MARCIN SMOLIK
Zakład Metodyki Nauczania Języka Angielskiego, Instytut Anglistyki,
UMCS Lublin

KU PORÓWNYWALNOŚCI OCENIANIA. WYKORZYSTANIE WIELOASPEKTOWEGO MODELU RASCHA (MODEL *FACETS*) W PROCESIE SZKOLENIA EGZAMINATORÓW: JUŻ RZECZYWISTOŚĆ CZY JESZCZE SCIENCE-FICTION?

Wprowadzenie

Wieloaspektowy model Rascha[1] (ang. *Multi-faceted Rasch Model* – MFRM)[2] jest rozszerzeniem jednoparametrowego modelu Rascha o dodatkowe **aspekty** (ang. *facets*). Poza typowymi dla testów wyboru aspektami *egzaminowani* oraz *jednostki testu*, w wieloaspektowym modelu Rascha mogą również pojawić się takie, jak: *egzaminatorzy* (wówczas do równania dodany jest parametr określający poziom surowości egzaminatorów) lub *zastosowana skala oceniania* (wówczas do równania dodany jest parametr określający poziom wyzwania, jakie stanowi osiągnięcie wyniku *k* raczej niż *k-1*).

Ujmując rzecz ogólnie, wieloaspektowy model Rascha *pozwala na modelowanie prawdopodobieństwa odpowiedzi na podstawie addytywnej kombinacji aspektów*[3]. Model ten opiera się na logistycznej transformacji zaobserwowanych wyników na skalę logitową[4] i został zastosowany w programie *FACETS*[5]. Pro-

---

[1] Linacre J.M., *Many-facet Rasch measurement*, MESA Press, Chicago, IL 1989; McNamara T., *Measuring Second Language Performance*, Addison Wesley Longman Limited, Harlow 1996; Bond T.G., Fox C.M., *Applying the Rasch model: Fundamental measurement in the human sciences*, Lawrence Erlbaum Associates, Mahaw, NJ 2001.

[2] Tłumaczenia nazw z języka angielskiego podaję za *Angielsko-polsko-słoweńskim glosariuszem terminów z zakresu testowania biegłości językowej*, Universitas, Kraków 2004.

[3] Ibidem, s. 107.

[4] Logit – jednostka miary stosowana w teorii wyniku jednostki testu (skrót od ang. **log** odds un**its**), otrzymywana z logarytmu naturalnego ilorazu prawdopodobieństwa sukcesu do prawdopodobieństwa porażki (szanse logarytmiczne – *log odds*). Skala logitowa jest skalą *przedzia-*

gram ten oblicza wartość parametru dla każdego elementu w ramach każdego aspektu (ang. *facet*)[6], a otrzymane wartości wyrażane są w logitach. Oprócz wartości w logitach, dla każdego elementu określone są: błąd standardowy (określający miarę precyzji oszacowania parametru dla danego elementu) oraz statystyki dopasowania (określające, jak dobrze dane pasują do przewidywań modelu pomiaru).

1. Analiza wyników egzaminów z wykorzystaniem programu *FACETS*

FACETS był z powodzeniem stosowany do analizy wyników różnego rodzaju egzaminów performancji (praktycznych), np. egzaminów sprawdzających sprawności pisania i mówienia w języku obcym[7], egzaminów ustnych i pisemnych sprawdzających przygotowanie do zawodu lekarza[8], egzaminów dla aktorów[9]. Jest również pewna grupa badań, w których wieloaspektowy model Rascha został zastosowany z myślą o wykorzystaniu wyników analiz podczas szkoleń dla egzaminatorów[10]. Omówieniu niektórych z zastosowanych w tych badaniach rozwiązań poświęcona będzie dalsza część artykułu.

łową, z czego wynika wiele korzyści (McNamara T., op. cit., s. 165; Bond T.G., Fox C.M., op. cit., s. 29).

5 Linacre J.M., *FACETS Rasch measurement computer program, version 3.57.0*, Winsteps.com, Chicago 2004; taż, A user's guide to FACETS, Rasch-model computer programs, Winsteps.com, Chicago 2004.

6 Aspekt (ang. facet) to np. „egzaminatorzy". W ramach aspektu występują poszczególne elementy – w tym przypadku konkretne osoby (pani X, pan Y, etc.).

7 Np.: Bachman L.F., Lynch B.K., Mason M., *Investigating variability in tasks and rater judgements in a performance test of foreign language speaking*, „Language Testing" 1995, nr 12, s. 238–257; Brown A., *The effect of rater variables in the development of an occupation-specific language performance test*, „Language Testing" 1995, nr 12(1), s. 1–15; Engelhard, Jr. G., *Examining rater errors in the assessment of written composition with a many-faceted Rasch model*, „Journal of Educational Measurement" 1994, nr 31(2), s. 93–112; Lumley T., *Assessment criteria in a large-scale writing test: what do they really mean to the raters*?, „Language Testing" 2002, nr 19(3), s. 246–276; Lynch B.K., McNamara T.F., *Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants*, „Language Testing" 1998, nr 15(2), s. 158–180; Myford C.M., Wolfe E.W., *When raters disagree, then what: examining a third-rating discrepancy resolution procedure and its utility for identifying unusual patterns of ratings*, „Journal of Applied Measurement" 2002, nr 3(3), s. 300–324.

8 Np.: Lunz M.E., Stahl J.A., *Judge consistency and severity across grading periods*, „Evaluation & the Health Professions" 1990, nr 13(4), s. 425–444; Lunz M.E., Wright B.D., *Linacre J.M., Measuring the impact of judge severity on examination scores*, „Applied Measurement in Education" 1990, nr 3(4), s. 331–345.

9 Linacre J.M., Engelhard, Jr. G., Tatum D.S., Myford C.M., *Measurement with judges: many-faceted conjoint measurement*, „International Journal of Educational Research" 1994, nr 21(4), s. 569–577.

10 Np.: Cushing-Weigle S., *Using FACETS to model rater training effects*, „Language Testing" 1998, nr 15(2), s. 263–287; Wigglesworth G., *Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction*, „Language Testing" 1993, nr 10, s. 305–335; Wilson M., Case H., *An examination of variation in rater severity over time: a study in rater*

2. Poziom surowości egzaminatora

Jedną z niewątpliwych zalet programu *FACETS* jest forma graficzna, w jakiej przedstawiane są dane wyjściowe otrzymane w wyniku przeprowadzonej analizy. Najbardziej przemawiająca wizualnie, bogata w informację i stosunkowo łatwa do zinterpretowania jest Tabela 6., zwana *All Facet Vertical „Rulers"* (*Pionowe „Linijki" Wszystkich Aspektów*), która stanowi pewnego rodzaju „mapę", umożliwiającą jednoczesne spojrzenie na wzajemne relacje pomiędzy wszystkimi elementami wszystkich aspektów (ang. *facets*) dzięki umiejscowieniu ich na wspólnej skali, której jednostką jest logit. Poniżej przedstawiono przykładową tabelę (odpowiednio sformatowaną i uproszczoną) dla trzech aspektów wymyślonego egzaminu, którymi są: egzaminowany, egzaminator, kryterium (tab. 1.).

Kolumna pierwsza (Measr) zawiera skalę w logitach, która stanowi punkt odniesienia dla wszystkich aspektów sytuacji egzaminacyjnej, co ułatwia czynienie porównań pomiędzy poszczególnymi elementami każdego aspektu, jak również pomiędzy elementami różnych aspektów. Druga kolumna (Examinee) zawiera listę 7 egzaminowanych uczniów, których imiona zostały umieszczone obok oszacowanego dla nich przez program (w skali logitowej) poziomu wiedzy (umiejętności). Egzaminatorów przedstawiono w kolumnie trzeciej (Examiner) na oszacowanym dla każdego z nich poziomie surowości. Kolumna czwarta (Item) natomiast przedstawia kryteria, według których oceniani byli uczniowie w moim wymyślonym egzaminie, uszeregowane według poziomu wyzwania, jakie stanowiły dla egzaminowanych uczniów.

Tab. 1. Przykładowa tabela otrzymana w wyniku analizy wyników egzaminacyjnych z wykorzystaniem programu *FACETS*

Measr	Examinee	Examiner	Item
	wysokie wyniki	*surowy*	*trudne*
2	Marysia, Zosia	Bożena, Jurek	gramatyka
1	Janek	Marcin, Marek, Basia	słownictwo
0	Paweł	Grzegorz	wymowa
-1	Kasia, Michał		organizacja wypowiedzi
-2	Piotr	Lidia	interakcja
	niskie wyniki	*pobłażliwy*	*łatwe*
Measr	Examinee	Examiner	Item

Dla organizatorów egzaminu czy dla trenerów przeprowadzających kursy szkoleniowe dla egzaminatorów informacje zawarte w tabeli podobnej do przedstawionej wyżej stanowią bezcenne źródło wiedzy o egzaminatorach jako grupie,

drift, [w:] M. Wilson, G. Engelhard, Jr. (red.), *Objective Measurement: Theory into Practice*, Volume 5, Alex Publishing Corporation, Stamford, Conn. 2000, s. 113–135

ale także o każdej osobie indywidualnie. Jeśli po analizie wyników egzaminu okaże się, iż poziom surowości większości egzaminatorów jest wysoki (sytuacja taka ma miejsce w tab. 1., gdzie poziom surowości większości egzaminatorów wynosi +1 lub +2), organizator egzaminu może przedsięwziąć odpowiednie kroki, mające na celu uświadomienie egzaminatorom, iż są w swoich ocenach nazbyt surowi.

Ponieważ samo mówienie o tym może być mało przekonywające, prowadzący szkolenie może wspomóc się przykładem, wykorzystując tab. 1. Przykład ów mógłby być następujący. Załóżmy, że każde z podanych w tab. 1. kryteriów (Item) oceniane jest w skali od 0 do 5, przy czym 3 pkt to niezbędne minimum, określające najniższy zadawalający poziom. W takiej sytuacji Paweł ma 50% szans na to, aby otrzymać 3 pkt w kryterium „wymowa", jeśli oceniany jest przez Grzegorza (wszystkie elementy mają oszacowaną wartość 0 logitów). Jeżeli natomiast Paweł miałby pecha i byłby egzaminowany przez Bożenę lub Jurka (poziom surowości +2), wówczas jego szanse uzyskania 3 pkt w tym kryterium spadają do 12%[11]! Powyższy przykład, choć może niezbyt wyszukany, unaocznia, jak ogromy wpływ na wynik egzaminu może mieć różnica w poziomie surowości indywidualnych egzaminatorów.

Jak pokazują wyniki wielu badań, niezależnie od rodzaju i przedmiotu egzaminu, egzaminatorzy różnią się pomiędzy sobą poziomem surowości[12]. Badacze są również zgodni co do tego, że kursy szkoleniowe mogą jedynie nieznacznie zmniejszyć te różnice, z pewnością zaś niemożliwe jest ich wyeliminowanie[13]. Wyjątkiem – lecz jakże istotnym z punktu widzenia organizatora egzaminu – są osoby oceniające prace w sposób skrajnie surowy lub skrajnie pobłażliwy[14] – osoby te „dopasowują" się do grupy. Cushing-Weigle[15], porównując poziom surowości, jaki reprezentowali kandydaci na egzaminatorów przed i po szkoleniu, w trakcie którego wykorzystała tabelę podobną do tab. 1., zaobserwowała, iż szkolenie wpłynęło na obniżenie się ich poziomu surowości, co nie jest w żaden sposób jednoznaczne z wyrównaniem się tego poziomu w całej grupie. Podobnie jednak jak Lumley i McNamara[16], zauważyła, iż osoby, które cechowała skrajna surowość, obniżyły swoje wymagania do akceptowalnego poziomu.

---

[11] McNamara T., op. cit., s. 166.

[12] Lunz M.E., Stahl J.A., op. cit.

[13] Warto tu również zaznaczyć, iż badania dowodzą, że poziom surowości, jaki charakteryzuje każdego egzaminatora, ulega nieznacznym zmianom w czasie, jednak zakres tych zmian jest sprawą bardzo indywidualną (ibidem; Lumley T., McNamara T.F., *Rater characteristics and rater bias: implications for training*, „Language Testing" 1995, nr 12(1), s. 54–71; Wilson M, Case H., op. cit). Lunz i Stahl (Lunz M.E., Stahl J.A., *Impact of examiners on candidate scores: an introduction to the use of multifacet Rasch model analysis for oral examinations*, „Teaching and Learning in Medicine" 1993, nr 5(3), s. 174–18) podają jednak, iż poziom surowości obserwowanych przez nich egzaminatorów nie zmieniał się.

[14] Lumley T., McNamara T.F., *Rater characteristics and rater bias: implications for training*, „Language Testing" 1995, nr 12(1), s. 54–71.

[15] Cushing-Weigle S., op. cit.

[16] Lumley T., McNamara T.F., op. cit.

W analizie z wykorzystaniem wieloaspektowego modelu Rascha różnice w ogólnym poziomie surowości pomiędzy egzaminatorami nie są jednak kwestią problematyczną – dopóki egzaminatorzy są w swoim poziomie surowości konsekwentni. W modelu tym poziom wiedzy (umiejętności) egzaminowanego jest szacowany dopiero po ustaleniu średniej surowości wszystkich egzaminatorów i średniej trudności zadań, co pozwala matematycznie zrekompensować różnice w poziomie surowości występujące pomiędzy poszczególnymi egzaminatorami, eliminując tym samym źródło systematycznego błędu pomiaru i zapewniając nieobciążone oszacowanie poziomu wiedzy (umiejętności) każdego egzaminowanego. Celem szkoleń powinno być zatem wyeliminowanie źródeł **losowego** błędu pomiaru.

3. Wykorzystanie programu *FACETS* do badania *nastawienia* egzaminatorów

Oprócz **ogólnego** poziomu surowości charakteryzującego każdego egzaminatora, *FACETS* umożliwia zbadanie, czy w przyznawanych przez niego wynikach nie pojawiają się pewne wzorce pobłażliwości bądź surowości w odniesieniu do konkretnej grupy kandydatów (np. mężczyzn) lub pewnych kryteriów oceniania (np. gramatyki). Taka interakcja pomiędzy egzaminatorem a pewnym elementem jakiegoś aspektu, bądź aspektem ogólnie, określana jest w programie *FACETS* słowem *bias*, które – aby odróżnić je od *test bias*, czyli tendencyjności testu – pozwolę sobie przetłumaczyć jako ***nastawienie***. Egzaminator może być zatem przychylnie bądź nieprzychylnie nastawiony do jednego lub kilku aspektów (lub tylko elementów) sytuacji egzaminacyjnej, co prowadzi do zmiany w stosunkowo stałym poziomie surowości, który go charakteryzuje. W programie *FACETS* można zbadać potencjalne interakcje pomiędzy egzaminatorem a każdym innym elementem każdego aspektu.

Dla każdej określonej w modelu potencjalnej interakcji (aspekt *egzaminator* x inny aspekt), program oblicza jej wielkość wraz ze standardowym błędem. Wyrażony w logitach wynik jest dodatkowo przekształcany na ocenę standardową (ang. z-score) poprzez podzielenie go przez błąd pomiaru. Oba wyniki wskazują zarówno zakres, jak kierunek *nastawienia*. *Nastawienie* o wartości ujemnej wskazuje na to, że egzaminator był bardziej pobłażliwy niż to przewiduje model, wziąwszy pod uwagę wszystkie posiadane na temat tego egzaminatora informacje. *Nastawienie* o wartości dodatniej wskazuje na to, iż egzaminator był bardziej surowy. Przyjmuje się, że jeżeli wartość wyniku z mieści się w przedziale <-2;+2>, to egzaminator nie wykazał ani pozytywnego, ani negatywnego *nastawienia*.

Wigglesworth[17] dokonała analizy *nastawienia* egzaminatorów oceniających ustną część australijskiego egzaminu dla imigrantów[18]. Egzamin ten przeprowa-

[17] Wigglesworth G., op. cit.

dzany był wówczas w dwóch formach: bezpośredniej i półpośredniej. Zadania w obu tych formach, choć nieidentyczne, były dość podobne i oceniane były na podstawie tych samych kryteriów według skali sześciostopniowej. Dane do eksperymentu zostały zebrane dwukrotnie dla tych samych egzaminatorów: pierwszy raz podczas przeprowadzania pilotażu, drugi raz po przeprowadzeniu testu żywego. Wszyscy egzaminatorzy oceniali wszystkie nagrane rozmowy. Wyniki przyznane przez egzaminatorów **podczas pilotażu** zostały poddane analizie z wykorzystaniem programu *FACETS*. Celem jednej z analiz było prześledzenie ewentualnych zmian w poziomie surowości egzaminatorów w interakcji z: (1) formą egzaminu oraz (2) skalą oceniania. Otrzymane rezultaty zostały wykorzystane do przeprowadzenia szkolenia uzupełniającego dla całej grupy egzaminatorów oraz do przeprowadzenia rozmów z każdym egzaminatorem indywidualnie[19].

Do wykorzystania w rozmowach indywidualnych badaczka stworzyła dla każdej osoby *Mapy Oceniania* (*Assessment Maps*), które w graficznej formie przedstawiały wyniki analizy *nastawienia* egzaminatorów, nakierowanej na badanie interakcji, jakie zaszły pomiędzy poszczególnymi egzaminatorami a kryteriami oceniania poszczególnych zadań. Pozwoliło to na zbadanie różnic w sposobie oceniania egzaminów w obu formach (bezpośredniej i półpośredniej), jak również na zbadanie, czy dany egzaminator nie wykazuje tendencji do oceniania wypowiedzi bardziej surowo lub bardziej pobłażliwie w pewnych tylko kryteriach. Przykładową *Mapę Oceniania* dla jednego z egzaminatorów przedstawiono na rys. 1.

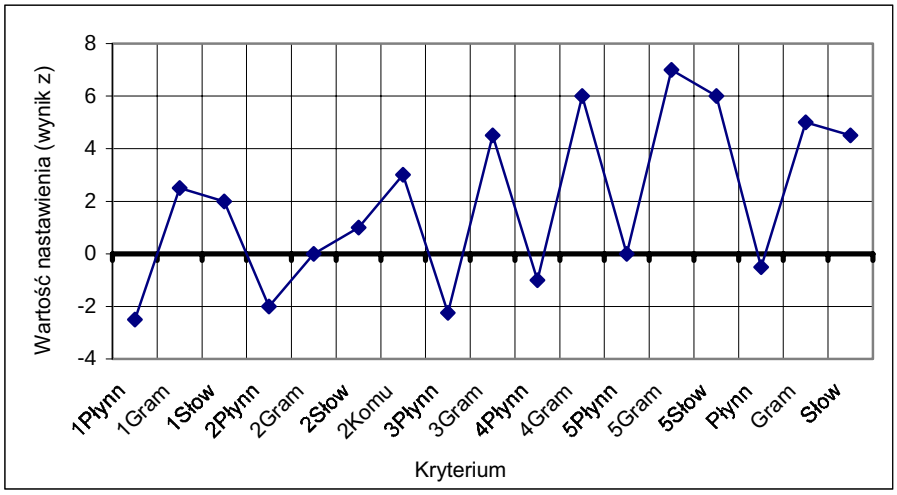
Na osi *x* przedstawiono kryteria oceniania. Egzamin składał się z 5 zadań, w których oceniane były: płynność, gramatyka, słownictwo i komunikacja (choć nie wszystkie zadania były oceniane według wszystkich kryteriów). Na podstawie ocen wystawionych za poszczególne zadania egzaminatorzy wystawiali ocenę ogólną za płynność, gramatykę oraz słownictwo (trzy ostatnie punkty na osi *x*).

Na osi *y* przedstawiono poziom *nastawienia* (bias) egzaminatora do poszczególnych kryteriów, wyrażony za pomocą not w skali *z*. Jeśli założymy, o czym pisałem już wyżej, że akceptowalny zakres wartości *nastawienia* mieści się w granicach od -2 do +2, to wykres przedstawiony na rys. 1. nie napawa optymizmem. Jedynie w przypadku 7 z 17 kryteriów (40%) oceny egzaminatora są zgodne z tym, co przewiduje model. W stosunku do pozostałych kryteriów egzaminator wykazuje nieprzychylne nastawienie, oceniając wypowiedzi uczniów w tych kryteriach bardziej surowo niż w pozostałych, biorąc pod uwagę

[18] McNamara T., op. cit., rozdz. 4.

[19] Badanie Wigglesworth było oparte na badaniu opisanym w niepublikowanym referacie Lunz i Stahla (za: McNamara, op. cit.), którzy zaobserwowali, że egzaminatorzy lepiej reagują na informację zwrotną na temat ich pracy, jeśli (1) otrzymują ją indywidualnie oraz (2) ich zachowanie nie jest porównywane do zachowania innych egzaminatorów. Z takim podejściem zdaje się nie zgadzać McNamara (McNamara T., *Language Testing*, Oxford University Press, Oxford 2000, s. 44), który twierdzi, że pewna doza presji społecznej wynikającej z poczucia wstydu i zakłopotania z powodu oceniania niezgodnie z grupą jest konieczna, aby skrajnie pobłażliwi bądź skrajnie surowi egzaminatorzy dążyli do zredukowania niepożądanych różnic.

ogólny wzorzec jego oceniania oraz oceny tych samych uczniów, które wystawili im inni egzaminatorzy.



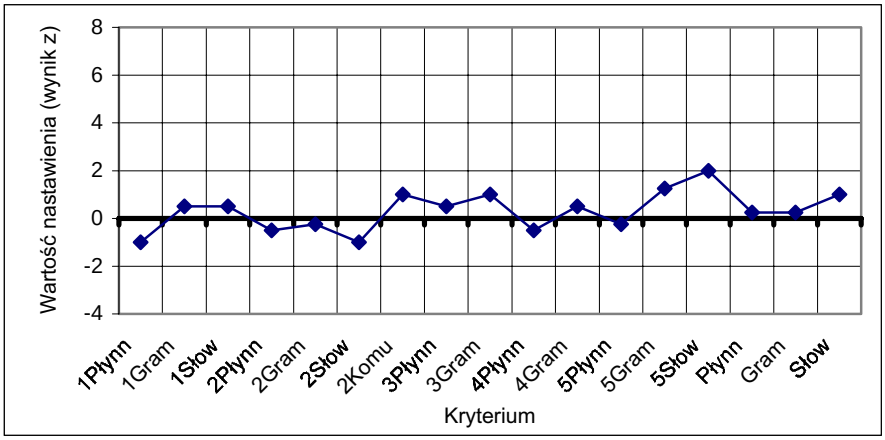
Rys. 1. Przykładowa *Mapa Oceniania* dla zadań w egzaminie półpośrednim dla wybranego egzaminatora po badaniu pilotażowym[20]

Wykres wyraźnie pokazuje, iż poziom surowości egzaminatora rośnie pod koniec egzaminu oraz że egzaminator ten wykazuje tendencję do bardziej surowego oceniania wypowiedzi uczniów w kryterium gramatyki, niezależnie od zadania. Z drugiej strony, ten sam egzaminator okazuje ogromną pobłażliwość w stosunku do kryterium płynności.

Wigglesworth przygotowała podobne wykresy dla obu form egzaminu dla każdego egzaminatora i przekazała im swoje wnioski (w sposób możliwie jak najmniej onieśmielający) po to, aby sprawdzić, czy dostarczenie egzaminatorom bardzo konkretnej informacji zwrotnej na temat ich pracy przyczyni się do wyeliminowania lub chociażby zmniejszenia niestałości w ich sposobie oceniania wypowiedzi uczniów.

Po zakończeniu szkolenia ci sami egzaminatorzy wzięli udział w ocenianiu egzaminu żywego, a przyznane przez nich oceny ponownie zostały poddane analizie *nastawienia* w programie *FACETS*. U większości autorka zaobserwowała znaczną poprawę: egzaminatorzy oceniali wszystkie wypowiedzi we wszystkich kryteriach z podobnym, charakterystycznym dla każdego poziomem surowości. Mówiąc inaczej, zanikły niepożądane interakcje, a egzaminatorów cechowała większa stałość oceniania. Wnioski takie wyciągnęła Wigglesworth, porównując *Mapy Oceniania* dla tych samych egzaminatorów utworzone na podstawie wyników egzaminu pilotażowego oraz egzaminu żywego. Rys. 2. przedstawia *Mapę Oceniania* egzaminu żywego dla tego samego egzaminatora, którego *Mapę Oceniania* dla egzaminu pilotażowego przedstawiono na rys. 1.

[20] Wigglesworth G., op. cit.



Rys. 2. Przykładowa *Mapa Oceniania* dla zadań w egzaminie półpośrednim dla wybranego egzaminatora po egzaminie żywym[21]

Różnica w poziomie *nastawienia* egzaminatora jest wyraźna. Wyniki przyznawane przez niego w każdym kryterium nie odbiegają od tego, co przewiduje model. Nie ma znaczących różnic w poziomie surowości, z jakim egzaminator ocenia wypowiedzi pod względem poprawności gramatycznej oraz płynności. Wigglesworth wysuwa zatem wniosek, iż wysokiej jakości informacja zwrotna może wyraźnie wpłynąć na sposób, w jaki egzaminatorzy oceniają egzamin, podnosząc znacznie rzetelność testu oraz trafność interpretacji otrzymywanych przez uczniów wyników. Analiza *nastawienia*, konkluduje Wigglesworth, *pozwala trenerom stworzyć profil każdego egzaminatora poprzez uzyskiwanie ciągłej informacji o jego sposobie oceniania*[22].

Zbierana na przestrzeni kilku sesji egzaminacyjnych informacja na temat egzaminatora może pomóc trenerom i organizatorom egzaminu dokładnie określić indywidualne potrzeby każdego z egzaminatorów, co przyczyni się do przekazania mu jak najbardziej wartościowej informacji zwrotnej. Wartościowej, ponieważ nie ogólnej i ogólnikowej, a zindywidualizowanej i konkretnej; opartej nie na trenerskim „widzimisię", ogólnych statystykach i subiektywnych obserwacjach, ale na rzetelnej analizie ilościowej i jakościowej[23].

Podsumowanie

Analiza wyników egzaminów performancji (praktycznych) z wykorzystaniem programu *FACETS* niewątpliwie dostarcza nam wielu informacji na temat

[21] Ibidem.
[22] Ibidem, s. 318; tłum. moje.
[23] T. McNamara (*Measuring...*, op. cit., s. 148) przytacza wyniki późniejszego badania przeprowadzonego przez Lunt, Morton i Wigglesworth, które były już mniej imponujące. Zastosowana przez Wigglesworth metoda nie przyniosła porównywalnych skutków w drugim badaniu.

wiedzy (umiejętności) kandydatów, rzetelności zastosowanego narzędzia pomiaru i schematu oceniania, ale również – a może przede wszystkim – na temat tego aspektu egzaminów performancji (praktycznych), który, jak twierdzi Lumley, *znajduje się w centrum całego procesu*[24] – czyli na temat egzaminatora. Uzyskane w wyniku analizy dane, odpowiednio przetworzone i dostosowane do poziomu wiedzy niespecjalistów, mogą stanowić niezwykle cenną informację zwrotną nie tylko dla organizatora egzaminu, ale również dla samych egzaminatorów, szczególnie tych, którym zależy na ciągłym kształceniu i podnoszeniu jakości swojej pracy. Wartościowa informacja zwrotna, podana w przystępny i zrozumiały sposób, może wpłynąć i – jak pokazują wyniki opisanych powyżej badań – wpływa na podniesienie jakości pracy egzaminatorów poprzez zwiększenie stałości, a więc także porównywalności oceniania. Wykorzystanie w procesie szkolenia egzaminatorów najnowszych dostępnych metod analiz może zdecydowanie ułatwić ten proces.

Zanim jednak wieloaspektowy model Rascha stanie się powszechnie stosowanym narzędziem wśród osób zaangażowanych w szkolenie egzaminatorów, minie jeszcze na pewno wiele czasu. Czas ten będzie wypełniony eksperymentami i próbami, nakierowanymi na lepsze zrozumienie samego narzędzia oraz możliwości jego wykorzystania. Na razie więc – aby odpowiedzieć na pytanie postawione w tytule tego artykułu – wykorzystanie *MFRM* w szkoleniu egzaminatorów to chyba jednak bardziej *science-fiction* niż rzeczywistość, choć ta sytuacja może nie potrwać już długo.

[24] Lumley T., op. cit., s. 267.