

Sławomir Sapanowski

Okręgowa Komisja Egzaminacyjna w Łodzi

O pewnym aspekcie konstrukcji testu wielostopniowego

„Test sprawdzający wielostopniowy różni się od testu jednostopniowego *hierarchią wymagań*, to jest porządkiem, w którym spełnienie wyższych wymagań obejmuje (bezwzględnie lub z wysokim prawdopodobieństwem) spełnienie niższych wymagań. [...] Gdy uczeń spełnił wymagania wyższego poziomu (np. rozszerzające), a nie spełnił wymagań niższego poziomu (np. podstawowych), jego wynik jest *niestopniowalny*, to znaczy, nie spełnia wymagań przypisanych do jakiegokolwiek stopnia szkolnego”. Tak na temat testu wielostopniowego pisał prof. Bolesław Niemierko (1999, s. 72).

Szkolne testy sprawdzające opanowanie wiedzy i umiejętności powinny być, z założenia, testami wielostopniowymi. Niestety takimi z reguły nie są. Przyczyny tego są złożone i wielorakie. Na pewno jest to spowodowane brakiem dobrych przykładów takich sprawdzianów oraz powszechnego dostępu do opracowań dotyczących tworzenia wielostopniowych testów, jak również pewną ignorancją nauczycieli tworzących swoje testy.

Przykładem takiego postępowania nauczycieli może być dość często spotykana praktyka szkolna polegająca na stosowaniu sprawdzianów składających się z pięciu zadań o podobnym stopniu trudności (!). Nauczyciel wystawia ocenę na zasadzie – liczba poprawnie rozwiązanych zadań odpowiada ocenie, jaką uczeń uzyskuje z testu. System taki posiada jednak cechę całkowicie go dyskwalifikującą. Zastosowany w klasie, w której uczniowie posiadają na tyle wysokie umiejętności, że potrafią rozwiązać wszystkie zadania z testu, będą mieć co prawda różne oceny, ale zróżnicowanie ocen wynikać jednak będzie jedynie z różnic między tempem pracy reprezentowanym przez uczniów. A zatem ocena odzwierciedlać będzie tempo pracy, a nie rzeczywisty poziom osiągnięć ucznia.

W sposób naturalny powstaje pytanie: czy istnieje idealny test wielostopniowy? Oczywiście, nie! Czy możemy zbliżyć się do ideału? Badania pokazują, że tak. Jak tego dokonać? Temu ma służyć niniejsze opracowanie.

Rozpoczynając konstrukcję testu wielostopniowego, powinniśmy określić wymagania, które uczeń musi spełnić, aby dany poziom osiągnąć (mówiąc kolokwialnie „zaliczyć”) oraz liczbę tychże stopni¹, czyli określić taksonomię wymagań (zakres oraz trudność zadań, których rozwiązanie będzie konieczne do osiągnięcia zakładanego poziomu wymagań). Możemy tutaj skorzystać z wielu opracowań określających zakres wiadomości i umiejętności przewidywanych na danym poziomie rozwojowym ucznia (Niemierko, 1999, 2005).

¹ W przypadku testów szkolnych naturalną liczbą poziomów jest 6, które odpowiadają ocenom szkolnym.

W szkolnych systemach oceniania (WSO) najczęściej stosuje się jednak sposób przeliczania wyniku testu na ocenę szkolną na podstawie przelicznika procentowego². Liczba uzyskanych punktów w teście wyrażana jest w skali procentowej i ta determinuje ocenę szkolną. Nie dziwi zresztą takie podejście, bowiem nauczyciel, który chciałby ten stan zmienić, nie dysponuje ani wiedzą, ani możliwościami, aby przeciwstawić się obowiązującemu i ugruntowanemu systemowi oceniania w danej placówce oświatowej.

Zastanówmy się zatem, jak powinien wyglądać test wielostopniowy zbliżony do ideału. Otóż naturalne wydaje się, że powinniśmy wymagać od testu, aby uczeń, który otrzymał „czwórkę”, czyli przekroczył trzeci stopień (dwa pierwsze to „dwójka” i „trójka”), rozwiązał wszystkie zadania przypisane do stopni niższych. Z oczywistych powodów tak nie jest. Zdający, nawet najlepszy, może pomylić się w rozwiązywaniu zadania łatwego, a ten o niskim potencjale udzielić poprawnej odpowiedzi na zadanie trudne (co przy zadaniach zamkniętych zdarza się nader często).

Zdefiniujmy współczynnik *stopniowości testu* (S_n), który posłuży do oceny przydatności badanego testu. Niech wskaźnik ten określa, **jaka część zdających, która osiągnęła stopień n lub wyższy, rozwiązała wszystkie zadania ze stopni niższych**. Stopniowość testu może przyjmować wartości od 0 do 1. Wartość 0 odpowiada sytuacji, gdy wszyscy poddani testowaniu popełnili błędy na niższych poziomach. Z kolei stopniowość równa 1 oznaczać będzie, że każdy rozwiązujący test uczestnik, który spełnił wymagania stopnia n , nie popełnił błędów na stopniach niższych. Wtedy mamy do czynienia z idealnym testem wielostopniowym (przynajmniej jeśli chodzi o stopień n). Oczywiście dla innych stopni współczynnik stopniowości może się różnić.

Sprawdźmy zatem, jak wskaźnik stopniowości testu funkcjonuje w praktyce. Niestety wyniki testów szkolnych, które z reguły przeprowadzane są na małych liczebnie populacjach, nie nadają się do opracowań statystycznych ze względu na zbyt duże przedziały ufności oszacowania badanych parametrów. Dlatego w dalszej analizie wykorzystamy wyniki egzaminu maturalnego z matematyki na poziomie podstawowym³ z rocznika 2017 dla ponad 20 000 uczniów zdających na obszarze działania OKE Łódź.

W kolejnych etapach analizy pomięto zawartość treściową zadań oraz umiejętności sprawdzane przez kolejne zadania. Jest to niewątpliwie ciekawe zagadnienie, ale wykraczające poza ramy niniejszego opracowania. Ponadto jeśli interesuje nas poziom wykonania, to w przypadku zadań ocenianych dychotomicznie jest to proste, ale w arkuszu egzaminacyjnym pojawiają się zadania wielopunktowe. Oczywiście uczeń, który otrzymał za rozwiązanie 0 punktów, oraz ten, który uzyskał punktację maksymalną, zostaną przypisani odpowiednio do kategorii „nie rozwiązał” lub „rozwiązał” zadanie. Problem pojawia się w przypadku uczniów z punktacją częściową. Możemy uniknąć tego problemu, wykorzystując

² W niniejszym opracowaniu również zastosowano zasadę, że o stopniu osiągnięć ucznia decyduje suma uzyskanych punktów bez względu na to, które zadania zostały rozwiązane poprawnie, a które nie.

³ Egzamin maturalny z matematyki (mimo że nie jest konstruowany z myślą o wielostopniowości) jest *de facto* testem jednostopniowym – nie zdał/ zdał.

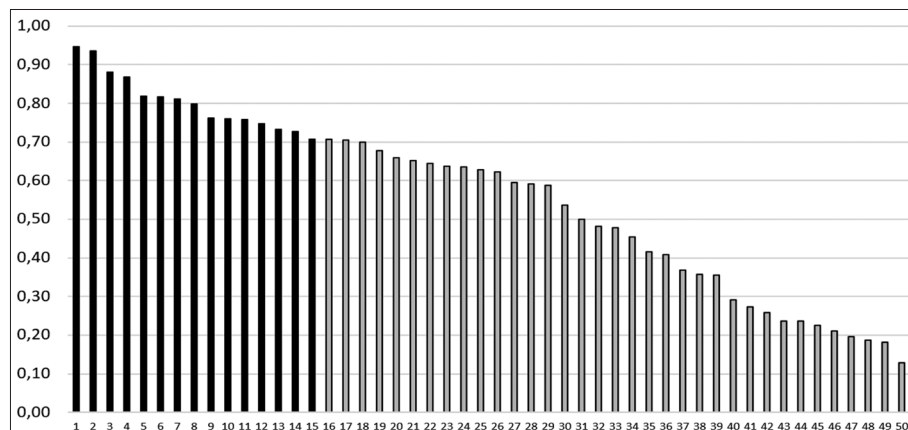
fakt, iż egzaminatorzy podczas oceny zadań otwartych stosują holistyczny system oceniania. A to oznacza, że uczeń, który otrzymał przykładowo 3 punkty za swoje rozwiązanie, musiał jednocześnie spełnić kryteria oceny zadania przewidziane zarówno za 1 punkt, jak i za 2 punkty. W związku z tym wyniki za zadania otwarte zostały przekształcone (rozbite) na poszczególne poziomy rozwiązania zgodnie ze schematem przedstawionym dla przykładowego zadania, którego pełne, poprawne rozwiązanie premiowane było przyznaniem 4 punktów (tab. 1).

Tabela 1. Rozkład wyników oceny zadania 33

Nr zad.	Poziom			
	I	II	III	IV
33	33.1	33.2	33.3	33.4
Ocena holistyczna	Ocena rozwiązania na danym poziomie			
4	1	1	1	1
3	1	1	1	0
2	1	1	0	0
1	1	0	0	0
0	0	0	0	0

W efekcie suma punktów uzyskana przez zdającego nie ulega zmianie, a cały arkusz będzie się składał z 50 zadań ocenianych dychotomicznie.

Kolejnym etapem dostosowania wyników do potrzeb analizy jest sortowanie zadań od najłatwiejszego do najtrudniejszego oraz zamiana oryginalnej numeracji na numerację od 1 do 50. Wtedy zadanie o numerze 1 będzie zadaniem najłatwiejszym, o numerze 50 – najtrudniejszym. Rozkład łatwości zadań w teście maturalnym z matematyki na poziomie podstawowym⁴ wyglądać będzie następująco (wykres 1):



Wykres 1. Rozkład łatwości zadań w teście maturalnym z matematyki na poziomie podstawowym 2017 roku

⁴ Dane z egzaminu przeprowadzonego w roku 2017.

Na wykresie 1 wyróżniono 15 najłatwiejszych zadań, czyli tych, które powinny być rozwiązane przez abiturientów, którzy zdali maturę (uzyskali wynik równy lub wyższy niż 30%). Powtórzmy tę procedurę dla wyników z matury w roku 2015 i 2016. W tabeli 2 przedstawiono współczynniki stopniowalności testu maturalnego z matematyki na poziomie S_{15} ⁵.

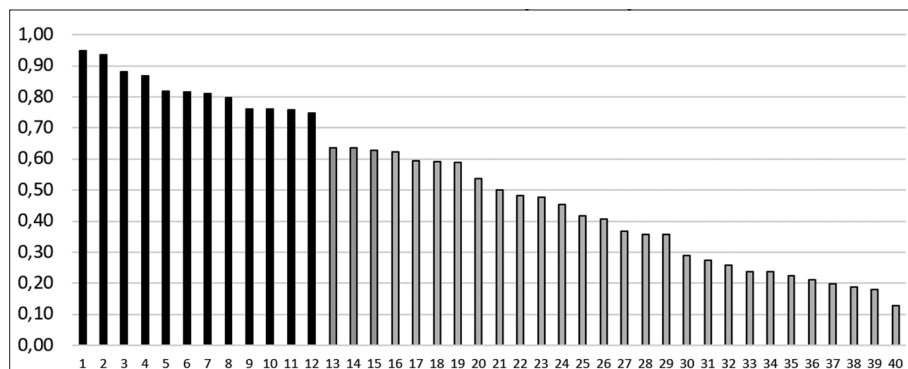
Tabela 2. Współczynniki stopniowalności testu maturalnego z matematyki

Rocznik	2015	2016	2017
Stopniowalność S_{15}	0,31	0,29	0,33

Okazuje się, że współczynnik stopniowalności testu (tab. 2) w kolejnych latach jest na podobnym poziomie. Jego wartość jest jednak daleka od naszych oczekowań w stosunku do dobrego testu wielostopniowego.

Obliczając współczynnik stopniowalności dla wszystkich możliwych stopni⁶ i zestawiając je z łatwościami zadań, daje się zauważyć pewną prawidłowość. Wzrost stopniowalności następuje wtedy, gdy między kolejnymi zadaniami (uporządkowanymi od najłatwiejszego do najtrudniejszego) występują znaczące różnice w ich łatwości. Zależność tego typu nasuwa hipotezę, że stopniowalność testu może się zwiększyć, jeżeli na granicy stopni⁷ różnica łatwości zadań będzie duża.

W celu sprawdzenia, czy hipoteza jest prawdziwa⁸, spróbujemy usunąć z testu 10 zadań, od numeru 13 do numeru 22. W ten sposób test składał się będzie z 40 pytań, a stopień zaliczenia/zdania utrzymany na poziomie 30% będzie wymagał rozwiązania co najmniej 12 zadań. Tym sposobem nie zmienimy w sposób znaczący struktury testu (w dalszym ciągu sukces egzaminacyjny polega na uzyskaniu 30% punktów). Rozkład łatwości dla tak zmodyfikowanego testu przedstawia wykres 2.



Wykres 2. Rozkład łatwości dla zmodyfikowanego testu

⁵ Indeks 15 oznacza, że 15 punktów (15 rozwiązanych zadań) jest wynikiem, który oznacza spełnienie wymagań tego poziomu i zaliczenie/zdanie matury.

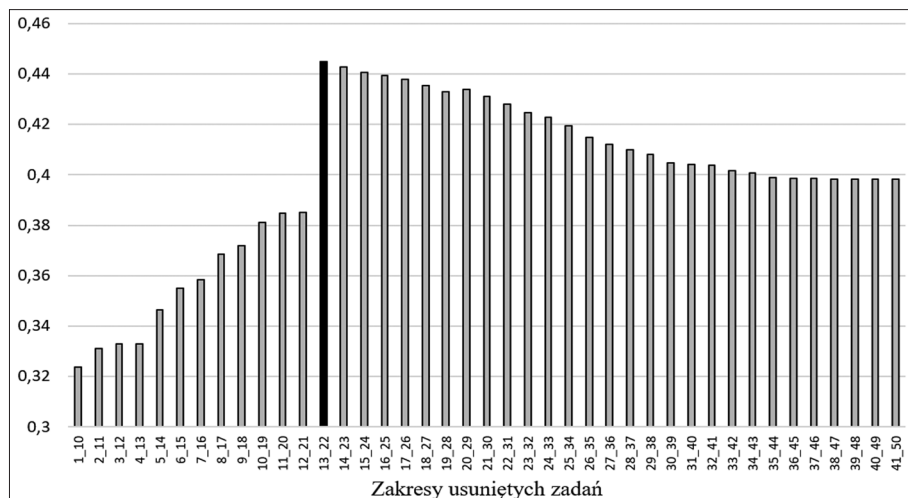
⁶ Implikuje to konieczność obliczenia 50 współczynników stopniowalności.

⁷ Przez granicę stopni rozumiemy minimalną liczbę punktów (zadań) koniecznych do osiągnięcia stopnia wyższego.

⁸ Prawdziwa nie w sensie statystycznym, ale czy w ogóle taka zależność ma miejsce.

Usunięcie zadań spowodowało wyraźną różnicę w łatwości zadania 12 i 13 wynoszącą 0,11. W efekcie nowa wartość stopniowalności wzrasta do 0,44.

Naturalnie pojawia się tutaj pytanie: czy samo skrócenie testu nie powoduje zmiany wartości stopniowalności? Sprawdźmy zatem, jak zmienia się stopniowalność testu w zależności od tego, które zadania usuniemy z testu. W tym celu obliczymy szukane wartości dla kolejnych testów, z których usuwać będziemy kolejne 10 zadań, poczynając od zadań o numerach 1 do 10, a kończąc na usunięciu zadań 41–50. Na wykresie przedstawiono wyniki proponowanych analiz.



Wykres 3. Wiarygodność testu po usunięciu 10 zadań

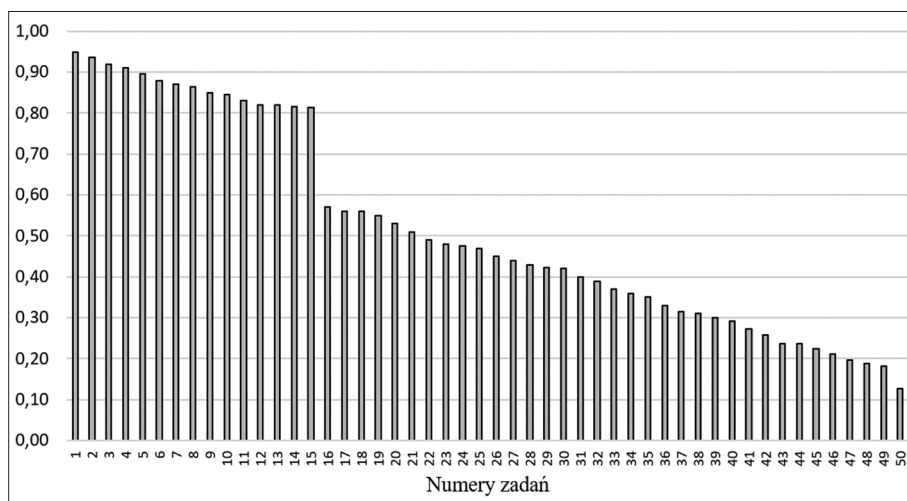
Odpowiedź na wcześniej postawione pytanie nie jest jednoznaczna. W większości wypadków wiarygodność jest większa niż w teście wyjściowym, ale najwyższa wartość i jednocześnie największy wzrost daje się zauważyć po usunięciu zadań 13–22, czyli tych, które w strukturze testu znajdują się „na granicy” stopni.

Możemy teraz pokusić się o zaproponowanie struktury testu wielostopniowego o wysokiej wartości stopniowalności⁹:

1. Łatwość zadań sprawdzających umiejętności na danym stopniu powinna być zbliżona.
2. Zadania sprawdzające umiejętności z różnych stopni powinny znacząco różnić się wskaźnikiem łatwości.

Przykładowa struktura łatwości zadań w teście jednostopniowym, o progu zaliczenia 30%, przedstawiona została na wykresie 4.

⁹ Kolejny raz należy przypomnieć, że abstrahujemy od zawartości treściowej zadań oraz od umiejętności, które są sprawdzane. Proponowana struktura testu jest tylko jednym z elementów planu testu, który należy uwzględnić, aby test spełniał wszystkie zakładane wymogi.



Wykres 4. Przykładowa struktura łatwości zadań w teście jednostopniowym

Z braku danych empirycznych nie możemy określić wartości współczynnika stopniowości dla tak skonstruowanego testu. Możemy się jedynie spodziewać, że będzie wyższy.

Pojawia się tutaj kolejna trudność, jeśli chodzi o adaptację powyższych założeń w praktyce szkolnej. Otóż test zbudowany z 50 zadań jest zbyt długi, aby mógł być zastosowany w ramach jednostki lekcyjnej. Żaden uczeń nie jest w stanie rozwiązać wszystkich pięćdziesięciu zadań w czasie 45 minut. Spróbujmy zatem zaproponować strukturę testu możliwego do zastosowania w warunkach szkolnych. Nasz test musi być testem o pięciu stopniach osiągnięć oraz powinien być złożony z 5 lub 10 zadań (po 1 lub 2 zadania na poziomie).

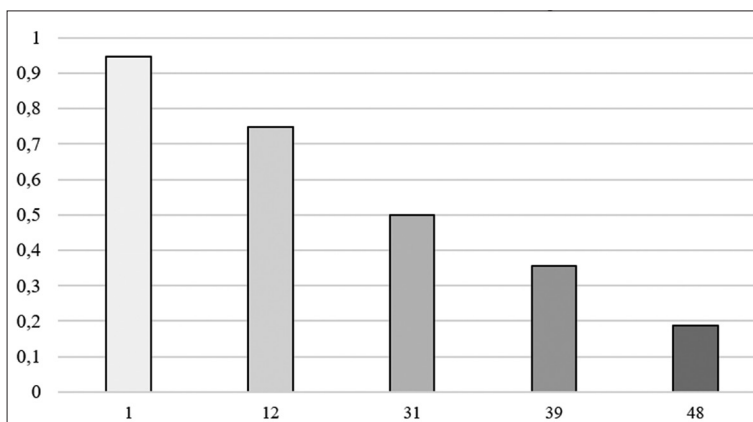
Wybermy zatem z arkusza maturalnego z matematyki 5 zadań (krótsza wersja testu), a następnie 10 (wersja dłuższa) w taki sposób, aby założenia co do łatwości zadań oraz ich struktury były spełnione. W tabeli 3 przedstawiono dane oraz zadania zgodne z powyższymi warunkami¹⁰.

Tabela 3. Przykładowe przedziały łatwości zadań i oceny szkolne

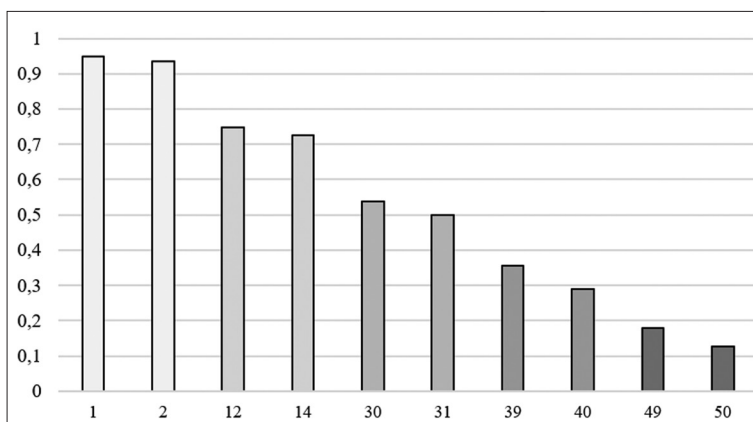
Stopień	Ocena szkolna	Łatwość zadań	Numery wybranych zadań	
			wersja krótsza	wersja dłuższa
I	dopuszczający	powyżej 0,9	1	1, 2
II	dostateczny	0,75–0,70	12	12, 14
III	dobry	0,55–0,50	31	30, 31
IV	bardzo dobry	0,35–0,30	39	39, 40
V	celujący	poniżej 0,2	48	49, 50

¹⁰ Przedziały łatwości zadań zostały dobrane arbitralnie. Tworząc test, możemy je modyfikować, pamiętając należy jednak o konieczności wyraźnego zróżnicowania łatwości pomiędzy stopniami.

Rozkłady łatwości zadań dla wybranych testów przedstawiają się następująco (wykresy 5 i 6):



Wykres 5. Struktura łatwości zadań w teście (wersja krótsza)



Wykres 6. Struktura łatwości zadań w teście (wersja dłuższa)

Różnice łatwości zadań przypisanych do różnych (kolejnych) stopni wahają się od 0,14 do 0,25 dla testu krótszego oraz od 0,11 do 0,19 dla testu dłuższego.

Tabela 4. Wiarygodność testu

Poziom	Wiarygodność testu	
	wersja krótsza (pięciozadaniowa)	wersja dłuższa (dziesięciozadaniowa)
I	0,98	0,93
II	0,92	0,78
III	0,87	0,79
IV	0,91	0,88
V	1,00	1,00

Różnice te nie wydają się zbyt duże, a jednak są istotne. Powodują one bowiem, że stopniowalności testu obliczane dla kolejnych stopni przyjmują wartości, które można przyjąć za co najmniej zadowalające (tab. 4). Należy jednocześnie zauważyć, że dla ostatniego poziomu (najwyższego) stopniowalność będzie zawsze równa 1, dlatego że dotyczy uczniów, którzy rozwiązali poprawnie wszystkie zadania z testu, w tym również zadania z poziomów niższych.

Otrzymane wyniki wskazują, że w teście pierwszym, krótszym, mamy co najmniej 87% pewności, że uczeń, który osiągnął stopień n , rozwiązał wszystkie zadania ze stopni niższych. W teście drugim taka pewność kształtuje się powyżej 78%. Widzimy tutaj wyraźnie, że konstrukcja testu szkolnego powinna opierać się na założeniu istotnej różnicy łatwości zadań przypisanych do konkretnych poziomów osiągnięć uczniów. Tak skonstruowane testy szkolne powinny dostarczyć podstaw do rzetelnego określenia, jaką ocenę szkolną przypisać uczniowi kończącemu cykl kształcenia.

Pamiętajmy jednocześnie, że dobry test charakteryzuje się nie tylko wysoką stopniowalnością, ale przede wszystkim trafnością i rzetelnością. A tego niestety w testach szkolnych bardzo często brakuje.

Bibliografia

- Ciżkowicz K., Ochenduszek J., *Pomiar sprawdzający wielostopniowy. Poradnik konstruktora i analityka*, WSP, Bydgoszcz 1986.
- Niemierko B., *Pomiar sprawdzający w dydaktyce*, Wydawnictwo Naukowe PWN, Warszawa 1990.
- Niemierko B., *Między oceną szkolną a dydaktyką*, WSiP, Warszawa 1991.
- Niemierko B., *Pomiar wyników kształcenia*, WSiP, Warszawa 1999.
- Niemierko B., *Ocenianie szkolne bez tajemnic*, WSiP, Warszawa 2005.