.51 | -65.05 | -101.61 |
| $R^2$ | 0.59 | 0.59 | 0.59 | 0.59 |
| Liczba gimnazjów | 25 | 25 | 25 | 25 |
| Liczba uczniów | 734 | 734 | 734 | 734 |

Błędy standardowe w nawiasach. ***, **, * oznaczają odpowiednio istotność na poziomie 1%, 5%, 10%

Kolejna tabela przedstawia oszacowania zwykłej regresji liniowej, która w podstawowym modelu dla gimnazjów w Polsce służy do obliczania EWD jako średniej z reszt (średniej z różnic między wynikiem rzeczywistym a przewidywanym przez regresję dla uczniów danego gimnazjum). Regresję oszacowano bez korekty na błąd pomiaru oraz z korektą dwoma metodami. Pierwsza metoda to ponownie regresja z wynikami próbnego sprawdzianu jako zmienną instrumentalną dla wyników sprawdzianu (IV). Druga metoda koryguje macierz wariancji–kowariancji zwykłej regresji liniowej o z góry założoną rzetelność pomiaru wyników sprawdzianu (procedura -eivreg-, patrz opis w: StataCorp, 2007). Oszacowano trzy regresje dla rzetelności na poziomie 0.77 (takiej jak korelacja między próbnym i ostatecznym sprawdzianem), 0.85 (wartość statystyki Alfa Cronbacha) oraz 0.90. Ponownie oceny parametrów dla poszczególnych zmiennych ulegają istotnej zmianie. Korekta metodą zmiennych instrumentalnych daje podobny współczynnik nachylenia sprawdzianu jak wcześniejsze modele. Korekta o założoną rzetelność równą 0.77 także daje podobne oszacowania. Przyjęcie wyższych rzetelności obniża interesujący nas współczynnik zgodnie z oczekiwaniami.

**Tabela 2. Modele EWD opierające się na zwykłej regresji liniowej bez i z korektą błędu pomiaru wyników sprawdzianu przez zmienne instrumentalne oraz przez model z założoną rzetelnością pomiaru**

| Model EWD | Regresja liniowa | | Regresja z założoną rzetelnością pomiaru sprawdzianu równą | | |
|---|---|---|---|---|---|
| | Bez korekty | Z korektą IV | 0.77 | 0.85 | 0.90 |
| Płeć | 1.86*<br>(0.81) | 1.05<br>(0.85) | 0.98<br>(0.62) | 1.34<br>(0.70) | 1.53*<br>(0.74) |
| Dysleksja w gimnazjum | 8.16***<br>(1.51) | 7.42***<br>(1.59) | 7.35***<br>(1.16) | 7.68***<br>(1.32) | 7.86***<br>(1.39) |
| Dysleksja w szkole podstawowej | -4.95**<br>(1.59) | -4.34**<br>(1.67) | -4.28***<br>(1.22) | -4.56***<br>(1.38) | -4.70**<br>(1.46) |
| Wynik sprawdzianu | 1.26***<br>(0.04) | 1.61***<br>(0.05) | 1.64***<br>(0.04) | 1.48***<br>(0.04) | 1.40***<br>(0.04) |
| Stała | -67.45 | -102.11 | -105.34 | -89.79 | -81.50 |
| $R^2$ | 0.59 | 0.55 | 0.76 | 0.69 | 0.66 |
| Liczba gimnazjów | 25 | 25 | 25 | 25 | 25 |
| Liczba uczniów | 734 | 734 | 734 | 734 | 734 |

Błędy standardowe w nawiasach. ***, **, * oznaczają odpowiednio istotność na poziomie 1%, 5%, 10%

Powyższe oszacowania pokazują, że uwzględnienie błędu pomiaru znacząco zmienia oceny parametrów w regresji, a przez to zapewne wpływa na predykcje EWD gimnazjów. Zgodnie z teorią, skorygowanie o błąd pomiaru podnosi nachylenie wyników sprawdzianu, co jak pokazaliśmy na wcześniejszym przykładzie (Wykres 1.), obniża EWD szkół z uczniami o wysokich wynikach sprawdzianu i podnosi EWD szkół z uczniami o niskich wynikach. Wykres 1. przedstawiał jednak jaskrawy przykład dwóch szkół o wyjątkowo niskiej i wysokiej średniej ze sprawdzianu. Nie jest jednak jasne, na ile korekta o błąd pomiaru wpływa na większość szkół.

Częściowej odpowiedzi na pytanie o wpływ błędu pomiaru na EWD gimnazjów udzielają tabele poniżej. Pierwsza z nich (Tabela 3.) zawiera współczynniki korelacji między EWD liczonym różnymi metodami przed i po korekcie błędu pomiaru sprawdzianu metodą zmiennych instrumentalnych. Jak widać korelacje są bardzo wysokie i minimalnie wyższe po korekcie błędów pomiaru.

**Tabela 3. Współczynniki korelacji między EWD liczonym różnymi metodami przed i po korekcie błędów pomiaru metodą zmiennych instrumentalnych**

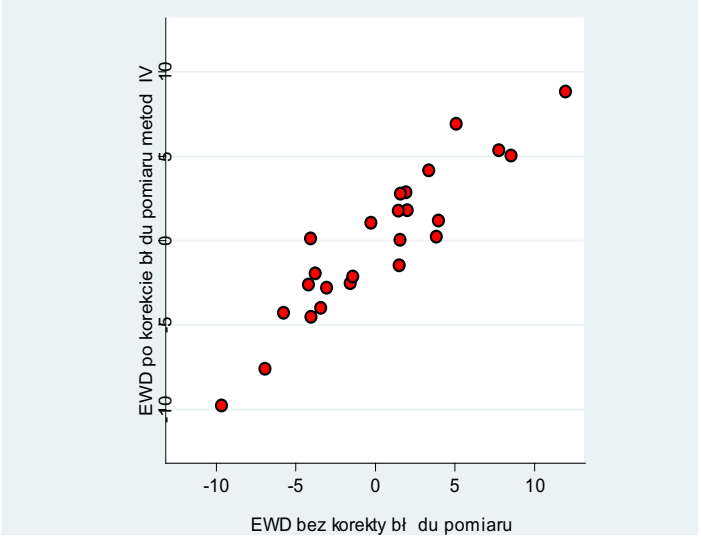
| | bez korekty | | po korekcie | |
|---|---|---|---|---|
| | *rednia reszt* | *efekt stały* | *rednia reszt* | *efekt stały* |
| | | Pearsona | | |
| efekt stały | 1.00 | | 1.00 | |
| efekt losowy | 0.96 | 0.96 | 1.00 | 1.00 |
| | | Spearmana | | |
| efekt stały | 0.99 | | 0.99 | |
| efekt losowy | 0.98 | 0.98 | 0.99 | 1.00 |
| | | Kendalla | | |
| efekt stały | 0.94 | | 0.95 | |
| efekt losowy | 0.91 | 0.90 | 0.96 | 0.99 |

Tabela 4. zawiera współczynniki korelacji dla EWD liczonego każdą z trzech podstawowych metod przed i po korekcie błędu pomiaru sprawdzianu. Tym razem korelacja określa zmianę dla tej samej metody po wprowadzeniu korekty na błąd pomiaru, a nie zmianę w spójność EWD liczonej różnymi metodami, jak poprzednio. Korelacje można uznać za dość wysokie, jednak dla tak małego zbioru korelacje rangowe oznaczają, że sporo szkół zmienia swoje miejsce w „rankingu". Błąd pomiaru ma więc widoczny wpływ na oceny EWD. Trzeba podkreślić, że wpływ ten jest dużo większy niż wynikający ze zmiany metody liczenia EWD (np. przejścia od efektów stałych do losowych).

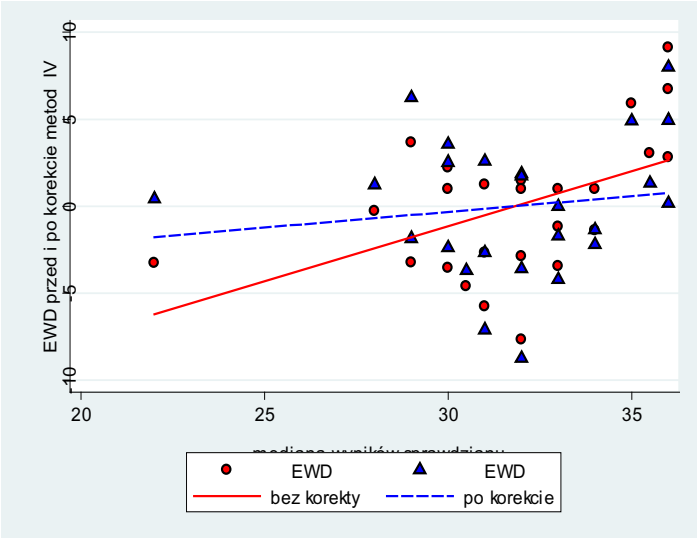
**Tabela 4. Korelacja między EWD dla różnych metod przed i po korekcie błędu pomiaru sprawdzianu**

| *Metoda EWD* | *metoda korekty bł du pomiaru* | współczynniki korelacji (n=25) | | |
|---|---|---|---|---|
| | | *Persona* | *Spearmana* | *Kendalla* |
| rednia z reszt | zmienna instrumentalna | 0.93 | 0.92 | 0.77 |
| | korekta dla r=0.77 | 0.92 | 0.90 | 0.75 |
| | korekta dla r=0.85 | 0.97 | 0.96 | 0.86 |
| | korekta dla r=0.90 | 0.99 | 0.98 | 0.91 |
| efekt stały | zmienna instrumentalna | 0.92 | 0.90 | 0.76 |
| efekt losowy | zmienna instrumentalna | 0.94 | 0.93 | 0.83 |

Zależność między EWD przed i po korekcie błędów pomiaru pokazują dwa wykresy poniżej, gdzie wykorzystano EWD liczone jako efekt stały przed i po korekcie błędów pomiaru metodą zmiennych instrumentalnych. Wykres 3. pokazuje rozrzut EWD liczonego bez korekty względem EWD liczonego z korektą. Wykres 4. pokazuje zależność między EWD przed i po korekcie a medianą wyników sprawdzianu uczniów gimnazjum. Widać, że przed korektą zależność EWD od poziomu umiejętności uczniów na progu nauki w gimnazjum jest dość silna i pozytywna. Po korekcie zależność ta staje się nieistotna statystycznie.




**Wykres 3. EWD (efekt stały) przed i po korekcie błędu pomiaru metodą zmiennych instrumentalnych**




**Wykres 4. Zależność między EWD (efekt stały) przed i po korekcie, a poziomem uczniów na wejściu (medianą wyników sprawdzianu w gimnazjum)**

Analiza przeprowadzona powyżej sugeruje, że korekta błędu pomiaru wyników sprawdzianu może mieć kluczowe znaczenie dla uzyskania wiarygodnych ocen EWD, które nie zależałyby od poziomu uczniów na wejściu do szkoły. Oczywiście chodzi nam jedynie o pozorną zależność spowodowaną wpływem na EWD błędu pomiaru wyników sprawdzianu. Nawet po korekcie takiej zależności nie można wykluczyć, a nawet można się jej spodziewać, bowiem trudno polemizować z zasadnością hipotezy, że szkoły bardziej efektywne mogą przyciągać uczniów o wyższych wynikach sprawdzianu. Naszym celem jest próba określenia, czy po korekcie błędu pomiaru wyników sprawdzianu korelacja między średnim poziomem uczniów w szkole na wejściu a oceną EWD jest wciąż istotna statystycznie, czy też w całości wynika z braku korekty w podstawowym modelu.

Niestety, analizy przedstawione powyżej dotyczą niewielkiej liczby szkół i są trudne do uogólnienia na całą populację zarówno ze względu na niewielki rozmiar próby, jak i na jej nielosowy charakter. Jednak tylko dla tej próby możemy określić rzetelność metodą test–retest. Załóżmy jednak, że dla całej populacji uczniów piszących sprawdzian w 2003 roku rzetelność testu była podobna lub nieco wyższa, chociażby dlatego, że na wyniki próbnego sprawdzianu wpływa niższa motywacja uczniów i szkół, niż w przypadku egzaminu końcowego. Przyjmując prawdopodobne poziomy rzetelności, możemy określić, na ile na modele EWD dla całej populacji są zależne od błędu pomiaru wyników sprawdzianu.

Tabela 5. poniżej zawiera oszacowania zwykłej regresji liniowej EWD dla niemal całej populacji uczniów piszących sprawdzian w 2003 roku i egzamin gimnazjalny w 2006 roku (dla mniej niż 10% uczniów nie można było połączyć wyników (por. Dolata (red.), 2007). Równania skorygowano dla założonej rzetelności pomiaru równej 0.77, 0.80 i 0.90. Nachylenie wyników sprawdzianu jest w całej populacji znacznie większe niż w badanej wcześniej grupie szkół. Jednak korekta o błąd pomiaru dodatkowo podwyższa współczynnik nachylenia sprawdzianu, dla r=0.80 wynosi on 1.76, a przed korektą 1.41. Zmieniają się także wartości innych współczynników. Tak więc dla całej populacji korekta o błąd pomiaru może mieć znaczący wpływ dla ocen EWD.

**Tabela 5. Modele EWD z korektą na błąd pomiaru sprawdzianu 2003 dla całej populacji uczniów zdających egzamin gimnazjalny w 2006 roku**

| Model EWD regresji liniowej | Bez korekty | Korekta dla rzetelno ci równej | | |
|---|---|---|---|---|
| | | 0.77 | 0.80 | 0.90 |
| Płe | 1.13*** (0.03) | 0.33*** (0.02) | 0.46*** (0.02) | 0.83*** (0.03) |
| Dysleksja w gimnazjum | 4.84*** (0.06) | 4.89*** (0.04) | 4.88*** (0.05) | 4.86*** (0.06) |
| Dysleksja w szkole podstawowej | -4.57*** (0.07) | -4.65*** (0.05) | -4.64*** (0.06) | -4.60*** (0.07) |
| Wynik sprawdzianu | 1.41*** (0.00) | 1.83*** (0.00) | 1.76*** (0.00) | 1.57*** (0.00) |
| Stała | -83.50 | -125.65 | -118.76 | -99.15 |
| $R^2$ | 0.63 | 0.81 | 0.78 | 0.70 |
| Liczba gimnazjów | 6359 | 6359 | 6359 | 6359 |
| Liczba uczniów | 480990 | 480990 | 480990 | 480990 |

Błędy standardowe w nawiasach. ***, **, * oznaczają odpowiednio istotność na poziomie 1%, 5%, 10%

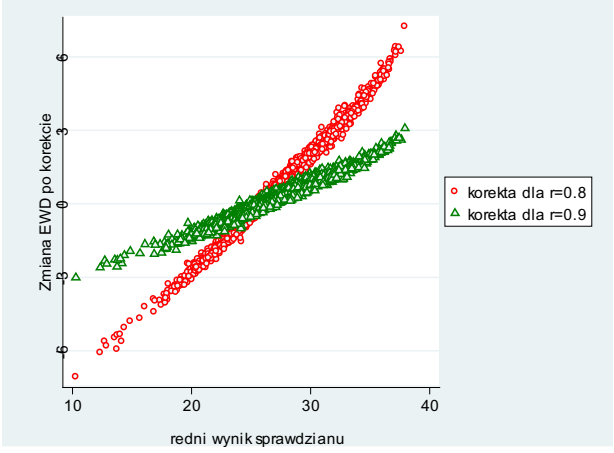
Tabela 6. zawiera współczynniki korelacji między EWD liczoną jako średnia z reszt równania regresji przed i po zastosowaniu korekty z założonymi poziomami rzetelności pomiaru. Dodatkowo przedstawiono współczynniki korelacji z medianą wyników sprawdzianu wśród uczniów danego gimnazjum. W porównaniu z wcześniejszą regresją usunięto gimnazja z mniej niż 10 uczniami, aby zwiększyć odporność analizy na niewielką liczbę obserwacji.

**Tabela 6. Współczynniki korelacji Pearsona między EWD (liczonym jako średnia reszt) przed i po korekcie ze względu na błąd pomiaru oraz medianą wyników sprawdzianu uczniów danego gimnazjum**

| *EWD jako rednia reszt* | bez korekty | Mediana sprawdzianu |
|---|---|---|
| bez korekty | - | 0.24 |
| korekta dla r=0.77 | 0.91 | -0.17 |
| korekta dla r=0.80 | 0.94 | -0.11 |
| korekta dla r=0.90 | 0.99 | 0.09 |

Widać, że uwzględnienie korekty ze względu na błąd pomiaru powoduje, że poziom umiejętności uczniów na wejściu do gimnazjum przestaje korelować z EWD. Jest to prawdą dla każdego z przyjętych poziomów rzetelności. Można więc przypuszczać, że widoczna w nieskorygowanych o błąd pomiaru ocenach EWD pozytywna korelacja z poziomem uczniów na wejściu wynika jedynie z nieuwzględnienia błędu pomiaru w regresji liniowej. Choć w całej populacji relacja ta nie jest silna, to jednak systematycznie podnosi EWD szkół ze zdolniejszymi uczniami i zaniża EWD szkół z uczniami słabszymi. Dla niektórych szkół uwzględnienie korekty błędu pomiaru ma spore znaczenie.

Ostatni wykres pokazuje zależność między różnicą w EWD liczonym bez korekty i z korektą (dla r=0.8 oraz r=0.9) a średnim wynikiem sprawdzianu wśród uczniów gimnazjum. Zależność ta wynika co prawda w bezpośredni sposób z oszacowanych współczynników regresji przedstawionych w Tabeli 5., jednak uzmysławia zarówno systematyczność zmiany w EWD szkół po uwzględnieniu korekty na błąd pomiaru, jak i znaczny rozmiar tego wpływu dla szkół przyjmujących uczniów o bardzo niskim lub bardzo wysokim poziomie umiejętności po szkole podstawowej. Widać, że zarówno zaniżenie, czyli przyjęcie r=0.8, gdyby prawdziwa rzetelność była równa 0.9, jak i zawyżenie rzetelności przez przyjęcie odwrotnych założeń, ma duże znaczenie dla niektórych szkół.




**Wykres 5. Zmiana EWD po korekcie na rzetelność pomiaru sprawdzianu r=0.8 i r=0.9 względem średniego wyniku sprawdzianu uczniów gimnazjum**

## Podsumowanie

Artykuł analizuje kwestię błędu pomiaru umiejętności uczniów w modelowaniu edukacyjnej wartości dodanej. W pierwszej części omówiono podstawowe pojęcia, klasyczne podejście do błędu pomiaru i teoretyczny wpływ na oszacowania regresji liniowej. W drugiej części zastosowano kilka metod korekty błędów pomiaru w wynikach sprawdzianu dla modeli EWD gimnazjów w Polsce. W badaniu wykorzystano dane z pilotażu sprawdzianu w powiecie oświęcimskim, gdzie połączono wyniki sprawdzianu próbnego z ostatecznymi wynikami sprawdzianu oraz egzaminu gimnazjalnego dla ok. 800 uczniów uczęszczających do kilkudziesięciu gimnazjów. Przed analizą oba pomiary sprowadzono do skali normalnej tak, aby skala pomiaru nie miała wpływu na uzyskane rezultaty. Zastosowano dwie metody korekty błędów pomiaru. Pierwsza, to metoda regresji ze zmiennymi instrumentalnymi, gdzie za instrument dla wyników sprawdzianu posłużył wcześniejszy egzamin próbny. Druga metoda polegała na określeniu rzetelności sprawdzianu przez obliczenie korelacji między wynikami obydwu testów (próbnego i końcowego), która wyniosła 0.77. Następnie oszacowano równania regresji, z góry zakładając taki, a także nieco wyższe poziomy rzetelności. Dla obydwu metod uzyskane rezultaty pokazują, że wpływ błędu pomiaru na oceny EWD jest widoczny i większy niż wybór samej metody szacowania EWD.

Pokazano także, że po skorygowaniu o błąd pomiaru znika pozytywna zależność między EWD a poziomem umiejętności uczniów gimnazjum po szkole podstawowej. Bez korekty istnieje pozytywna zależność, tzn. szkoły z uczniami o wyższych wynikach sprawdzianu mają też wyższą EWD. Po korekcie, niezależnie od metody i przyjętych poziomów rzetelności, zależność ta przestaje być istotna statystycznie. Analizy powtórzono dla pełnej populacji uczniów kończących naukę w gimnazjach w 2006 roku. Także tu stwierdzono, że dla prawdopodobnych poziomów rzetelności sprawdzianu (od 0.8 do 0.9) skorygowanie równań regresji o błąd pomiaru powoduje, że zależność między EWD a poziomem uczniów na wejściu znika. Dla niektórych szkół różnica między EWD liczonym bez korekty i po uwzględnieniu błędu pomiaru jest znaczna, szczególnie dla szkół z uczniami o bardzo niskim lub bardzo wysokim poziomie wyników sprawdzianu. Pokazuje to, że uwzględnienie błędu pomiaru w modelach EWD ma duże znaczenie, jednak niesie ze sobą także sporo ryzyko błędu, o ile zakładana rzetelność będzie daleka od prawdziwej.

**Bibliografia:**

1. Dolata R., *Edukacyjna wartość dodana jako metoda oceny efektywności nauczania na podstawie egzaminów zewnętrznych*, Centralna Komisja Egzaminacyjna 2007.
2. Baum C. F., *Introduction to Modern Econometrics Using Stata*, Stata Press, 2006.
3. Greene W. H., *Econometric Analysis*, Wydanie V, Prentice–Hall, 2003.
4. Jakubowski M., *Implementing value-added models of school assessment,* European University Institute RSCAS 2008/06 working paper, 2008.
5. Ladd H., Walsh R., *Implementing value-added measures of school effectiveness: Getting the incentives right*, Economics of Education Review, 2002, s. 21, 1–17.
6. McCaffrey D., Lockwood J., Koretz M., Hamilton L., *Evaluating Value–Added Models for Teacher Accountability*, Rand Corporation MG–158, 2005.
7. Raudenbush S. W., Bryk A., *Hierarchical Linear Models***,** wydanie II, Sage Publications, 2002.
8. StataCorp, *Stata Statistical Software. Release 10*, StataCorp, 2007.
9. Szmigel K., Rappe A., *Przydatność uproszczonej (staninowej) metody szacowania wartość dodanej osiągnięć uczniów, szkół i jednostek administracyjnych oświaty"* [w:] *Holistyczne i analityczne metody diagnostyki edukacyjnej, perspektywy informatyczne egzaminów szkolnych* pod red. B. Niemierko i G. Szyling, Gdańsk 2005.
10. Wooldridge J. M., *Econometric Analysis of Cross Section and Panel Data,* MIT Press, 2002.