NEODIDAGMATA XX
POZNAŃ 1991

BOLESŁAW NIEMIERKO
Wyższa Szkoła Pedagogiczna
w Bydgoszczy

W POSZUKIWANIU TECHNOLOGII PISANIA ZADAŃ TESTÓW SPRAWDZAJĄCYCH OSIĄGNIĘCIA UCZNIÓW

ABSTRACT. Niemierko Bolesław, *W poszukiwaniu technologii pisania zadań testów sprawdzających osiągnięcia uczniów* (In the search of the technology of writing achievement tests checking pupils' results), „Neodidagmata" XX, Poznań 1991, Adam Mickiewicz University Press, pp. 27 - 52. ISBN 83-232-0302-4. ISSN 0077-653X. Received: June 1986.

In his article the author discussed the following questions:
1) description and evaluation of an attempt at constructing „technology" of writing tasks of tests of school achievements and
2) the contemporaneously changing views on the value of tasks of multiple choice.

Bolesław Niemierko, Wyższa Szkoła Pedagogiczna, 85 - 064 Bydgoszcz, ul. Chodkiewicza 30, Polska-Poland.

Celem tego artykułu jest przedstawienie współczesnego stanu badań nad zadaniami testów osiągnięć szkolnych. Obejmie on zasady konstrukcji, procedury wytwarzania, ocenę jakości i przydatność takich zadań. Sięgnie też do zadawnionego sporu o wartość zadań wyboru wielokrotnego i możliwość kontrolowania zjawiska zgadywania odpowiedzi przez uczniów.

Zgodnie z dominującą obecnie tendencją w pomiarze dydaktycznym (por. Niemierko 1981), skupimy uwagę na zadaniach testów sprawdzających (*criterion-referenced tests*), to jest testów, które reprezentują wymagania programowe określonego przedmiotu nauczania. Wymagania programowe – a nie wyniki innych badanych – tworzą układ odniesienia wyniku każdego badanego testem sprawdzającym (Niemierko 1982).

Skrótem określenia „zadania testu sprawdzającego" jest termin zadanie sprawdzające, oznaczający zadanie testowe, które mierzy opanowanie wybranego elementu treści nauczania ujętego planem testu sprawdzającego. Termin ten będzie stosowany dla podkreślenia związku czynności nauczanej z czynnością mierzoną zadaniem.

Wzrost znaczenia pojedynczego zadania testowego, jaki nastąpił w epoce pomiaru sprawdzającego w oświacie, zaowocował dążeniem do obiektywizacji procesu konstruowania zadań. W 1970 roku John Bormuth przedstawił pionierskie studium *O teorii zadań testów osiągnięć* (Bormuth 1970), a w dwanaście lat później Gale Roid i Thomas Haladyna (1982) opublikowali *Technologię pisania zadań*. Na tej drugiej pozycji oprę przegląd nowych rozwiązań w zakresie konstrukcji zadań sprawdzających.

I. TECHNOLOGIA PISANIA ZADAŃ

Techniki konstruowania i oceniania zadań, składające się na technologię pisania zadań, Roid i Haladyna uważają za „bezpośrednie przedłużenie ruchu pomiaru sprawdzającego" (1982, s. 7). To dziedzictwo powoduje uzależnienie nowej technologii od stopnia sprecyzowania dziedziny sprawdzających czynności. Precyzowanie dziedzin czynności, do którego technologia pisania zadań także się przyczynia, stanowi „najpoważniejsze wyzwanie wyłaniającej się technologii" (s. 236).

Roid i Haladyna wypowiadają się entuzjastycznie o proponowanych przez Shoemakera (1975) standaryzowanych dziedzinach zadań. Pokładają wielkie nadzieje w komputeryzacji procesów konstrukcyjnych nie tylko w zakresie budowy trzonu zadania, lecz także w doborze dystraktorów do zadań zamkniętych.

Cel swoich wysiłków wyjaśniają autorzy *Technologii*... następująco: „Metody pisania zadań, opisane i zilustrowane w kolejnych rozdziałach [książki], dają podstawę tworzenia wielkich ilości zadań przydatnych w trzech rodzajach działań: (a) testowaniu osiągnięć szkolnych; (b) ocenianiu programów nauczania i (c) badaniach naukowych. [...] Wszelka technologia zmierza do zwiększenia produktywności jakiegoś procesu przez dostarczanie narzędzi lub procedur umożliwiających redukcję operacji naturalnych. Technologia pisania zadań jest zbiorem odrębnych metod, które mogą być zastosowane przez konstruktora testu do wyprodukowania większej liczby zadań wysokiej jakości niż bez tych metod. Jak, po prostu, każde przedsiębiorstwo podnosi produktywność dzięki nowym narzędziom, tak konstrukcja testu staje się bardziej efektywna dzięki nowym narzędziom pisania zadań" (tamże, s. 5).

Na podstawie czterech środkowych rozdziałów (6-9) omawianej książki oraz odpowiednich tekstów źródłowych scharakteryzujemy następujące metody pisania zadań testowych:

1) transformacje gotowego tekstu,
2) zdania projektujące,
3) schematy i wzorce zadań,
4) rozbiór pojęć teoretycznych.

Bormutha transformacje tekstu podręcznika

Dla Johna Bormutha (1970) universum zadań testowych ma fundament w tekście zawierającym sprawdzaną wiedzę, a więc w typowym podręczniku uczniowskim. Ponieważ tekst taki jest zbudowany ze zdań, technologia pisania zadań polega na przekształcaniu zdań orzekających w pytania.

Bormuth nie szczędzi krytki dotychczasowym koncepcjom i procedurom konstrukcji zadań testowych, stwierdzając, że „są one zdefiniowane wyłącznie subiektywnym życiem prywatnym autora testu, co czyni testowanie osiągnięć czymś niewiele lepszym od czarnej magii" (s. 2). Wynik tej sztuki „zależy niemal całkowicie od introspekcyjnego wglądu, wytrwałości, pomysłowości i finezji literackiej autora testu" (s. 9).

Zadania wyprodukowane przez transformacje tekstu podręcznika uważa Bormuth za zdefiniowane operacyjnie, co znaczy, że zbiór operacji prowadzących do ich wytworzenia jest zobiektywizowany i dostępny kontroli „publicznej".

Istnieją, zdaniem Bormutha, cztery główne „kontrasty" między operacyjnymi i tradycyjnymi metodami uzyskiwania zadań testowych (s. 10-13).

1. Autor zadania nie ma wpływu na dobór wyrażeń. Dwaj osobno pracujący autorzy stosujący te same operacje do tego samego tekstu powinni otrzymać identyczne zadania. Co więcej, proces wytwarzania zadań powinien być całkowicie zautomatyzowany, nie tyle dla usprawnienia, ile dla zapewnienia jednoznaczności operacji.

2. Mierzona czynność jest dokładnie określona. Etykietki nadawane zadaniom wytwarzanym tradycyjnymi metodami bywają mylące. Na przykład zadanie „$9 \cdot 12 = ?$" może wymagać tylko przypomnienia, obliczenia lub też posłużenia się prawem rozdzielności mnożenia względem dodawania (odejmowania), zależnie od treści nauczania, z którą zadanie jest związane. Wyprowadzenie zadania z tekstu podręcznika daje możliwość natychmiastowego sprawdzenia, o jaką czynność chodzi.

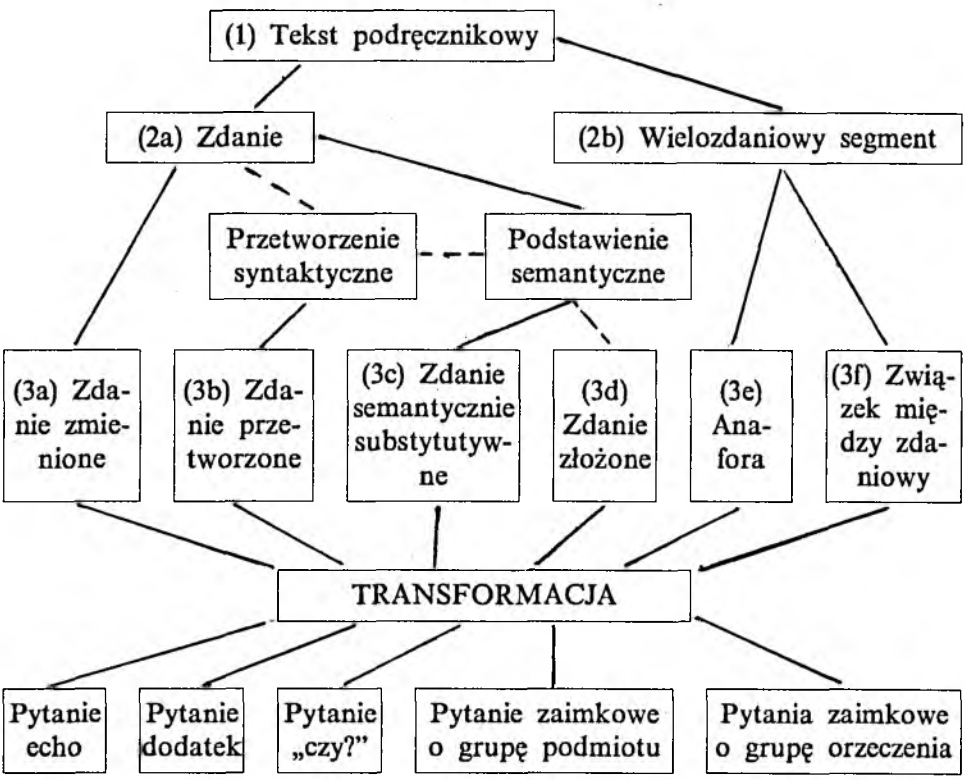
3. Autor testu nie decyduje o tym, czy napisać określone zadanie. Tradycyjny autor odrzuca zadania, które wydają mu się „banalne, zbyt złożone, zbyt proste, przegadane lub z innego względu niepożądane". W podejściu operacyjnym decyzja o napisaniu zadania jest podejmowana z góry: w toku planowania testu i przez (warstwowe) losowanie elementów dziedziny czynności. Racjonalność tej decyzji podlega kontroli zewnętrznej.

4. Zadania testu są jednoznacznie powiązane z nauczaniem. Przedtem autor zadań i powołani eksperci oceniali ten związek intuicyjnie, a więc mało dokładnie. Metody operacyjnie pozwalają na uzyskanie logicznej stosowności (*logical relevance*) zadania wobec nauczania, polegają na tym, że „zarówno zadanie, jak i prawidłowa odpowiedź, mogą być wyprowadzone z określonego fragmentu nauczania przez zastosowanie zbioru operacji, które

mogą być (a) uogólnione na różne rodzaje nauczania i (b) obiektywnie opisane" (s. 34).

Bormuth przyznaje, że używa terminu „transformacje" w nieco innym znaczeniu niż lingwistyka, zwłaszcza generatywna gramatyka transformacyjna Noama Chomsky'ego, na której się opiera. Transformacji na zadanie (*item transformation*) podlega segment tekstu – zwykle prozaicznego, ale także przedstawionego w innej symbolice, np. matematycznej – poprzez zaproponowane w „teorii zadań testów osiągnięć" operacje.

Materiał i wyniki transformacji na zadania, rozważane przez Bormutha, podaję w postaci schematu według Helmuta Rupprechta (1972), który napisał zwięzłe studium krytyczne o teorii Bormutha.



Rys. 1. Formy zadań według Bormutha i ich wyprowadzenie z tekstu podręcznika

Najbardziej charakterystyczne dla metody Bormutha jest transformowanie zdań wyjętych z tekstu (1) jako „zdania bazowe" dla zadań, w nie zmienionej formie (3a). By dotrzeć do pełnego sensu zdań podręcznikowych koniecznie bywają jednak „parafrazy" syntaktyczne, na przykład polegające na zbudowa-

niu kilku zdań prostych ze zdania rozwiniętego, oraz podstawienia semantyczne, sprowadzające się do operowania synonimami. Pierwszy z tych zabiegów prowadzi do zdań przetworzonych (3b), drugi prowadzi do zdań semantycznie substytutywnych (3b), a obydwa zabiegi zastosowane łącznie prowadzą do zadań, które Bormuth nazywa „złożonymi" (*compound*).

Bormuth zdaje sobie sprawę, że pojedyncze zdania nie wyczerpują informacji zawartej w tekście. Jego procedury nie sięgają jednak dalej niż „anafory" (3e), to jest zdanie zbudowane przez dołączenie wyrażeń użytych w tekście wcześniejszym lub późniejszym, sygnalizowanych zwykle przez zaimki osobowe („on"), zaimki wskazujące („ten") i przysłówki zaimkowe („tam") w zdaniu bazowym. Związki międzyzdaniowe (3f) wymienia tylko dla „wartości heurystycznej", gdyż „bardzo mało wiemy o tym, jak formalne właściwości języka korelują z tymi związkami i czy te związki są w ogóle jakoś strukturalnie sygnalizowane" (s. 53). Możliwymi sygnałami mogłyby być: pozycje zdań w tekście, porządek czasowy, informacje o przyczynowości i podporządkowaniu.

Transformacje Bormutha prowadzą do następujących form zadań testowych:

1. Pytanie – echo (*echo item*), stanowiące pierwszy sposób weryfikacji zdania. Przykład (Bormutha): „Te jabłka były zielone?".

2. Pytanie – dodatek (*tag item*), stanowiące drugi sposób weryfikacji zdania (mało naturalny dla współczesnej polszczyzny). Przykład: „Chłopiec pracuje, nieprawdaż?".

3. Pytanie „czy?" (*yes/no item*), stanowiące trzeci sposób weryfikacji zdania. Przykład: „Czy goście zjedli?".

4. Pytanie zaimkowe (*wh-item*) w dwu głównych odmianach:
   (a) o grupę podmiotu i
   (b) o grupę orzeczenia.

Na przykład zdanie bazowe „Chłopiec odbił piłkę" może być transformowane m.in. następująco: (a) „Kto odbił piłkę?"; (b) „Co odbił chłopiec?"

Trzy początkowe formy zadań, sprawdzające prawdziwość informacji, odpowiadają dotychczasowym zadaniom typu „prawda – fałsz" (Niemierko 1975a, s. 39-41). Pytania zaimkowe tworzą zadanie krótkiej odpowiedzi (tamże, s. 33-34), a po zaopatrzeniu ich w kilka odpowiedzi – zadanie wyboru wielokrotnego (tenże, s. 37-39). To ostatnie przekształcenie Bormuth pochwala (s. 43), ale nie podaje żadnych operacyjnie zdefiniowanych reguł dobierania dystraktorów (odpowiedzi nieprawidłowych).

Oceniając metodę Bormutha, Roid i Haladyna (1982) zauważają, że posługiwanie się podręcznikami i innymi materiałami pisemnymi jest w szkole częste. Do transformacji na zadania testowe nadaje się jednak tylko segment tekstu (a) bardzo ważny i (b) zwięźle napisany. W typowym tekście dydaktycznym jest wiele zdań zawierających wtręty, powtórzenia, nazbyt szczegółowe

objaśnienia i dość dowolne przykłady, a „każde zdanie, choćby nie wiadomo jak banalne i źle napisane może być przekształcone na pytanie" (s. 93).

Jak wyliczono (Diederich 1970), pełne zastosowanie metody Bormutha do transformacji podręcznika fizyki złożonego z 16 tysięcy zdań doprowadziłoby do zbudowania 960 tysięcy zadań testowych, a więc do ilości zdecydowanie nieekonomicznej bez względu na jakość tych zadań.

Roid i Haladyna proponują ulepszenie metody Bormutha, polegające na zastąpieniu losowego doboru zdań do transformacji na zadania testowe doborem celowym, dokonywanym według:

a) oceny ekspertów;

b) listy częstotliwości słów języka angielskiego (wybiera się zdania zawierające rzeczowniki i przymiotniki najrzadsze) lub

c) dwu powyższych zasad łącznie.

Procedura ulepszonej metody Bormutha obejmuje cztery kolejne kroki (Roid i Haladyna 1982, s. 98-108):

1) odsianie z tekstu zdań dydaktycznie nieistotnych;

2) wybór dydaktycznie istotnych zdań kluczowych. Wymaga to klasyfikacji zdań według ekspertów lub list częstotliwości, a niekiedy zredagowania (streszczenia) tekstu przez ekspertów;

3) transformowanie zdań na pytanie o „główny rzeczownik" (*keyword noun*) zdania bazowego z zachowaniem reszty słownictwa tego zadania;

4) konstrukcję dystraktorów do zadań wyboru wielokrotnego.

Autorzy *Technologii...* zauważają, że swobodny dobór dystraktorów przez autorów zadań prowadzi do dużych różnic trudności zadań. Automatyczny dobór dystraktorów wyrównuje te różnice w dół, ułatwiając zadanie. Zalecana jest następująca metoda doboru dystraktorów, pozwalająca zachować niektóre zalety pracy autorskiej:

1) sporządzenie listy głównych rzeczowników wybranego tekstu;

2) poklasyfikowanie tych rzeczowników na siedem kategorii (Frederiksen 1975):

I. ożywione („owad", „John"),
II. symboliczne–dynamiczne („film", „gra"),
III. symboliczne–statyczne („obraz", „list"),
IV. rzeczowe–dynamiczne („wiatr", „hałas"),
V. rzeczowe–statyczne („skała", „dom"),
VI. abstrakcyjne–dynamiczne („miłość", „nadzieja"),
VII. abstrakcyjne–statyczne („długość", „wielkość");

3) wylosowanie odpowiedniej liczby rzeczowników z tej kategorii, do której należy prawidłowa odpowiedź na zadanie. Rzeczowniki te będą dostatecznie atrakcyjnymi dystraktorami (Roid i Haladyna 1982, s. 106).

Zarzutem, jaki można postawić proponowanym ulepszeniom metody Bormutha, jest utrata pełnej operacyjności (przez wprowadzenie ekspertów) dla stosunkowo niewielkich zysków, polegających na wyeliminowaniu (a) zadań marginalnych treściowo i zbyt łatwych oraz (b) dystraktorów logicznie niespójnych z trzonem zadania.

Metoda Bormutha wyrosła ze statystycznych badań psychologicznych nad pamięcią i zrozumieniem słów (Ebbinghaus 1987). Jej wcześniejszą, i w pewnym stopniu równoległą, postacią są testy zamknięte (*close tests*), w których opuszczone słowa – zwykle dokładnie co piąte, z pominięciem pierwszego i ostatniego akapitu danego testu – odgaduje się wyłącznie z kontekstu (Taylor 1953). Testy takie uważa się za narzędzie pomiaru rozumienia czytanego tekstu przez uczniów oraz „czytelności" (*readability*) różnych tekstów. Są zwykle znacznie trudniejsze (Bormuth 1968) i mniej rzetelne od klasycznych testów tego rodzaju. Ulepszoną wersją testów zamkniętych są testy labiryntowe (MAZE tests), w których uczeń wybiera jedno z kilku podanych słów wypełniających lukę (por. „wybór wielokrotny w tekście", Niemierko 1975b, s. 190-191). Testy te są zalecane przez specjalistów amerykańskich (Nitko 1983, s. 239; Roid i Haladyna 1982, s. 239), zwłaszcza gdy losowane są tylko słowa najbardziej znaczące (użyte w funkcji rzeczownikowej).

Żadna z metod badania znajomości testów nie nadaje się do pomiaru osiągnięć wybiegających poza elementarne rozumienie tych testów. Transformacje i uzupełnienie tekstu podręcznikowego sprowadzają cele nauczania do odtwarzania materiału nauczania (Rupprecht 1972, s. 113). Praktyka zamiany wybranych zdań podręcznika na „obiektywne" zadania typu „prawda – fałsz" i zadania „z luką" znana jest i potępiana – jako skrajny materializm pomiarowy – od dawna (por. Niemierko 1975a, s. 35 i 40). Z tego względu trudno byłoby kogokolwiek w Polsce nakłaniać do budowania takich zadań.

Jakie jest więc znaczenie gramatycznej metody zdaniowej Bormutha i pokrewnych metod transformacji tekstu podręcznikowego na zadania testowe? Bezpośrednio, niewielkie. Tylko omawianie czytanek w najniższych klasach szkoły podstawowej wymaga podobnej szczegółowości odtwarzania tekstu. Potem uwaga czytelnika przenosi się na układy wielozdaniowe. Tymczasem, jak zauważył sam Bormuth (1970, s. 55), „w miarę jak pytania są formułowane na coraz wyższych piętrach logicznej struktury treści, odpowiedzi stają się coraz dłuższe. Na wysokich piętrach nie sposób odróżnić tych bardzo długich odpowiedzi od samodzielnych rozpraw (*essay responses*). A specjaliści zwykli przykładać większą wagę do wiedzy sprawdzanej pytaniami z wyższych pięter struktury logicznej. Często nazywają te pytania docieraniem do myśli przewodniej danego tekstu lub podobnie".

Ponieważ pogodzenie idei testowania osiągnięć wysokich kategorii celów nauczania z automatyzacją produkcji zadań testowych jest na razie niewyob-

rażalne, „teoria zadań testów osiągnięć" Johna Bormutha, mimo ciepłego przyjęcia przez specjalistów pomiaru sprawdzającego, pozostaje ważnym, ale negatywym doświadczeniem technologii pisania zadań.

Guttmana zdanie projektujące

Louis Guttman (1969) zaproponował technologię pisania zadań testowych akcentującą, przeciwnie niż teoria Bormutha, najogólniejszy sens wybranego zakresu treści nauczania. Sens ten wyrażają zdania projektujące (*mapping sentences*), wiążące główne elementy treści ujęte w klasy zwane „aspektami" (*facets*). Zbudowanie odpowiedniej liczby zdań projektujących jest równoznaczne ze sprecyzowaniem dziedziny nauczanych czynności, a zarazem z operacyjnym zdefiniowaniem zadań stosowanego testu sprawdzającego.

Wkrótce stwierdzono (Berk 1978), że typowe poprawnie sformułowane cele nauczania mogą być wykorzystane jako zdania projektujące. Aby tak było, nauczana czynność powinna być wieloaspektowa, to jest modyfikowana sytuacją, materiałem i ewentualnie innymi warunkami jej wykonywania.

Prosty przykład zdania projektującego zaczerpnę z polskiego programu nauczania matematyki w klasie VII szkoły podstawowej (*Program* ..., 1984, s. 25; dodano numerację aspektów): „Uczniowie rozpoczynający naukę w klasie VII powinni umieć (1a) sprawnie wykonywać (2a-d) rachunki w zakresie liczb wymiernych (3a) całkowitych i (3b) ułamkowych, (4a) nieujemnych i (4b) ujemnych oraz (1b) stosować te umiejętności [...] w obliczaniu wartości nieskomplikowanych wyrażeń algebraicznych".

W powyższym przykładzie wyróżniono następujące aspekty umiejętności matematycznych niezbędnych do podjęcia nauki w klasie VII:

1) rodzaj wielkości, na których wykonywane są działania: liczbowe w arytmetyce, literowe i liczbowe w algebrze;

2) rodzaj działań: tu przyjęto, iż chodzi o cztery podstawowe działania (dodawanie, odejmowanie, mnożenie i dzielenie) z osobna, ale szersza klasyfikacja działań matematycznych i ich połączeń (na poziomie progu klasy VII) wymagałaby osobnego zdania projektującego;

3-4) zbiory liczb: całkowite nieujemne (naturalne), całkowite ujemne, ułamkowe nieujemne, ułamkowe ujemne, przy czym współwystępowanie liczb dodatnich i ujemnych w jednym działaniu nie jest tu odróżnione.

Zacytowane zdanie projektuje nam $2 \cdot 4 \cdot 4 = 32$ zbiory zadań sprawdzających względnie jednorodnych co do mierzonej umiejętności matematycznej.

Entuzjaści metody zdań projektujących (Engel i Martuze 1976) podkreślają logiczny związek treści zadań produkowanych tą metodą z treścią nauczania oraz możliwość dobierania, także komputerowo, dystraktorów dostatecznie bliskich prawidłowej odpowiedzi. Roid i Haladyna (1982, s. 132 i n.) są ostrożniejsi. Zauważają, że:

1) łatwiej o wyraziste, uzgodnione między specjalistami, struktury treści w matematyce i naukach przyrodniczych niż w naukach humanistycznych i społecznych;

2) budowanie zdań projektujących i katalogowanie elementów w poszczególnych aspektach może pochłaniać niezwykle wiele wysiłku;

3) brak jeszcze systematycznych badań nad tą metodą.

Zaletą metody Guttmana jest orientacja na ogólne cele nauczania, a nie – jak w przypadku transformacji zdań podręcznikowych – na szczegółowy materiał nauczania. Nieuchronnym kosztem tej zmiany podejścia jest nie w pełni operacyjne zdefiniowanie zadań testu, gdyż „różne osoby mogą mieć różne koncepcje tej samej dziedziny, co doprowadzi do różnych zdań projektujących, aspektów i elementów" (Roid i Haladyna 1982, s. 143).

Jak stwierdzają Roid i Haladyna (tamże, s. 143), zdania projektujące „wymagają dalszych badań, które awansują tę technologię pisania zadań testowych z interesującego pomysłu na praktyczną metodologię".

Hively'ego schematy zadań

Schematy zadań narodziły się z doświadczeń w konstruowaniu zadań równoległych, to jest zadań testowych różniących się tylko szczególnym materialem (danymi liczbowymi, jednostkowymi faktami), oraz w zastosowaniu komputerów do przechowywania i wytwarzania takich zadań.

Według H. G. Osbourna (1968, s. 97), schemat zadań ma następujące właściwości:

„1. generuje zadania o stałej strukturze syntetycznej,

2. obejmuje jeden lub więcej elementów zmiennych,

3. definiuje klasę zadań testowych przez wyszczególnienie zbiorów podstawień (*replacement sets*) jako elementów zmiennych".

Największe zasługi w rozwijaniu schematów zadań położył Wells Hively (1973), którego prace eksperymentalne zyskały powszechne uznanie. Hively'ego schematy zadań były bardzo szczegółowe, gdyż obejmowały dziewięć następujących pozycji:

I. Opis ogólny: 1-3 zdanie objaśniające (a) sytuację, w której uczeń zostanie postawiony i (b) oczekiwaną od niego czynność.

II. Właściwości bodźca i reakcji: szczegółowa charakterystyka sytuacji i czynności ucznia obejmująca właściwości:

a) wspólne dla wszystkich podklas (*cells*) zadania, to jest dla wszystkich odmian nie naruszających jego schematu;

b) odróżniające te podklasy między sobą;

c) zmienne wewnątrz podklas.

Hively zauważył, że najtrudniejsze jest zwykle określenie wspólnych (stałych) właściwości dla schematu zadania, zwłaszcza we wczesnych etapach analiz i w szerszych dziedzinach sprawdzanych czynności. Im szersza jest dziedzina, tym więcej właściwości bodźca i reakcji musimy ustalić, aby odróżnić schematy zadań między sobą (tamże, s. 30).

III. Macierz podklas zadania: zestawienie podklas zadania w postaci tabeli o kilku lub kilkunastu polach. Podklasy mogą być zróżnicowane typami przedmiotów (rekwizytów), wielkościami i stosunkami liczbowymi, materiałem językowym i podobnymi czynnikami sytuacyjnymi.

IV. Szkielet zadania (*item form shell*): wyszczególnienie stałych składników zadania: wyposażenia, wskazówek dla prowadzącego testowanie, tekstu zadania dla ucznia.

V. Opis wyposażenia: szczegółowy opis urządzeń i materiałów, które mają być udostępnione uczniowi oraz karty odpowiedzi (pozycji na karcie), które ma wypełnić.

VI. Schemat podstawień: wyszczególnienie sposobów dobierania (np. w pary) podstawionych elementów, to jest przedmotów, nazw, liczb itp. dla utworzenia kolejnych podklas zadania.

VII. Zbiory podstawionych elementów: wykaz przedmiotów, obrazów, nazw, liczb itp. służących jako elementy wymienne w zadaniu.

VIII. Rejestrowanie odpowiedzi: sposób protokółowania (a) zachowania ucznia w toku rozwiązywania zadania oraz (b) udzielonej odpowiedzi. Często obejmuje sporządzanie szkicu i zapisy symboliczne potrzebne do pogłębionych analiz czynności uczniów.

IX. Zasady punktowania: wykaz niezbędnych cech prawidłowej odpowiedzi, wywiedziony z próbnych zastosowań testu. „Ogólnie biorąc, zespół badawczy zyskał przekonanie, że najdogodniej jest ustalać zasady punktowania empirycznie, w toku kolejnych zastosowań testu, a nie próbować tworzyć zasad z powietrza przed zgromadzeniem danych pilotażowych" – napisał Hively (s. 33).

Wszystkie zadania Hively'ego były otwarte (krótkiej odpowiedzi), a testowanie odbywało się w zasadzie indywidualnie ze skrupulatnością eksperymentu naukowego. Dzięki precyzyjnie określonym regułom podstawień zadania wyprodukowane według jednego schematu okazywały się na ogół zadowalająco jednorodne (Macready i Mervin 1973), zwłaszcza w obrębie podklasy zadania.

Roid i Haladyna (1982, s. 121) dostrzegają następujące zalety schematów zadań jako technologii pisania zadań:

1) skracają czas budowania testu (po wstępnym zainwestowaniu czasu w wytworzenie schematu);

2) wiernie definiują dziedziny sprawdzanych czynności;

3) mogą korzystać z mocy i szybkości komputerów (w zakresach wymagających liczenia, jak statystyka lub księgowość);

4) mogą być zaprogramowane na komputery, które będą konstruować i drukować losowe wersje zadań równoległych.

Schematy zadań znalazły pewne zastosowanie w naukach ścisłych oraz w nauczaniu wspomaganym komputerowo, m.in. w szkołach wojskowych i innych wyższych uczelniach zawodowych. Najłatwiej je stosować, gdy podstawianymi elementami są liczby; bywają użyteczne przy „mnogości terminów technicznych, które uczeń czasami myli" (tamże, s. 122). Inne źródła elementów wymienionych to „poprawne i niepoprawne przykłady dla pewnych pojęć i zasad" (s. 123), a także wybrane układy fizyczne, np. obwód elektryczny z podłączonymi przyrządami.

Brak przykładów udanego zastosowania metody Hively'ego w przedmiotach humanistycznych i „artystycznych". Co gorsza, okazała się ona uciążliwa dla większości konstruktorów testów nawet w matematyce i fizyce. Jak pisze James Popham (1975, s. 136): „poziom sprecyzowania dziedziny czynności jest tak wysoki, że, wyjąwszy zastosowanie jednego lub dwu schematów do przećwiczenia nowej techniki generowania zadań, niewielu konstruktorów ma cierpliwość do pracy z tymi hiperszczegółowymi opisami". Tymczasem „wyszczególnienia dziedzin, które nie są intensywnie wykorzystywane przez decydentów oświatowych są nieużyteczne. A mało jest cech tak niezbędnych jak zwięzłość do skłonienia zapracowanych pedagogów do spożytkowania wyników badań, opisów dziedzin lub po prostu czegokolwiek dostępnego na piśmie" (tamże, s. 138).

Popham zaproponował modyfikację schematów zadań w kierunku rozluźnienia ich struktury i uproszczenia zapisu. Nazwał to „strategią ograniczonej ostrości" (1978, s. 117). Jego wyszczególnienia testu (*test specifications*) obejmują pięć następujących pozycji:

I. Opis ogólny: kilkuzdaniowy opis czynności, której opanowanie było celem nauczania. W opisie tym, podobnie jak w schematach Hively'ego, wyróżnia się bodziec (sytuację) i reakcję ucznia.

II. Przykładowe zadania: zadanie wyboru wielokrotnego (rzadziej – w innej formie) wraz z niezbędnym fragmentem ogólnej instrukcji testowania (o sposobie rozwiązywania zadań i przedstawiania odpowiedzi). Popham dopuszcza ewentualność, iż „najbardziej zajęci" konstruktorzy testów mogą chcieć wykorzystać tylko opis ogólny i przykładowe zadanie, by przystąpić jak najszybciej do własnych działań. Zaleca, by nie oznaczać prawidłowej odpowiedzi, co może wciągnąć użytkownika wyszczególnień do dalszego czytania i zapobiega zbyt pośpiesznej dyskwalifikacji zadania, opartej na powierzchownym zrozumieniu jego konstrukcji (s. 124).

III. Właściwości bodźca: opis materiału (językowego, liczbowego, faktograficznego), jaki może być wykorzystany w zadaniach, oraz „absolutnie niezbędne" wskazówki co do ich budowy. Przedstawiając właściwości bodźca należy kierować się wyczuciem potrzeb użytkowania, zdrowym rozsądkiem

VIII. Uwagi

Umiejętność może być sprawdzana pisemnie zbiorowo.

Wzorce umiejętności z innych dziedzin niż arytmetyka nie mogły być tak zwięzłe. Już w zakresie matematycznych umiejętności praktycznych w nauczaniu początkowym pojawia się konieczność osobnego scharakteryzowania czynności naturalnych, jak „rozpoznawanie w otoczeniu odcinków prostopadłych i równoległych", „posługiwanie się monetami i banknotami od 1 złotego do 100-złotowych", „mierzenie pojemności w litrach", „odczytywanie temperatur dodatnich na termometrze", od podobnych czynności symulowanych w zadaniach pisemnych, jedynie możliwych do zastosowania w testowaniu zbiorowym (Siterska 1987).

W chwili pisania tego artykułu brak nam jeszcze rodzimych doświadczeń w stosowaniu wzorców umiejętności przez inne osoby niż ich autorzy.

Produkowanie zadań dotyczących pojęć

Częstym zarzutem wobec technologii pisania zadań testowych jest ograniczenie jej do zapamiętanych wiadomości i wyćwiczonych umiejętności. Ten sam zarzut bywa wysuwany wobec wszelkich zadań testów osiągnięć szkolnych.

Zagadnienia testowania rozumienia treści nauczania podjął Richard Anderson (1972). Przeglądając liczne testy osiągnięć stwierdził on, że prawidłowe odpowiedzi na zadania otwarte zawierają dziesięciokrotnie częściej słowa występujące w oryginalnym tekście (podręcznika) niż odpowiednie synonimy. Nie daje to pewności, czy badani rozumieli podawaną informację. „Zatem – konkluduje Anderson (s. 163) – najlepszym posunięciem jest umieszczenie parafrazy tekstu w trzonie zadania, a nie oczekiwanie jej w odpowiedzi ucznia". Jako parafraza określa Anderson zdanie równoważne merytorycznie zdaniu oryginalnemu (podręcznikowemu), ale nie mające żadnych słów o znaczeniu rzeczownikowym (wyrazów samodzielnych) wspólnych z nim.

Anderson zaproponował następującą procedurę tworzenia parafraz zasad i praw naukowych, których zrozumienie przez ucznia jest niezbędne:

1. Zamień każdy termin ogólny w podręcznikowym sformułowaniu zasady lub prawa na odpowiednią nazwę jednostkową.
2. Podstaw synonimy w miejsce pozostałych słów o znaczeniu rzeczownikowym.
3. Sprawdź, czy uzyskany tekst nie ma wspólnych wyrażeń z którymkolwiek ze zdań (przykładów) pełniących w podręczniku rolę objaśnień danej zasady lub prawa naukowego.

Większość zasad i praw naukowych da się łatwo przedstawić w postaci zdań warunkowych (w formie „jeżeli – to"). Stwarza to możliwość sprawdzania

III. Wymagania programowe

Matematyka. Nauczanie początkowe. Osiągnięcie konieczne – „arytmetyka".

IV. Sytuacja sprawdzania

a. Tekst pisemny

Zapis działania w wierszu.

b. Wyposażenie specjalne

(niepotrzebne)

c. Instrukcja

„Wykonaj mnożenie pisemne".

V. Przebieg sprawdzania

a. Obserwacja czynności ucznia

Analiza zapisu dokonanego przez ucznia:

1) prawidłowość zapisu,
2) poprawność obliczenia

b. Ocena czynności ucznia

Czynność jest opanowana, gdy obliczenie jest poprawne i zapis działania jest prawidłowy. Nie są brane pod uwagę:

- zewnętrzna staranność zapisu,
- kształtność cyfr,
- błędne umieszczenie (lub brak) znaku „x"
- (inne) drobne uchybienia.

c. Zapis oceny

Podwójny zapis:

1) wykonanie czynności liczbą „1", niewykonanie (błędne wykonanie) liczbą „0",
2) rodzaj ewentualnego błędu, np.:
   „Błędne przepisanie liczb",
   „Błędne podpisanie liczb",
   „Wykonanie innego działania",
   „Błąd w zakresie tabliczki mnożenia",
   „Błąd w dodawaniu",
   „Błąd w przekraczaniu progu dziesiątek (setek, tysięcy)".

VI. Przykład zadania sprawdzającego

Wykonaj mnożenie pisemne:

$253 \cdot 3$

VII. Przykładowe elementy wymienne zadania sprawdzającego

116;2	115;6	126;4	115;5	121;7
460;2	224;4	307;3	401;3	200;8

i wewnętrzną dyscypliną, gdyż „niechlujne myślenie autora wyszczególnień zaowocuje bezsensownymi wskazówkami" (s. 124). Sporo wyjaśnia zamieszczone wcześniej przykładowe zadanie.

IV. Właściwości reakcji: objaśnienie sposobu udzielania odpowiedzi na zadanie (w danej formie) oraz konstrukcji (ewentualnych) dystraktorów i ich uporządkowanie w zadaniu. Najważniejsze jest tu dostatecznie precyzyjne odróżnienie odpowiedzi prawidłowej od odpowiedzi nieprawidłowych (niepełnych, błędnych, nieadekwatnych), które mogą być dystraktorami.

V. Uzupełnienie: miejsce na szczegółowe listy wymiennych elementów treści zadań i dokładniejszą informację o materiale nauczania, którego zadanie dotyczy.

Pophem nie pragnął stworzyć technologii pisania zadań i nieustannie podkreślał znaczenie inteligencji konstruktora testu w posługiwaniu się wyszczególnieniami. Dążył do takiego opisu dziedziny czynności, by „niezależni wykazywali wysoką zgodność w rozpoznawaniu, czy poszczególne zadania testowe rzeczywiście mierzą czynność opisaną w danej dziedzinie" (1975, s. 138).

Roid i Haladyna nie wprowadzili metody Pophama do swego podręcznika technologii pisania zadań. Słuszniejsze byłyby ją uznać za metodę semitechnologiczną (półtechnologiczną), obejmując tą nazwą podejścia skutecznie porządkujące procedury konstrukcyjne zadań bez prób zautomatyzowania tych procedur. Wiązki zadań równoległych wytwarzane metodami semitechnologicznymi cechują się rzetelnością „zadziwiająco wysoką" (Ebel 1979, s. 282), co stanowi argument na rzecz stosowania tych metod.

Jedną z zalet Pophama wyszczególnień testu jest ich uniwersalność. Dobrze nadają się do przedmiotów humanistycznych, a także do dziedziny motywacyjnej, na co autor metody przedstawia przekonywające dowody (1978, rozdział 9) w postaci opisów takich szczegółowych dziedzin, jak „preferencje muzyczne", „ocenianie ludzi jako indywidualności" i „przestrzeganie zasad bezpieczeństwa".

Pierwszą próbą zastosowania podejścia semitechnologicznego do konstrukcji zadań testowych w Polsce jest sporządzony przeze mnie wzorzec umiejętności, którego budowę przedstawię wraz z przykładem dostarczonym przez Władysławę Siterską (1987):

I. Nazwa

Mnożenie liczb trzycyfrowych przez jednocyfrowe sposobem pisemnym.

II. Opis ogólny

Uczeń otrzymuje zapis działania w wierszu. Działanie polega na pomnożeniu liczby całkowitej trzycyfrowej przez liczbę jednocyfrową z jednokrotnym przekroczeniem progu dziesiątek, setek lub tysięcy. Uczeń zapisuje podane liczby jedna pod drugą i wpisuje pod kreską wynik mnożenia.

rozumienia tych zasad i praw przez ich zastosowanie w następujący sposób (tamże, s. 153):

A. Do przypadku mieszczącego się w poprzedniku zdania warunkowego uczeń konstruuje (wybiera) następnik.

B. Do przypadku mieszczącego się w następniku zdania warunkowego uczeń konstruuje (wybiera) poprzednik.

Propozycje Andersona znalazły uznanie specjalistów pomiaru sprawdzającego (Roid i Haladyna 1982, s. 91), mimo iż parafrazowanie tekstów i dobieranie przykładów odbiega dość daleko od rygorów technologicznych. Sam autor tych pomysłów uważał, że wiążą one pomiar dydaktyczny z nauczaniem, a niedostatek takiego związku we wcześniejszych sprawozdaniach z badań testowych krytykował niezwykle ostro. „Procedury obecnie stosowane do konstrukcji i opisu testów osiągnięć – napisał (1972, s. 168) – stanowią śmietnik (*a mess*). Wnioski o metodach, czynnikach i procedurach z trudem mogą być brane poważnie, gdy nie wiemy co test mierzy. Drastyczne działania muszą być podjęte". Istotnie, następne lata przyniosły przewartościowanie stanowisk wielu badaczy pedagogicznych w kierunku sprawdzania osiągnięć przewidzianych programami nauczania i wyższej rangi analizy treści nauczania.

Systematyczne podejście do testowania znajomości pojęć przedstawili D. W. Tiemann i S. M. Markle (1978). Ich zdaniem, każde pojęcie teoretyczne (*concept*) musi być nauczane i sprawdzane poprzez wiele przykładów, gdyż

1) reprezentuje obszerną klasę przedmiotów, zdarzeń, idei lub relacji, a

2) każdy z tych desygnatów wskazuje pewne cechy „krytyczne", wspólne, decydujące o przynależności do danej klasy, oraz cechy zmienne, odróżniające poszczególne desygnaty między sobą.

Tiemann i Markle proponują tworzenie list przykładów i „nieprzykładów" desygnatów każdego pojęcia oraz podzielenie obu tych list na dwie części, z których jedna będzie wykorzystana do nauczania pojęcia, a druga – do sprawdzania jego zrozumienia przez uczniów. W tym drugim przypadku pozycje listy powinny być dobierane losowo.

Operowanie gotowymi listami pozwala, według Tiemanna i Markle'a, na diagnozę dwu składowych procesów rozumienia pojęcia:

1) uogólnienia, to jest „zdolności przyporządkowania danej nazwy nowym prawdziwym desygnatom tego pojęcia" (Roid i Haladyna 1982, s. 150), oraz

2) różnicowania, to jest „przyporządkowania innej nazwy, gdy przykład nie jest desygnatem danego pojęcia, mimo iż ma pewne cechy wspólne jego desygnatom" (tamże, s. 151). Najlepsze do tego celu są „bliskie nieprzykłady", bardzo podobne do jednego z prawdziwych desygnatów pojęcia, ale nie posiadające jednej z jego cech krytycznych.

Analizy Tiemanna i Markle'a dowodzą, że sięgnięcie „ponad jednostkowe fakty" – będące dążeniem technologów pisania zadań (Roid i Haladyna 1982, s.

145) – nie uwalnia pomiaru dydaktycznego od szczegółowej informacji. Wydaje się nawet, że oderwanie się od materiału podręcznikowego zwiększa ilość tej informacji, przynajmniej na etapie precyzowania dziedzin czynności i systematycznego produkowania zadań.

II. OTWARTE CZY ZAMKNIĘTE?

Pomiar sprawdzający, mający umożliwić oszacowanie stopnia opanowania określonej dziedziny czynności przez uczniów, stawia dylemat rodzaju zadań testowych w nowym świetle. Jak wiadomo, zadania otwarte (rozprawki, krótkiej odpowiedzi, z luką) wymagają od badanego samodzielnego sformułowania odpowiedzi, podczas gdy w zadaniach zamkniętych (typu „prawda – fałsz", wyboru wielokrotnego, na dobieranie) wybiera on jedną z gotowych odpowiedzi. Powstaje pytanie, na ile ten drugi rodzaj zadań zniekształca, a przede wszystkim – ułatwia, wykonanie mierzonej czynności przez badanego. Przedtem nie miało ono większej doniosłości, gdyż można było przyjąć, że czynniki formalne oddziałują na wszystkich badanych w przybliżeniu podobnie, a więc różnicowanie ich osiągnięć jest zakłócone.

Warto zauważyć, że formułowanie odpowiedzi przez ucznia także może prowadzić do zniekształcenia obrazu czynności, i to zarówno przez nieudolność językową ucznia, jak przez upiększanie odpowiedzi pewnymi terminami. „Testy pisemne – napisał Robert Ebel (1979, s. 48) – mocno zależą od słów. Słowa są zgrabnymi i niezbędnymi narzędziami myślenia i porozumiewania się, ale reprezentują tylko środki, a nie cele uczenia się. Ich użyteczność dla nas zależy od naszej wiedzy niewerbalnej o tym, co one symbolizują". Tropiąc nawyki „bezsensownej werbalizacji" w nauczaniu, autor ten wyraża pogląd, iż „większość uczniów, a także większość ludzi dorosłych, rozróżnia i stosuje więcej słów, zwrotów i nawet stereotypowych całych zadań, niż jasno rozumie" (tamże). Zadania otwarte mierzą zawsze płynność słowną (szerzej: płynność operowania symbolami), ale nie podejmowano dotychczas poważniejszych badań nad podobnym źródłem zniekształceń obrazu opanowanych czynności, włączając na ogół tę zdolność do definicji dziedziny czynności, na zasadzie „wie i umie o tym powiedzieć".

Ułomność zadań zamkniętych jest poważniejsza, gdyż uczeń może niekiedy wykonać całkiem inną czynność niż przewidziana w planie testu, a mimo to trafić na prawidłową odpowiedź. Z grubsza biorąc, możliwe są cztery strategie rozwiązywania zadań wyboru wielokrotnego:

1. Strategia samodzielnego formułowania odpowiedzi (*frontal attack*), polegająca na rozwiązaniu zadania na podstawie informacji zawartej w trzonie zadania i porównaniu własnej odpowiedzi z odpowiedziami podanymi w teście. Gdy trzon wszystkich zadań jest samoistny, to w zwykłych warunkach uczniowie rozwiązują do 75 procent zadań według tej strategii. Oszacowanie to

jest dokonane na podstawie kilku badań w zakresie matematyki, przeprowadzonych w Polsce (m. in. Nowik 1984).

2. Strategia eliminacji dystraktorów, polegająca na kolejnym odrzuceniu sfalsyfikowanych lub zbyt mało subiektywnie prawdopodobnych odpowiedzi. Jest to najlepsza z możliwych strategii przy niesamoistnym (zależnym od zbioru odpowiedzi) trzonie zadania, zwłaszcza gdy chodzi o wybór najlepszej, a nie – bezwzględnie prawdziwej, odpowiedzi. Stosują ją także badani o wiedzy częściowej oraz chcący uniknąć trudu samodzielnego rozwiązywania zadania o samoistnym trzonie. W zwykłych warunkach uczniowie rozwiązują do 52 procent zadań według tej strategii, ale liczba ta jest wyższa dla słabszych uczniów.

3. Strategia analiz formalnych, polegająca na stosowaniu technik pozamerytorycznego porównania odpowiedzi, w tym przez wykorzystanie ukrytych wskazówek (np. długość lub występowanie pewnych słów) oraz interpretację konstruktora (jego sposobu „ukrywania" prawidłowej odpowiedzi, produkowania dystraktorów). Im test jest lepiej skonstruowany, tym zakres skutecznych analiz formalnych jest mniejszy. Bywa tak jednak, że prawidłową odpowiedź łatwo wskazać nie czytając trzonu zadania. W klasycznym przypadku, jeżeli kolejne odpowiedzi zawierają elementy: (1) ab, (2) bc, (3) cd, (4) ef, to odpowiedzią prawidłową jest „bc", gdyż zawiera elementy najczęściej powtarzające się w czterech odpowiedziach.

4. Strategia „ślepego zgadywania" (*blind guessing*), polegająca na kierowaniu się przeczuciem, przypadkiem lub inną pozaintelektualną zasadą w wyborze odpowiedzi. W zwykłych warunkach strategię tę stosują tylko najsłabsi uczniowie, i to po wyczerpaniu innych możliwości, co sięga ogółem 10 procent rozwiązywanych zadań.

Mimo ograniczonego zasięgu występowania, strategie 2 – 4 są źródłem trosk pedagogów, nie tylko z powodu nieprawnie zdobytych punktów, lecz także z powodu demoralizacji (strat motywacyjnych) uczniów, przyzwyczajanych do omijania trudności i pozorowania wiedzy.

Stanowisko anglosaskich specjalistów pomiaru dydaktycznego wobec obniżenia trafności testów osiągnięć przez dopuszczenie strategii innych niż samodzielne formułowanie odpowiedzi przez uczniów, zwłaszcza strategii analiz formalnych i „ślepego zgadywania", jest – z naszego (środkowoeuropejskiego) punktu widzenia – bardzo niefrasobliwe.

Bruce Choppin (1974) zestawił wyniki testowania ponad 10 tysięcy trzynastolatków w 14 krajach, w tym w Polsce, zadaniami matematycznymi w różnej formie. Pod względem tendencji do zgadywania, mierzonej stosunkiem oszacowania liczby zadań zgadywanych do oszacowania liczby zadań, na które uczeń nie znał prawidłowej odpowiedzi, najwyżej uplasowały się Japonia, kraje anglojęzyczne, Holandia i Szwecja. Choppin tak to skomentował (s. 41): „Pierwszym czynnikiem jest etos społeczny. Systemy socjalistyczne wydają się

zniechęcać do zgadywania lub może raczej powinno się powiedzieć, że systemy niesocjalistyczne zachęcają do tego. Polska, Węgry, Włochy, Finlandia, Indie i Chile plasują się nisko w tabelach zgadywania. To nie dziwi, gdy weźmie się pod uwagę międzykrajowe różnice w praktyce pedagogicznej. Podczas gdy uczniowie w Stanach Zjednoczonych i paru innych krajach są otwarcie nauczani, jak dawać sobie radę z zadaniami zamkniętymi i jakie strategie przyjmować, aby uzyskać najwyższe wyniki, uczniowie w krajach socjalistycznych są wychowywani ku innym celom".

Pogląd Choppina wymaga dwu uzupełnień:

1. Wskazana grupa krajów nie popierających „zgadywania" jest skrajnie niejednorodna:

a. pod względem średniego poziomu osiągnięć matematycznych, mierzonego zadaniami otwartymi, kraje socjalistyczne i Włochy wyprzedzają wszystkie kraje anglojęzyczne (Anglia, Australia, Nowa Zelandia, Szkocja i USA), a z pozostałych krajów tylko Japonia i Holandia mają wynik trochę wyższy (s. 25):

b. Finlandia, Indie i Chile zamykają stawkę krajów w tej tabeli.

2. Pojęcie zgadywania jest rozumiane w tych analizach bardzo szeroko. Ponieważ jego wskaźnikiem jest pewne przetworzenie liczby popełnionych błędów (wzór będzie podany w następnym punkcie), obejmuje ono nie tylko strategię 4 („ślepego zgadywania"), ale i wszystkie poprzednie, gdy bywają nieskuteczne. W ten sposób ustalana tendencja do zgadywania rozwiązań zadań otwartych w matematyce (s. 40) wyniosła od 35 procent (Chile) do 80 procent (Japonia), a dla Polski wyniosła 50 procent! Chodzi więc raczej o podejmowanie ryzyka, rzeczywiście intensywnie wspierane w rozwiniętych krajach kapitalistycznych.

Opanowanie kompletu strategii rozwiązywania zadań testowych (wszelkich form) nazwane jest obyciem testowym (*test-wiseness*) ucznia. „Taksonomia" obycia testowego jest dość rozbudowana (Millman i in. 1965). Obejmuje kategorie wychowawcze cenne, a mianowicie:

A. wykorzystanie czasu (niezwłoczne rozpoczynanie, pozostawianie trudniejszych zadań na koniec) i

B. unikanie błędów (przestrzeganie instrukcji, sprawdzanie rozwiązań), a także kategorie wychowawcze – z naszego punktu widzenia – wątpliwe i szkodliwe (podaje je we własnym porządku);

C. wczuwanie się w intencje konstruktora (udzielanie odpowiedzi według oczekiwań konstruktora testu i na zakładanym przez niego poziomie ścisłości):

D. analizy formalne (porównywanie odpowiedzi pod względem logicznym, poszukiwanie związku z odpowiedziami na inne zadania);

E. wykorzystanie ukrytych wskazówek (tropienie wszelkich niezręczności konstruktora w doborze dystraktorów) i

F. zgadywanie („Zawsze zgaduj, jeżeli nie przewidziano kary za błędy").

Stwierdzono (w USA) wzrost obycia testowego uczniów wraz ze szczeblem szkoły (Slahter i in. 1970), inteligencją i wynikami pomiaru osiągnięć językowych (Sarnacki 1979), ale nie tak silny, by rezygnować z zajmowania się tym obyciem jako odrębnym zjawiskiem. Na ogół specjaliści amerykańscy proponują (tamże, s. 274), by (1) kształcić konstruktorów testów w umiejętności oczyszczania zadań z ułatwień w stosowaniu strategii innych niż samodzielne formułowanie odpowiedzi i (2) ćwiczyć uczniów o małym obyciu testowym w zakresie wszystkich kategorii przedstawionej wyżej „taksonomii". Tę drugą funkcję spełniają w pewnym stopniu łatwo dostępne w krajach anglosaskich przewodniki dla zdających egzaminy testowe (np. Feder 1979). Wypełniają je przykłady zadań oraz rady jak przechytrzyć konstruktora testu.

W Polsce należy zalecić stosowanie zadań otwartych w nauczycielskich testach sprawdzających oraz stosowanie zadań zamkniętych w typowych masowych badaniach osiągnięć szkolnych. To drugie jest przede wszystkim koniecznością organizacyjną. W powojennej historii szkolnictwa w Polsce znane są przykłady badań, które nagromadziły ogromne ilości swobodnych wypowiedzi uczniów, przekraczające możliwości odczytania i interpretacji (Okoń 1970, rozdz. III). Autokodowanie odpowiedzi na zadania wyboru wielokrotnego i elektroniczne sczytywanie kart odpowiedzi bywa niezastąpione w testowaniu na szeroką skalę.

O zawodności poprawki na zgadywanie

Niedogodność zadań zamkniętych w pomiarze sprawdzającym polega na tym, że wyników tych zadań nie potrafimy przekształcić na wyniki odpowiednich zadań otwartych. Dzieje się tak dlatego, że – w zwykłych warunkach – nie potrafimy odtworzyć strategii rozwiązywania poszczególnych zadań przez uczniów.

Spośród licznych prób skorygowania efektu pozamerytorycznych strategii rozwiązywania zadań zamkniętych (zob. Diamond i Evans 1973; Niemierko 1975b, s. 119-124), większość uwagi teoretyków skupia klasyczna poprawka na zgadywanie (negatywna poprawka na losowy wybór odpowiedzi), stanowiąca iloraz liczby błędnych odpowiedzi i liczby dystraktorów (uzasadnienie podaje: Niemierko 1975b, s. 120). Jej założeniem jest pochodzenie wszystkich błędów ze „ślepego zgadywania" odpowiedzi (strategii 4). Wiadomo wszakże, że pozostałe strategie mogą także prowadzić do odpowiedzi nieprawidłowej, jako że istnieje „wiedza błędna" w postaci fałszywych wiadomości i luk w umiejętnościach uczniów. Z tego powodu klasyczna poprawka na zgadywanie okazuje się zwykle zbyt silna.

Bruce Choppin (1974) sformułował na podstawie scharakteryzowanych poprzednio szerokich badań wnioski, które mogą być streszczone następująco:

1. Strategie uczniów rozwiązujących trudniejsze (dla nich) zadania zamknięte (wyboru wielokrotnego i „prawda – fałsz") są tak „skrajnie złożone", że jest nieprawdopodobne, aby jakikolwiek prosty model matematyczny pozwolił je opisać. Strategie te różnią się między krajami, nadto zależą od uzdolnień, a zapewne także od innych cech osobowości uczniów.

2. Tendencja do zgadywania – mierzona stosunkiem (oszacowania) liczby zadań zgadywanych do (oszacowania) liczby zadań, na które uczeń nie uznał prawidłowych odpowiedzi – wzrasta wraz z poziomem osiągnięć uczniów, co znaczy, iż uczniowie bardziej zaawansowani chętniej podejmują ryzyko niż uczniowie słabsi, zwłaszcza w zadaniach z dużą liczbą odpowiedzi do wyboru i w zadaniach otwartych (s. 43).

3. Nie można spodziewać się wzrostu trafności i rzetelności wyników testowania na skutek zastosowania poprawki. Jednakże ze względu na różnice tendencji do zgadywania wskazane byłoby zastosowanie 1/3 lub 1/4 klasycznej poprawki na zgadywanie, a ze względu na upodobnienie średniego wyniku zadań zamkniętych do średniego wyniku zadań otwartych – około 1/2 tej poprawki (wniosek z tabeli na s. 49). Niestety, rozpiętość różnic łatwości poszczególnych zadań zamkniętych i odpowiednich (równoległych treściowo) zadań otwartych prawdopodobnie wzrośnie po zastosowaniu poprawki.

4. Należy unikać zadań „typu wstecz" (*backward type items*), to jest takich, które mogą być łatwo rozwiązane przez podstawianie gotowych odpowiedzi (np. pierwiastków równania). Prawa rządzące doborem strategii rozwiązań są tu szczególnie zawiłe (obejmują posługiwanie się wiedzą częściową), a poprawki – mało skuteczne.

W Polsce zgrabne metodologicznie badania nad zgadywaniem odpowiedzi na zadania wyboru wielokrotnego przeprowadził Tadeusz Patrzałek (1982). Przebadał on dwie losowo równoległe próby po 250 uczniów dziesięcioma zadaniami polonistycznymi w dwu wersjach: krótkiej odpowiedzi – w pierwszej próbie i wyboru wielokrotnego – w drugiej próbie. Średnie wskaźniki łatwości odpowiednio 0,646 i 0,604, co dało różnicę bardzo istotną statystycznie (obliczenie własne – B. N.). Ponieważ uczniowie w zasadzie nie opuszczali zadań, do przekształcenia średniego wyniku zadań zamkniętych na średni wynik zadań otwartych potrzeba było 0,48 klasycznej poprawki na zgadywanie. Podobny wynik otrzymał T. Patrzałek w dwu innych badaniach. Tak to komentuje: „Okazało się zatem, że klasyczna poprawka na zgadywanie zamiast być za mała (taka jest na ogół opinia krytyków zadań wyboru), jest zbyt duża. Zadania wyboru są łatwiejsze od zadań otwartych, ale nie aż o tyle, na ile wskazywałoby statystyczne prawdopodobieństwo sukcesu w zgadywaniu".

Wobec widocznej zgodności wyników badań w dwu głównych przedmiotach nauczania szkolnego w Polsce, można przyjąć następującą regułę roboczą: połowa klasycznej poprawki na zgadywanie pozwala przekształcić średni wynik badania grupy uczniów testem złożonym z zadań

wyboru wielokrotnego na użyteczne oszacowanie średniego wyniku odpowiednich zadań otwartych. Reguła ta nie nadaje się jednak do interpretacji wyników testowania pojedynczych uczniów i wyników rozwiązywania pojedynczego zadania przez wielu uczniów.

Podnoszenie użyteczności zadań wyboru wielokrotnego

Obraz osiągnięć uczniów sprawdzanych zadaniami zamkniętymi jest zaćmiony przez niepełnowartościowe strategie rozwiązywania tych zadań. Liczne badania porównawcze, prowadzone także i w Polsce (Czarnota 1977; Zywer 1977) doprowadziły do ujawniania dwu statystycznych właściwości zadań wyboru wielokrotnego, ograniczającego przydatność takich zadań do szczegółowej diagnozy osiągnięć:

1. Wskaźniki łatwości zadań wyboru wielokrotnego są nie tylko, średnio biorąc, wyższe niż wskaźniki łatwości odpowiednich zadań krótkiej odpowiedzi, ale także mniej zróżnicowane. Trudne zadania otwarte stają się na ogół łatwiejsze w wersji wyboru wielokrotnego, a łatwe zadania otwarte często stają się nieco trudniejsze w wersji wyboru wielokrotnego. To drugie, dość zaskakujące, zjawisko będzie przedmiotem osobnej dyskusji. Bez względu na przyczyny, spadek wariancji wskaźników łatwości zadań mierzących różne czynności uczniów utrudnia odróżnienie czynności opanowanych od czynności wymagających dalszego nakładu pracy.

2. Moc różnicująca (korelacja wyników zadania z wynikiem testu) zadań wyboru wielokrotnego jest na ogół niższa niż moc różnicująca zadań krótkiej odpowiedzi. Wskutek tego testy złożone z zadań wyboru wielokrotnego są mniej rzetelne, jeżeli nie są znacznie dłuższe od testów złożonych z zadań otwartych.

Środki zaradcze wobec ujawnionych słabości zadań wyboru wielokrotnego są poszukiwane głównie w odpowiednim doborze i wykorzystaniu dystraktorów. Oto trzy kierunki poszukiwań:

1. Nadawanie dystraktorom specjalnej wartości diagnostycznej, opartej na wybranej teorii psychologicznej lub dydaktycznej. Na przykład Kurt Bergling (1974) dokonał analizy wybranych zadań testów przyrodniczych Międzynarodowego Stowarzyszenia Badań Osiągnięć Pedagogicznych „zorientowanej na teorię Piageta", to jest wyróżnił w nich odpowiedzi charakterystyczne dla stadiów rozwoju myślenia: przedoperacyjnego, operacji konkretnych i operacji formalnych. Wzorując się na nim, Ewa Guttmejer (1982) zbudowała testy rozumienia treści symbolicznej przez uczniów, składające się z zadań wyboru czterech interpretacji tej treści: faktycznej, baśniowej, refleksyjnej i symbolicznej (odpowiedź najwyżej punktowana). Taki dobór dystraktorów pomnaża możliwości jakościowej diagnozy opanowanych czynności, nadaje wszystkim odpowiedziom na zadanie pewne znaczenie pozytywne.

2. Dobieranie dystraktorów atrakcyjnych, to jest mających znamiona prawdziwości dla największej liczby badanych.

Najogólniej, im odpowiedzi na zadanie są bardziej jednorodne (wzajemnie bliskie), tym trudniej badanemu wskazać odpowiedź prawidłową (Ebel 1979, s. 159). Skojarzenia obrazów słów i pojęć w umyśle ucznia bywają jednak odległe od struktur programowych, a zręczny konstruktor zadań testowych potrafi nawiązać do tych skojarzeń. Okazuje się, że trzy metody doboru dystraktorów:

1) subiektywna, polegająca na osądzie doświadczonych autorów zadań;

2) frekwencyjna, polegająca na wykorzystaniu najczęstszych błędów popełnionych przez rozwiązujących odpowiednie zadanie otwarte w próbnym testowaniu, i

3) korelacyjna, polegająca na podobnym wykorzystaniu błędów najwyżej ujemnie korelujących z ogólnym wynikiem testowania, dają w przybliżeniu równoważne wyniki (Owens i in. 1970). Metoda frekwencyjna może być więc zalecona początkującym konstruktorom zadań wyboru wielokrotnego, niepewnym swojej interpretacji procesów myślowych uczniów.

Dystraktory nie pracują „równym frontem". Zwykle jeden z nich, rzadko – dwa, są tak atrakcyjne dla badanych, że pozostałe nie odgrywają większej roli. Zagadnienie to podjął Tadeusz Patrzałek (1982), wprowadzając pojęcie ekspresji odpowiedzi do wyboru, określone jako „moc oddziaływania (presja), jaką wywiera na przebieg rozwiązywania zadania zamkniętego sam sposób sformułowania odpowiedzi" (s. 85). Autor ten jest przekonany, że w „zadaniach polonistycznych odpowiedzi o słabej ekspresji mają często postać liczb lub dat, osobnych nazw lub nazwisk, wyrazów obcych (rzadko symboli); odpowiedzi sugestywne bywają raczej pełnymi zdaniami lub równoważnikami" (s. 86). Gdy konstruktorowi zadania uda się dobrać dystraktor (lub dystraktory) o dużej ekspresji, zadanie wyboru wielokrotnego bywa trudniejsze od odpowiedniego zadania otwartego, a strategie analiz formalnych i „ślepego zgadywania" nie są przez badanych stosowane.

Także modelowe analizy statystyczne wykazują, że zwiększanie liczby odpowiedzi do wyboru niewiele podnosi jakość zadań. Teoretycy (Grier 1975; Lord 1977) ustalili, że trzy odpowiedzi (odpowiedź prawidłowa i dwa dystraktory) są optymalne w tym sensie, iż 80 zadań potrójnego wyboru stanowi lepszy test niż 60 zadań poczwórnego wyboru i 120 zadań podwójnego wyboru. Brano pod uwagę oczekiwaną rzetelność testu. Okazało się przy tym, że dłuższe testy złożone z mniej rozbudowanych zadań lepiej różnicują badanych o wysokich wynikach, a krótsze testy złożone z bardziej rozbudowanych zadań lepiej różnicują badanych o niskich wynikach, częściej stosujących niepożądane strategie rozwiązywania zadań zamkniętych (Lord 1977, s. 36).

3. Rejestrowanie eliminacji dystraktorów, polegające na zróżnicowanym punktowaniu oznaczenia od jednego do wszystkich dystraktorów przez bada-

nego. Takie włączenie wiedzy częściowej badanego do dziedziny sprawdzanych czynności daje pewien (niewielki) przyrost trafności i rzetelności pomiaru (Collett 1971). Odwróceniem tego podejścia jest procedura „odpowiadaj, aż trafisz" (*answer-until-correct*), w której badany otrzymuje natychmiastową (np. po potarciu karty odpowiedzi gumką lub z komputera) informację o tym, czy kolejny wybór prawidłowej odpowiedzi na zadanie był trafny. Istnieją różne sposoby punktowania wyników obydwu odmian eliminacji (Frary 1980). Empiryczne próby stosowania tych sposobów dają „mieszane rezultaty" (Hanna 1977; Poizner i in. 1978).

III. PODSUMOWANIE

W artykule podjęto kwestie: (1) opisu i oceny prób budowania „technologii" pisania zadań testów osiągnięć szkolnych i (2) zmieniających się współcześnie poglądów na wartość zadań wyboru wielokrotnego.

W ostatnich dwudziestu latach z projektami zaawansowanych procedur wytwarzania zadań testowych wystąpili: J. Bormuth, L. Guttman, W. Hively i R. Anderson. Żadna z tych procedur nie nadaje się jeszcze do masowego zastosowania.

W związku z rozwojem teorii pomiaru sprawdzającego odżyły wątpliwości co do użyteczności zadań wyboru wielokrotnego i stosowanych poprawek na zgadywanie wyników tych zadań. Autor artykułu przedstawia strategię unikania niektórych wad takich zadań przez analizę ich rozwiązywania, zredukowaną (do 1/2) poprawkę na zgadywanie i podnoszenie jakości dystraktorów.

CYTOWANA LITERATURA

Anderson R. C., *How to construct achievement tests to assess comprehension*, „Review of Educational Research" 1972, s. 145-170.

Bergling K., *The development of hypothetico-deductive thinking in children*, Stockholm 1974.

Berk R. A., *The application of structural facet theory to achievement test construction*, „Educational Research Quaterly" 1978, s. 62-72.

Bormuth J. H., *Close test readability. Criterion-referenced scores*, „Journal of Educational Measurement" 1968, s. 189-196.

–, *On the theory of achievement test items*, Chicago 1970, University of Chicago Press.

Choppin B. H., *The correction for guessing on objective tests*, Buçuresti 1974, IEA.

Collett L. S., *Elimination scoring: An empirical evaluation*, „Journal of Educational Measurement" 1971, s. 209-214.

Czarnota A., *Porównanie trudności i mocy różnicującej zadań otwartych i zamkniętych* (praca magisterska, UMSC), Lublin 1977.

Diamond J., Evans W., *The correction for guessing*, „Review of Educational Research" 1973, nr 2.

Diederich P. B., *Bormuth's On the theory of achievement test items*, „Educational and Psychological Measurement" 1970, s. 1003-1005.

Ebbinghaus H., *Über eine neue Methode zur Prüfung geistiger Fähigkeiten und ihre Anwendung bei Schulkindern*, „Zeitschrift für Psychologie und Physiologie der Sinnenorgane" 1987, s. 401-457.

Ebel R. L., *Essentials of educational measurement*, Third edition, Englewood Cliffs 1979, Prentice – Hall.

Engel J. D., Martuza V. R., *A systematic approach to the construction of domain-referenced multiple-choice test items. Paper presented at the meeting of the APA*, Washington 1976.

Feder B., *The complete guide to taking tests*, Englewood Cliffs 1979, Prentice – Hall.

Frary R. W., *The effect of misinformation, partial information and guessing on expected multiple-choice test items scores*, „Applied Psychological Measurement" 1980.

Frederiksen C. H., *Representing logical and semantic structure of knowledge acquired from discourse*, „Cognitive Psychology" 1975, s. 371-458.

Grier J. B., *The number of alternatives for optimum test reliability*, „Journal of Educational Measurement" 1975, s. 109-113.

Guttman L., *Integration of test design and analysis. Proceedings of the 1969 Invitational Conference on Testing Problems*, Princeton 1969, ETS.

Guttmejer E., *Rozumienie treści symbolicznych przez dzieci z klas III – V*, Warszawa 1982, PWN.

Hanna G. S., *A study of reliability and validity effects of total and partial immediate feedback in multipe-choice testing*, „Journal of Educational Measurement" 1977, s. 1-7.

Hively W., Maxwell G., Rabehl G., Sension D., Lundin S., *Domain-referenced curriculum evaluation*, Los Angeles 1973, CSE.

Lord F. M., *Optimal number of choice per item – A comparison of four approaches*, „Journal of Educational Measurement" 1977, s. 33-38.

Macready G. B., Mervin J. C., *Homogeneity within item forms in domain-referenced testing*, „Educational and Psychological Measurement" 1973, s. 351-360.

Millman J., Bishop C. H., Ebel R. E., *An analysis of test-wiseness*, „Educational and Psychological Measurement" 1965, s. 707-726.

Niemierko B. (red), *ABC testów szkolnych*, Warszawa 1975, WSiP.

–, *Testy osiągnięć szkolnych. Podstawowe pojęcia i techniki obliczeniowe*, Warszawa 1975, WSiP.

–, *Rozwój teorii pomiaru dydaktycznego na świecie w latach 1960 – 1980*, „Kwartalnik Pedagogiczny" 1981, nr 1.

–, *Pomiar sprawdzający jako metoda badawcza pedagogiki*, „Studia Pedagogiczne" 1982, t. XLIV,

Nitko A. J., *Educational tests and measurement. An introduction*, New York 1983, Harcourt.

Nowik J., *Funkcjonowanie zadań wyboru wielokrotnego w sprawdzaniu osiągnięć szkolnych z matematyki*, „Oświata i Wychowanie", wersja B, 1984, nr 9.

Okoń W., *O postępie pedagogicznym*, Warszawa 1970, KiW.

Osbourn H. G., *Item sampling for achievement testing*, „Educational and Psychological Measurement" 1968, s. 95-104.

Owens R. E., Hanna G. S., Coppedge F. L., *Comparison of multiple-choice test using different types of distractor selection techniques*, „Journal of Educational Measurement" 1970, s. 87-90.

Patrzałek T., *O niektórych właściwościach polonistycznych zadań wyboru*, [w:] J. Kram i E. Polański (red.), *Z teorii i praktyki dydaktycznej języka polskiego*, Katowice 1982, Uniw. Śląski.

Poizner S. B., Nicewander W. A., Gettys C. F., *Alternative response and scoring methods for multiple-choice items: An empirical study of probabilistic and ordinal response modes*, „Applied Psychological Measurement" 1978, s. 83-96.

Popham W. J., *Educational evaluation*, Englewood Cliffs 1975, Prentice – Hall.

–, *Criterion-referenced measurement*, Englewood Cliffs 1978, Prentice – Hall.

Program szkoły podstawowej. Matematyka, klasy IV – VIII, Warszawa 1984, WSiP.

Roid G. H., Haladyna T. M., *A technology for test-item writing*, New York 1982, Academic Press.

Rupprecht H., *Konstruktion von Testaufgaben nach einem Verfahren von Bormuth*, [w:] K. J. Klauer i in., *Lehrzielorientierte Tests*, Düsseldorf 1972, Schwann.

Sarnacki R. E., *An examination of test wiseness in the cognitive test domains*, „Review of Educational Research" 1979, s. 252-259.

Shoemaker D. M., *Toward a framework for achievement testing*, „Review of Educational Research" 1975, s. 127-147.

Sitarska W., *Pomiar osiągnięć koniecznych w nauczaniu początkowym matematyki*, Bydgoszcz 1987, WSP.

Slahter M. J., Koehler R. A., Hampton S. H., *Grade level, sex, and selected aspects of test-wiseness*, „Journal of Educational Measurement" 1970, s. 119-122.

Taylor W. L., *Close procedure: A new tool for measuring readability*, „Journalism Quaterly" 1953, s. 415-433.

Tiemann P. W., Markle S. M., *Analyzing instructional content: A guide to instruction and evaluation*, Champaigne 1978, Stipes.

Zywer U., *Trafność i rzetelność testów osiągnięć szkolnych z matematyki w wersji z zadaniami otwartymi i zamkniętymi* (praca magisterska, UMCS, Lublin 1977).

BOLESŁAW NIEMIERKO

AUF DER SUCHE NACH EINEM OPTIMALEN KONZIPIERUNGSVERFAHREN DER TESTAUFGABEN ZUR ÜBERPRÜFUNG DER SCHÜLERLEISTUNGEN

Zusammenfassung

Im vorliegenden Aufsatz wurden folgende Fragen besprochen: 1) Beschreibung und Bewertung der Versuche des Entwerfens des Konzipierungsverfahrens der Testaufgaben zur Überprüfung der Schülerleistungen; 2) die Entwicklung der modernen Anschauungen von der Nützlichkeit der Wahlaufgaben.

In den letzten zwanzig Jahren wurden die Konzipierungsverfahren der Testaufgaben von J. Bormuth, L. Guttman, W. Hively und R. Anderson entwickelt. Keines von diesen läβt sich aber im allgemeinen anwenden.

Mit der Entwicklung der Theorie des Prüfverfahrens entstanden wieder neue Zweifel über die Nützlichkeit der Wahlaufgaben sowie der angenommenen Wahrscheinlichkeit der zufälligen Lösung dieser Aufgaben. Der Verfasser stellt jedoch dar, wie einige Mängel dieser Aufgaben durch die Analyse deren Lösungsprozesses, die Reduzierung der Wahrscheinlichkeit der zufälligen Lösung (auf 1/2) sowie Erhöhung der Qualität der Distraktoren zu beseitigen seien.

БОЛЕСЛАВ НЕМЕРКО

В ПОИСКЕ ТЕХНОЛОГИИ НАПИСАНИЯ ЗАДАНИЙ КОНТРОЛЬНЫХ ТЕСТОВ ДОСТИЖЕНИЙ УЧЕНИКОВ

Резюме

В статье ставятся вопросы: 1) описания и оценки попыток конструирования „технологии" написания тестов школьных достижений и 2) ныне изменяющихся взглядов на ценность заданий многократного выбора.

В последние двадцать лет с проектами продвинутых процедур производства тестовых заданий выступали: И. Бормут, Л. Гуттмен, В. Хайвли, Р. Андерсон. Ни одна из этих процедур не годится пока для массового применения.

В связи с развитием теории проверочного измерения вновь появились сомнения по отношению к пригодности заданий многократного выбора и применяемых корректур на угадывание результатов этих заданий. Автор статьи представляет стратегию избегания некоторых недостатков таких заданий путем анализа процесса их решения, редуцированного (до 1/2) исправления на угадывание и повышения качества факторов рассеянности.