XIII Konferencja Diagnostyki Edukacyjnej
Uczenie się i egzamin w oczach uczniów.
Łomża, 5 - 7.10.2007

Wojciech Malec
Instytut Filologii Angielskiej
Katolicki Uniwersytet Lubelski Jana Pawła II

Efekt metody w pomiarze sprawdzającym na przykładzie testowania kolokacji języka angielskiego

Wprowadzenie

Pełne zrozumienie efektu metody porównane zostało przez Charlesa Aldersona i in. do odnalezienia świętego Graala.[1] Wprawdzie wszyscy zgadzają się co do tego, że efekt ten istnieje, jednak sprawą ciągle otwartą jest jego dokładny wpływ na wyniki testu językowego. Wpływ ten może być inny dla każdej osoby testowanej. Bardzo często jest też nieprzewidywalny. Na przykład nie mamy pewności w jakim stopniu wyniki testu wielokrotnego wyboru odzwierciedlają wiedzę i umiejętności językowe ucznia, a w jakim stopniu jego spryt testowy (*testwiseness*). Ponadto, jeżeli nawet test językowy sprawdza tylko umiejętności językowe, pozostaje problem wyboru takich metod, które są najbardziej odpowiednie do testowania konkretnych umiejętności.[2]

Może też zaistnieć sytuacja, w której pewne metody mierzą mniej więcej tę samą umiejętność, ale każdą z nich charakteryzuje inny poziom trudności. Taki wpływ metody na wynik testu jest szczególnie istotny w przypadku pomiaru sprawdzającego. Dokładne poznanie tego wpływu jest niezbędne do poprawnej konstrukcji testów oraz trafnej interpretacji ich wyników.

---

[1] Alderson J.C., Clapham C., Wall D., *Language Test Construction and Evaluation*, Cambridge University Press, Cambridge 1995.

[2] Wątpliwość może budzić, na przykład praktyka popularna w Japonii, polegająca na stosowaniu techniki wyboru wielokrotnego do sprawdzania umiejętności wymawiania angielskich głosek (ibidem, s. 45).

1. Testy sprawdzające i testy różnicujące

W testowaniu różnicującym (*norm-referenced*) układem odniesienia wyniku konkretnej osoby testowanej są wyniki innych osób, natomiast w testowaniu sprawdzającym (*criterion-referenced*) tym układem odniesienia są wymagania programowe.[3] W testowaniu różnicującym, na podstawie rang uzyskanych przez uczniów, stwierdzamy, czyje umiejętności są lepsze lub gorsze w porównaniu z umiejętnościami pozostałych uczniów. Taki test nie może być ani zbyt łatwy, ani zbyt trudny. Gdyby wszyscy uczniowie zdobyli 100% punktów, test byłby bezużyteczny, gdyż na jego podstawie nie bylibyśmy w stanie wyłonić zwycięzcy konkursu, wybrać najlepszych kandydatów na studia, zaszeregować uczniów na odpowiedni poziom w nowej szkole językowej, itd. Test różnicujący dobrze spełnia swoją rolę wówczas, gdy jego wyniki charakteryzują się dużym stopniem rozproszenia, a ich rozkład przyjmuje postać krzywej normalnej.

Natomiast w przypadku testów sprawdzających, jeżeli tylko zadania testowe odpowiednio dobrze reprezentują treści nauczania i wymagania programowe, wspomniany przypadek uzyskania przez wszystkich uczniów maksymalnej liczby punktów nie świadczy o tym, że test został źle skonstruowany. Oznacza to po prostu, że uczniowie bardzo dobrze opanowali treści nauczania. Nie ma konieczności, aby rozkład wyników takiego testu był zbliżony do rozkładu normalnego – często bywa wypiętrzony i skośny.

Z powodu tych fundamentalnych różnic inne są wymagania w stosunku do właściwości psychometrycznych tych dwóch rodzajów testów. Nieco inne są też metody analizy zadań i oceny jakości narzędzi pomiaru.[4]

Wybór testów sprawdzających wydaje się oczywisty w przypadku klasowych testów osiągnięć, obejmujących materiał językowy poszczególnych elementów języka, ponieważ najlepiej nadają się do celów diagnostycznych. Testy różnicujące natomiast dobrze spełniają swoją rolę jako globalne i integracyjne testy biegłości (np. jako egzaminy międzynarodowe, wstępne, konkursowe), a także jako testy plasujące.[5]

Jeżeli chodzi o efekt metody, należy podkreślić, że nieco inny jego rodzaj będzie istotny w obu podejściach do pomiaru. Załóżmy, na przykład, że trzech studentów uzyskuje odpowiednio 40, 50 i 60 procent za tłumaczenie tekstu z języka polskiego na angielski, natomiast 50, 60 i 70 (czyli każdy o 10% więcej) za tłumaczenie tekstu z języka angielskiego na polski. Z punktu widzenia pomiaru różnicującego efekt metody jest w tym przypadku nieistotny, ponieważ metoda nie

---

[3] Niemierko B., *Pomiar wyników kształcenia*, Wydawnictwa Szkolne i Pedagogiczne, Warszawa 1999, s. 52–53.

[4] Np. Brown J.D., Hudson T., *Criterion-referenced Language Testing*, Cambridge University Press, Cambridge 2002.

[5] Por. Komorowska H., *Sprawdzanie umiejętności w nauce języka obcego. Kontrola – ocena – testowanie*, Fraszka Edukacyjna, Warszawa 2002.

wpływa tutaj na rangi wyników tych studentów: ten sam student uzyskuje najlepszy wynik bez względu na metodę i ten sam student uzyskuje wynik najsłabszy. Jeżeli tylko korelacja między wynikami uzyskanymi przy zastosowaniu jednej metody, a tymi uzyskanymi przy użyciu drugiej metody jest wysoka, nie ma znaczenia, która z nich zostanie wybrana – obie tak samo dobrze spełniają swoją rolę, którą jest uszeregowanie studentów od najlepszych do najsłabszych. Natomiast z punktu widzenia testów sprawdzających wspomniana różnica jest bardzo istotna, ponieważ może ona oznaczać zaliczenie lub niezaliczenie egzaminu albo testu.

Z powyższego wynika, że efekt metody nie może być badany w ten sam sposób w pomiarze różnicującym i sprawdzającym. W przypadku testów różnicujących klasyczną metodą analizy metod testowania jest oparta na korelacjach *macierz wielu cech – wielu metod* (*multitrait-multimethod matrix*).[6] W przypadku pomiaru sprawdzającego, oprócz korelacji, za pomocą których możemy stwierdzić, czy metody badają tę samą umiejętność, konieczny jest także test istotności różnic między średnimi.

2. Testowanie kolokacji

Kolokacja to łączliwość wyrazów, czyli ich skłonność do występowania obok siebie. W badaniach z wykorzystaniem korpusu siłę tej łączliwości określa się, porównując prawdopodobieństwo oczekiwane współwystępowania wyrazów z prawdopodobieństwem obserwowanym. Im większa różnica między tymi prawdopodobieństwami, tym większa siła kolokacji. W sensie lingwistycznym większa siła kolokacji pociąga za sobą większą nietransparentność semantyczną (np. *pay attention*, *foot the bill*) oraz niepodstawialność składniową: na przykład wyrazy *duży* i *wielki*, choć mają podobne znaczenie, nie zawsze mogą być użyte zamiennie w kolokacjach (np. *człowiek wielkiego/*dużego serca*, ale *duże/*wielkie zachmurzenie*).

Wiele wskazuje na to, że łączliwość wyrazów jest nie tylko faktem statystycznym, wynikającym z analizy konkretnych tekstów, ale że istnieje ona także w sensie psycholingwistycznym, w leksykonie mentalnym. Aitchison[7] zwraca uwagę na to, że leksykon mentalny pełen jest rozmaitych połączeń między wyrazami, spośród których kolokacje należą do najtrwalszych.[8] Dzięki temu, jak pod-

---

[6] Po raz pierwszy zaproponowana przez Campbella i Fiske (Campbell D.T., Fiske D.W., *Convergent and discriminant validation by the multitrait-multimethod matrix*, „Psychological Bulletin" 1959, nr 56, s. 81–105).

[7] Aitchison J., *Words in the Mind: An Introduction to the Mental Lexicon*, second edition, Blackwell Publishers, Oxford 1994.

[8] Psycholingwistyczny charakter łączliwości jednostek leksykalnych znajduje potwierdzenie w istnieniu tzw. efektów poprzedzania (*priming effects*) dla kolokacji typu czasownik-rzeczownik (Bonk W., *A Psycholinguistic Investigation of Collocation*, unpublished MA thesis, University of Colorado, Boulder 2003).

kreśla Hoey, użytkownicy języka potrafią z łatwością zrekonstruować tekst, w którym całe wyrazy zostały zastąpione lukami.[9]

Z uwagi na to testy kolokacji powinny wymagać udzielenia odpowiedzi polegających na odtworzeniu całych zdań lub tekstów, z których wybrane elementy kolokacji zostały uprzednio usunięte. Inaczej mówiąc, konstruktor testu sprawdzającego znajomość łączliwości wyrazów powinien dla każdej testowanej kolokacji dostarczyć odpowiedni (tzn. typowy) kontekst, oraz skonstruować takie zadanie testowe, w którym ten kontekst umożliwiałby osobom testowanym odtworzenie całej konstrukcji.

Nie ma wątpliwości co do tego, że sprawdzanie znajomości łączliwości wyrazów jest bardzo ważnym elementem nauczania języka obcego, a konstrukcja trafnych i rzetelnych testów kolokacji jednym z priorytetów w pracy nauczyciela. Przede wszystkim dlatego, że znajomość kolokacji jest jednym z najistotniejszych składników kompetencji językowej rodzimych użytkowników języka,[10] a błędy kolokacyjne wyraźnie zdradzają naleciałości obcojęzyczne.[11] Ponadto popularne testy leksykalne nie sprawdzają dobrze wiedzy kolokacyjnej, co w przekonywającym stopniu wykazał Schmitt.[12]

Warto również dodać, że nie ma w literaturze prac poświęconych testom kolokacji, które można by określić mianem testów sprawdzających.[13] Badania opisane w tym opracowaniu są próbą wypełnienia tej luki.

3. Aspekty metody testowania

Metoda testowania to, generalnie rzecz ujmując, sposób, w jaki przebiega cała procedura testowania. Metody mogą różnić się od siebie poleceniami, limitem czasu, schematem punktowania, kryteriami poprawności odpowiedzi, materiałami, które są bodźcem do udzielenia odpowiedzi, formą oczekiwanej odpowiedzi, itd. Te i inne aspekty procedury testowania składają się na model Bachmana,[14]

---

[9] Hoey M., *Patterns of Lexis in Text*, Oxford University Press, Oxford 1991, s. 154.

[10] Pawley A., Syder H., *Two puzzles for linguistic theory: Nativelike selection and nativelike fluency*, [w:] J.C. Richards, R.W. Schmidt (red.), *Language and Communication*, Longman, London 1983, s. 191–226.

[11] Korosadowicz-Strużyńska M., *Word collocations in FL vocabulary instruction*, „Studia Anglica Posnaniensia" 1980, nr 12, s. 109–120.

[12] Schmitt N., *The relationship between TOEFL vocabulary items and meaning, association, collocation and word-class knowledge*, „Language Testing" 1999, nr 16, s. 189–216.

[13] Walidację testu *różnicującego* mierzącego kolokacje przeprowadził Bonk (Bonk W., *Testing ESL learners' knowledge of collocations*, [w:] T. Hudson, J.D. Brown (red.), *A Focus on Language Test Development: Expanding the Language Proficiency Construct across a Variety of Tests*, Second Language Teaching and Curriculum Center, University of Hawai'i, Honolulu 2001, s. 113–142).

[14] Bachman L.F., *Fundamental Considerations in Language Testing*. Oxford University Press, Oxford 1990.

tzw. *test method facets*. To, co w tradycji polskiego pomiaru dydaktycznego określane jest jako forma zadania testowego odpowiada w tym modelu, mniej więcej, terminom *format of the input* oraz *format of the expected response*. *Input* dzieli się na dwa typy, ściśle związane z formą oczekiwanej odpowiedzi: (1) *item* (czyli zadanie, które wymaga wybrania jednej z opcji albo udzielenia odpowiedzi krótkiej, np. zadanie na dobieranie, transformacje), oraz (2) *prompt* (czyli zadanie, które wymaga odpowiedzi rozszerzonej, np. temat eseju). Testy kolokacji wykorzystane w niniejszych badaniach zbudowane były z zadań pierwszego typu (*item*).

Głównym celem badań było skonstruowanie trafnych i rzetelnych testów kolokacji oraz na podstawie ich wyników udzielenie odpowiedzi na pytanie, czy w pomiarze sprawdzającym forma zadania testowego (*item format*) ma wpływ na osiągnięcia osób testowanych.[15]

4. Metoda

4.1. Osoby badane

W badaniu pilotażowym wzięło udział 60 studentów (44 kobiety i 16 mężczyzn) podzielonych na 6 grup po 10 osób, a w badaniu głównym 51 studentów (37 kobiet i 14 mężczyzn) podzielonych na 3 grupy po 17 osób. W obu przypadkach byli to studenci pierwszego roku filologii angielskiej KUL. W testowaniu próbnym do badania głównego uczestniczyło 45 studentów (38 kobiet i 7 mężczyzn) drugiego i trzeciego roku filologii angielskiej NKJO w Tomaszowie Lubelskim.

4.2. Materiały

Najistotniejsze specyfikacje testów kolokacji to: cel, konstrukt oraz metody testowania:

1. Oprócz *celów* badawczych, *celami* dydaktycznymi testów kolokacji było sprawdzenie, w jakim stopniu studenci opanowali treści kształcenia w zakresie konkretnego elementu języka, dokonanie oceny ich osiągnięć oraz zmotywowanie studentów do systematycznej pracy. Takie cele najłatwiej osiągnąć za pomocą testów, które (a) mierzą jeden konstrukt, w których (b) przedmiotem pomiaru jest uprzednio wyselekcjonowany zestaw konkretnych kolokacji, i w których (c) udzielenie poprawnej odpowiedzi nie wymaga od studenta odnoszenia się do kontekstu szerszego niż jedno zdanie.[16]

---

[15] Dokładny opis tych badań znajduje się [w:] Malec W., *The Impact of Item Format on Test Performance in Criterion-referenced Assessment of Collocations*, rozprawa doktorska, Instytut Filologii Angielskiej KUL, Lublin 2006.

[16] Zob. Read J., *Assessing Vocabulary*, Cambridge University Press, Cambridge 2000; Read J., Chapelle C.A., *A framework for second language vocabulary assessment*, „Language Testing" 2001, nr 18, s. 1–32.

2. *Konstrukt* może być zdefiniowany albo na podstawie modelu teoretycznego kompetencji językowej, albo na podstawie programu nauczania języka.[17] W przypadku tych testów konstrukt był określony na podstawie programu nauczania PNJA (Praktycznej Nauki Języka Angielskiego).

3. Kolokacje można testować na wiele różnych *sposobów*. Z metodologicznego punktu widzenia, przy niezbyt licznej próbie osób badanych, zasadne mogło być porównanie tylko kilku metod, w dodatku takich, z którymi studenci byli wystarczająco oswojeni. Do badania pilotażowego wybranych zostało 6 form zadań: (1) zadanie z luką (*L*); (2) zadanie wyboru wielokrotnego (*WW*); (3) transformacje (z użyciem podanego wyrazu) (*TW*); (4) transformacje (uzupełnianie zdania) (*TZ*); (5) poprawianie błędu (*B*); (6) tłumaczenie (*T*). W badaniu głównym porównane zostały zadania *WW*, *TW* i *B*.

Testy skonstruowane były tak, aby każda z 30 wybranych kolokacji testowana była we wszystkich formach zadania. Ilustruje to Tabela 1.

Tab. 1. Zadania testowe dla kolokacji „couldn't be bothered".

Zadanie z luką (*L*)
It was getting dark, but he ___*couldn't be*____ bothered to put the lights on.

Zadanie wyboru wielokrotnego (*WW*)
It was getting dark, but he ____ bothered to put the lights on.
A wouldn't have B didn't feel C wasn't D couldn't be

Transformacje (z użyciem podanego wyrazu) (*TW*)
It was getting dark, but he was too lazy to put the lights on. BOTHERED
It was getting dark, but he couldn't be bothered to put the lights on.

Transformacje (uzupełnianie zdania) (*TZ*)
It was getting dark, but he was too lazy to put the lights on.
It was getting dark, but he couldn't be bothered to put the lights on.

Poprawianie błędu (*B*)
It was getting dark, but he ~~wasn't~~ COULDN'T BE bothered to put the lights on.

Tłumaczenie (*T*)
It was getting dark, but *he* ____*couldn't be*____ *bothered* to put the lights on. (*nie chciało mu się*)

[17] Bachman L.F., Palmer A.S., *Language Testing in Practice. Designing and Developing Useful Language Tests*, Oxford University Press, Oxford 1996.

Powstało w ten sposób kilka wersji tego samego testu. W każdej wersji znalazły się wszystkie kolokacje i wszystkie formy zadań.[18] Istotne jest to, że bez względu na formę zadania, każda konkretna kolokacja testowana była w tym samym kontekście zdaniowym, a rozwiązanie zadania wymagało rekonstrukcji zdania kontekstowego.[19]

4.3. Procedura

W przypadku obu badań zmienną niezależną była forma zadania, a zmienną zależną wynik testu. Liczba poziomów zmiennej niezależnej odpowiadała liczbie porównywanych form zadania. Jako że przy zastosowaniu grup niezależnych istnieje duże ryzyko wpływu niekontrolowanych zmiennych ubocznych związanych z osobami badanymi, wybrany został układ z powtarzanymi pomiarami.

Wadą takiego planu jest możliwość wystąpienia efektów przeniesienia. Aby je zrównoważyć, zastosowany został plan kwadratu łacińskiego, dodatkowo zmodyfikowany poprzez przypisanie każdej pozycji poszczególnych warunków eksperymentalnych innego zestawu kolokacji. Układ ten umożliwiał przeprowadzenie całego eksperymentu jednego dnia, co dodatkowo eliminowało wpływ zmiennych zakłócających, takich jak samopoczucie, motywacja, przygotowanie do testu, itd. (zob. Tabela 2.).

Tab. 2. Plan eksperymentalny (badanie pilotażowe).

	F1	F2	F3	F4	F5	F6		
G1	K1	K2	K3	K4	K5	K6	(A)	
G2	K6	K1	K2	K3	K4	K5	(B)	
G3	K5	K6	K1	K2	K3	K4	(C)	F1-6 – forma zadania G1-6 – grupa studentów K1-6 – zestaw kolokacji A-F – wersja testu
G4	K4	K5	K6	K1	K2	K3	(D)	
G5	K3	K4	K5	K6	K1	K2	(E)	
G6	K2	K3	K4	K5	K6	K1	(F)	

[18] W układzie zrównoważonym – tak, że każda kolokacja była testowana w każdej wersji testu, ale w innej formie zadania.

[19] Poprzez: (1) uzupełnienie luki (*L*), (2) wybranie prawidłowej odpowiedzi spośród czterech opcji (*WW*), (3) sparafrazowanie innego zdania o podobnym znaczeniu w taki sposób, aby użyty został podany wyraz (*TW*), (4) sparafrazowanie innego zdania o podobnym znaczeniu poprzez uzupełnienie podanego zdania (*TZ*), (5) zidentyfikowanie i poprawienie błędu (*B*), oraz (6) przetłumaczenie fragmentu zdania kontekstowego zawierającego testowaną kolokację (*T*).

Rezultaty tak zaplanowanego eksperymentu mogą ulec zniekształceniu w przypadku wystąpienia dużych różnic między zestawami kolokacji lub między grupami osób testowanych. Z tego powodu ważne było zachowanie zasad randomizacji (badanie pilotażowe), ewentualnie dokonanie systematycznego przyporządkowania osób badanych do grup eksperymentalnych na podstawie wyników wcześniejszych testów, a w przypadku kolokacji podzielenie ich na podstawie wersji próbnej testu[20] (badanie główne).

Wszystkie testy zostały przeprowadzone w ramach programu nauczania PNJA, co gwarantowało rzetelne podejście osób badanych do eksperymentu. Ponadto były one częścią tzw. „big testów", odbywających się dwa razy w semestrze, w dużej sali, dla całego roku jednocześnie, a więc w warunkach sprzyjających spełnieniu założenia niezależności pomiarów.

5. Wyniki

5.1. Analiza zadań i testów

Analiza ilościowa zadań polegała na obliczeniu łatwości każdego zadania (*item facility*) oraz jego mocy różnicującej (*item discrimination*). W tym drugim przypadku nie był to jednak współczynnik korelacji Pearsona, tak jak przy analizie testów różnicujących. Zamiast tego obliczone zostały dwa wskaźniki mocy różnicującej progowej (*cut-score indices*): ϕ (*item phi*) oraz *A* (*agreement statistic*).[21]

W przypadku badania pilotażowego zgodność wyników poszczególnych zadań z wynikami całego testu była dosyć niska, głównie dlatego, że test był bardzo trudny. Dla Wersji F, której nie zaliczył żaden student, obliczenie współczynnika ϕ nie było nawet możliwe ze względu na zero w mianowniku. Taka sytuacja byłaby nie do zaakceptowania w przypadku egzaminu końcowego. Wymagałaby dokładnej analizy wymagań programowych i skuteczności systemu kształcenia. Jednak w przypadku pierwszego „big testu" na pierwszym roku studiów, wyniki te można było uznać za dopuszczalne.

Testy z badania głównego były częścią trzeciego „big testu", przeprowadzonego w kwietniu, a więc w czasie, kiedy studenci już dobrze zapoznali się z poziomem wymagań. Dla większości zadań z tego testu wartość współczynnika *A* była większa od 0,50, a wartość ϕ znacznie większa od zera. Tylko kilka zadań wymagało weryfikacji, jednak analiza treści nie wykazała ich niezgodności z materiałem kształcenia.

---

[20] Wersja próbna testu składała się wyłącznie z zadań *WW*.
[21] Brown J.D., Hudson T., op. cit.

Rzetelność testów kolokacji oszacowana była na podstawie trzech współczynników. Pierwszy z nich to Φ[22] (*phi coefficient*). Jest to wskaźnik wewnętrznej zgodności testu sprawdzającego, odpowiednik *alfy* Cronbacha, albo *KR20*, stosowanych dla testów różnicujących. Podobnie jak w przypadku małego ϕ, wartość tego współczynnika była znacznie wyższa w badaniu głównym – od 0,81 (Wersja A) do 0,96 (Wersja C).

Aby określić rzetelność decyzji klasyfikacyjnych (*mastery/non-mastery decisions*), obliczane były dwa współczynniki[23] – phi lambda (Φ_λ) i kappa kwadrat (κ^2). Tym razem, zarówno w przypadku badania pilotażowego jak i badania głównego, wartości współczynników były bardzo wysokie.

Trafność konstrukcyjna ustalana była na podstawie korelacji wyników testów kolokacji z wynikami testów gramatycznych i leksykalnych przeprowadzonych tego samego dnia, także w ramach „big testu". Dla każdej wersji testu z badania pilotażowego współczynnik korelacji Pearsona był wyższy dla pary Kolokacje–Słownictwo niż dla par Kolokacje–Gramatyka i Gramatyka–Słownictwo. W badaniu głównym, ze względu na rozkład danych, obliczane było τ (tau) Kendalla. Także i w tym przypadku korelacje były najwyższe między wynikami testów kolokacji i słownictwa. Krótko mówiąc, zarówno testy kolokacji, jak i testy leksykalne sprawdzały materiał językowy zaczerpnięty z tego samego fragmentu programu nauczania słownictwa, co zostało to odzwierciedlone w wynikach testów.

Aby oszacować jednowymiarowość testów kolokacji (*unidimensionality*), przeprowadzona została analiza głównych składowych (*principal component analysis*). Według Browna i Hudsona, jeżeli jeden czynnik wyraźnie dominuje nad pozostałymi, test można uznać za jednowymiarowy.[24] W wyniku tej analizy wyłoniony został tylko jeden czynnik o wartości własnej (*eigenvalue*) większej od 1, a wartość własna drugiego czynnika była kilkakrotnie mniejsza. Należy jednak dodać, że jednowymiarowość testu była wyraźnie większa w przypadku badania głównego.

Warto wspomnieć też i o tym, że na jednowymiarowość testu istotny wpływ miał schemat punktowania. W przypadku, gdy test był punktowany 0–1, a jeden punkt przyznawany był także za odpowiedź zawierającą błąd gramatyczny, test charakteryzował się największą jednowymiarowością. Świadczy to o trafności konstrukcyjnej testów kolokacji, ponieważ uwzględnienie znajomości gramatyki w kryteriach poprawności odpowiedzi sprawiało, że test w mniejszym stopniu mierzył jeden konstrukt, tzn. znajomość łączliwości wyrazów.

---

[22] Brown J.D., *Short-cut estimates of criterion-referenced test consistency*, „Language Testing" 1990, nr 7, s. 77–97.
[23] Brown J.D., Hudson T., op. cit.
[24] Ibidem.

5.2. Testowanie hipotez eksperymentalnych

Analiza wariancji przeprowadzona została pomimo tego, że rozkłady niektórych danych wyraźnie odbiegały od rozkładu normalnego. Decyzja ta spowodowana była tym, że ANOVA jest bardzo elastyczna wobec odchyleń danych od normalności. Francuz i Mackiewicz, wykorzystując program *Monte Carlo* Brooks'a, wykazali, że prawdopodobieństwo popełnienia błędu I rodzaju, czyli odrzucenia hipotezy zerowej w sytuacji, gdy jest ona prawdziwa, tylko nieznacznie zwiększa się w przypadku, gdy dane pochodzą z rozkładu, który nie jest normalny.[25] Dla pewności jednak zastosowane zostały także testy nieparametryczne (ANOVA Friedman'a oraz testy Wilcoxon'a).

Ze wszystkich analiz statystycznych[26] wynika, że forma zadania jest czynnikiem, który istotnie wpływa na rezultaty testów kolokacji (badanie pilotażowe – $F(5, 295) = 21.49$, $p < 0{,}001$; badanie główne – $F(1.57, 78.32) = 96.55$, $p < 0{,}001$). Porównania wielokrotne oraz analiza skupień (w badaniu pilotażowym) wykazały, że cztery formy zadań – zadania z luką, transformacje (oba typy) oraz tłumaczenie – nie różnią się znacząco pod względem trudności. Wszystkie są natomiast trudniejsze od zadań wyboru wielokrotnego i łatwiejsze od tych, które wymagają poprawienia błędów.

6. Dyskusja wyników

Jednowymiarowość testów kolokacji – wynikająca m.in. z dodatniej korelacji między badanymi metodami testowania – świadczy o tym, że wszystkie metody (a ściślej: formy zadania) sprawdzały ten sam konstrukt. Zatem, z punktu widzenia pomiaru różnicującego, efekt metody byłby w tych testach mało istotny: zastosowanie każdej z metod doprowadziłoby do mniej więcej takiego samego uszeregowania studentów od najlepszych do najsłabszych.

Jednakże w pomiarze sprawdzającym efekt metody jest istotny wówczas, gdy wpływa na decyzje klasyfikacyjne (zaliczenie lub niezaliczenie testu przez studenta), tzn. wówczas, gdy metody testowania charakteryzują się różnym poziomem trudności. O ważkości różnic między trudnością metod testowania kolokacji z badania głównego świadczy fakt, że gdyby cały test składał się wyłącznie z zadań wyboru wielokrotnego, wówczas tylko jeden student nie osiągnąłby progu zaliczeniowego (60%). Gdyby natomiast test

---

[25] Francuz P., Mackiewicz R., *Liczby nie wiedzą, skąd pochodzą. Przewodnik po metodologii i statystyce nie tylko dla psychologów*, Wydawnictwo KUL, Lublin 2005; Brooks G.P., *Using Monte Carlo methods to teach statistics. The MC2G computer program*, „Understanding Statistics" 2003, nr 2, s. 137–150.

[26] Szczegółowe wyniki [w:] Malec W., op. cit.

zawierał tylko zadania polegające na poprawieniu błędów, wówczas prawie połowa studentów nie zaliczyłaby go.

Wspomniane wyżej różnice w trudności wytłumaczyć można tym, że zadania wyboru wielokrotnego z natury swojej sprawdzają rozpoznanie prawidłowej odpowiedzi (*recognition*), a więc wiedzę receptywną; zadania z luką, transformacje i tłumaczenie sprawdzają przypomnienie (*recall*), a więc wiedzę produktywną; natomiast poprawianie błędów sprawdza umiejętność, którą moglibyśmy określić jako kombinację rozpoznania i przypomnienia, jednak pytanie dlaczego ta właśnie umiejętność jest najtrudniejsza, pozostaje otwarte.

Podsumowanie

Lyle F. Bachman stwierdza, że niektóre osoby preferują testy wyboru wielokrotnego i uzyskują na nich lepsze wyniki, natomiast inne osoby wypadają lepiej na testach wymagających skonstruowania własnej odpowiedzi.[27] Można nawet znaleźć przykłady badań, w których zadania zamknięte wydają są dla niektórych osób trudniejsze niż zadania otwarte.[28] Należałoby jednak zapytać, na jakiej podstawie formułowane są takie wnioski? Oczywiście, jeżeli treść zadań będzie inna, a także inny schemat punktowania, wówczas zadanie zamknięte może być rzeczywiście trudniejsze od zadania rozszerzonej odpowiedzi. Być może dlatego wiele osób odpowiada w ankietach, że woli udzielić rozszerzonej odpowiedzi pisemnej lub ustnej, zamiast wypełniać test wyboru. Należy jednak podać w wątpliwość możliwość wystąpienia takiej sytuacji, w której rozpoznanie odpowiedzi jest trudniejsze niż wyprodukowanie *tej samej* odpowiedzi. Na to w każdym bądź razie wskazują wyniki tych badań, które zostały zaprojektowane w taki sposób, że każda metoda sprawdzała tę samą treść.

[27] Bachman L.F., *Statistical Analyses for Language Assessment*, Cambridge University Press, Cambridge 2004.

[28] Kennedy P., Walstad W.B., *Combining multiple-choice and constructed-response test scores: an economist's view*, „Applied Measurement in Education" 1997, nr 10, s. 359–375.