dr Tomasz Żółtak
Diagmatic

Wyniki IRT w zasięgu każdego diagnosty. Nowe możliwości obliczania i prezentacji wyników diagnozy w testach przeprowadzanych z wykorzystaniem tabletów i komputerów

Abstrakt

Przy konstruowaniu narzędzi diagnostycznych bardzo wiele uwagi poświęca się ich podstawom teoretycznym oraz własnościom pomiarowym. W cieniu tych, niewątpliwie bardzo istotnych, kwestii pozostaje dopracowywanie metod wykorzystania narzędzia przez diagnostę w taki sposób, aby było to dla niego proste, a uzyskiwane wyniki diagnozy jak najbardziej zrozumiałe i użyteczne. Wynika to oczywiście w dużej mierze z utrwalenia się standardów postępowania, związanych z obliczaniem wyników diagnozy (obliczanie „wyniku surowego" i przeliczanie go na normy staninowe lub centylowe). Należy jednak zwrócić uwagę, że standardy te powstały w czasach, kiedy jedyną dostępną formą prowadzenia testów były papierowe kwestionariusze, a zasób narzędzi wspomagających obliczanie wyników przez diagnostę ograniczał się do klucza odpowiedzi, tabeli przeliczeniowej, ołówka i ewentualnie kalkulatora. Tymczasem coraz szersze wykorzystanie do diagnozy edukacyjnej i psychologicznej narzędzi komputerowych daje w tym aspekcie zupełnie nowe możliwości, które pozwalają zarówno zwiększyć trafność i precyzję wyników, jak i uczynić ich obliczanie i dalszą analizę bez porównania prostszymi. W wystąpieniu zostaną one omówione na przykładzie rozwiązania informatycznego oraz testów do diagnozy motywacji i amotywacji, a także funkcjonowania społecznego uczniów, stworzonych przez firmę Diagmatic. W szczególności zaprezentowane zostanie wykorzystanie automatycznego obliczania wyników diagnozy przy pomocy estymatorów EAP, na podstawie wielowymiarowych modeli czynnikowych oraz metody graficznej prezentacji uzyskanych wyników.

Słowa kluczowe: testowanie komputerowe, IRT, modele wielowymiarowe, oszacowania EAP, normalizacja ekwikwantylową, graficzna prezentacja wyników diagnozy

Wprowadzenie

W ostatnich latach można zauważyć w Polsce pojawienie się narzędzi komputerowych (nazywanych też *testami w formie multimedialnej*) do celów diagnozy edukacyjnej i psychologicznej dzieci i młodzieży. Warto odnotować choćby stworzony w 2012 r. na zlecenie Instytutu Badań Edukacyjnych przez zespół Macieja Karwowskiego Test Umiejętności na Starcie Szkolnym (TUNSS) (Kaczan i Rycielski, 2012; Karwowski i Dziedziewicz, 2012) – chyba pierwszy

krajowy test w dziedzinie diagnostyki edukacyjnej przeprowadzany przy pomocy tabletów, który dodatkowo miał postać testu adaptatywnego[1] i wykorzystywał zaawansowane metody statystyczne (dokładnie: estymator największej wiarygodności) do obliczania wyników pomiaru. Z rozwiązań, których wykorzystanie promowane było w ostatnich latach znacznie szerzej, należy wymienić zestaw testów używanych w ramach narzędzia do diagnozy funkcjonalnej rozwoju społeczno-emocjonalnego uczniów w wieku 9–13 lat TROS-KA, przygotowany w latach 2016–2017 na zlecenie Ośrodka Rozwoju Edukacji przez zespół pod kierownictwem Ewy Domagały-Zyśk, Tomasza Knopika i Urszuli Ozwy (2017). Testy te dostępne są zarówno w formie papierowej, jak i jako aplikacja komputerowa. Rośnie też liczba możliwych do wykorzystania przez szkoły testów komputerowych, głównie w obszarze diagnozy umiejętności uczniów, ale również szerzej pojętych uzdolnień i funkcjonowania społecznego, tworzonych przez wydawców podręczników i materiałów dydaktycznych oraz dostawców dzienników elektronicznych. Można więc ocenić, że zakres wykorzystania narzędzi komputerowych do diagnozy edukacyjnej i psychologicznej w Polsce dynamicznie się rozszerza i trend ten będzie się w kolejnych latach utrzymywał.

W tym kontekście warto rozpatrzeć, jakie są nowe możliwości związane z wykorzystaniem narzędzi komputerowych w tym obszarze, ale także, czy w dostępnych dotychczas rozwiązaniach były one w pełni wykorzystane. W szczególności dotyczy to metod obliczania wyników pomiaru, jako że większość przygotowywanych narzędzi ogranicza się tu do implementacji dokładnie tych samych procedur, które stosowane są od dziesięcioleci w testach papierowych, tj. wykorzystania tablic z normami pozwalającymi przeliczyć prostą sumę punktów na rangi stenowe, staninowe lub centylowe.

W dalszej części tekstu najpierw ogólnie omówione zostały zalety i ograniczenia wykorzystania narzędzi komputerowych do diagnozy edukacyjnej i psychologicznej. W kolejnej części skoncentrowałem się na kwestii obliczania wyników pomiaru. Opisuję w niej ograniczenia dotychczas stosowanego podejścia oraz nowe możliwości, jakie daje w tym zakresie wykorzystanie aplikacji komputerowych. Następnie prezentuję implementację omówionych wcześniej rozwiązań dotyczących rozwiązania informatycznego przygotowanego przez firmę Diagmatic w ramach szerszego projektu, w wyniku którego opracowane zostały nowe narzędzia do diagnozy motywacji i amotywacji oraz funkcjonowania społecznego uczniów (por. Wysocka, 2020). Tekst kończy krótkie podsumowanie.

Zalety i ograniczenia diagnozy z wykorzystaniem narzędzi komputerowych

W literaturze coraz częściej wskazuje się na zalety wykorzystania narzędzi komputerowych do diagnozy edukacyjnej i psychologicznej. Pozwalają one na wykorzystanie bardziej zróżnicowanych formatów zadań/pytań, w tym zawierających elementy multimedialne lub interaktywne, co ma szczególne

[1] Tj. testu, w którym kolejne pytania dobierane są dla danego rozwiązującego na bieżąco, na podstawie dotychczas udzielonych odpowiedzi, tak aby odpowiadały jego poziomowi umiejętności. W porównaniu z sytuacją, gdy wszyscy badani rozwiązują ten sam zestaw zadań, pozwala to zwiększyć precyzję pomiaru przy tej samej liczbie zadań, które rozwiązuje każdy zdający.

znaczenie w diagnozie młodszych dzieci, które nie opanowały jeszcze umiejętności czytania (Kaczan i Rycielski, 2012: 171). Dodatkowo testy w takiej formie są dla większości uczniów bardziej atrakcyjne, a tym samym pozwalają zredukować problemy związane z deficytem motywacji uczestników (Domagała-Zyśk, 2017: 125). Podstawową zaletą jest jednak ogromne uproszczenie procesu przetwarzania wyników testu – odpowiedzi uczniów są zapisywane w systemie informatycznym, w ramach którego mogą zostać w automatyczny sposób przetworzone, bez konieczności mozolnego sprawdzania papierowych prac z użyciem klucza odpowiedzi i *ręcznego* sumowania punktów. Upraszcza to prowadzenie badań pilotażowych i standaryzacyjnych, a w przypadku wykorzystania już opracowanych narzędzi do diagnozy indywidualnej oznacza, że wynik pomiaru może zostać zaprezentowany diagnoście już w chwilę po tym, gdy uczeń skończył wypełniać test, bez konieczności samodzielnego dokonywania przez diagnostę jakichkolwiek obliczeń. Możliwe jest też tworzenie rozwiązań informatycznych wspomagających diagnostę w analizie uzyskanych wyników poprzez tworzenie ich porównań i wykresów oraz zestawianie ich w raporty.

Ograniczenia diagnozy przy użyciu narzędzi komputerowych wiążą się przede wszystkim z konieczną do tego infrastrukturą, przy czym problem ten ma kilka różnych wymiarów. Najbardziej oczywistym jest zapewnienie w szkole lub poradni psychologiczno-pedagogicznej dostępu do odpowiedniej liczby urządzeń, na których test mają wypełniać diagnozowani. Dodatkową komplikację stanowi to, że w zależności od narzędzia do jego wypełnienia może być potrzebny komputer lub tablet. Tablety są co prawda tańsze i dzięki ekranom dotykowym pozwalają na łatwiejszą interakcję z urządzeniem (w szczególności stanowią bardziej odpowiednie medium dla młodszych dzieci), jednak w polskich szkołach są rzadko dostępne, ze względu na ograniczony zakres zadań, do których mogą być wykorzystane. Z drugiej strony komputery są szeroko dostępne, ale nie udostępniają takiego poziomu interaktywności jak tablety.

Drugi problem to kwestia interoperacyjności sprzętu dostępnego w danej szkole lub poradni z oprogramowaniem wykorzystywanym do testowania, tj. zapewnienia, że da się to oprogramowanie uruchomić na konkretnych komputerach lub tabletach. Jest to temat złożony, a jednocześnie specjalistyczny, nie będę więc go tutaj rozwijał. Jeszcze do niedawna był to jednak poważny problem, zwłaszcza ze względu na wiek sprzętu komputerowego wykorzystywanego w niektórych szkołach. Obecnie jego znaczenie na szczęście maleje, wobec upowszechniania się dostępu do szybkich łącz internetowych – umożliwia to tworzenie aplikacji działających na zewnętrznych serwerach, z którymi osoba testowana łączy się za pośrednictwem interfejsu WWW, wykorzystując dowolną przeglądarkę internetową (konieczne jest przy tym oczywiście zapewnienie systemu gwarantującego bezpieczne zarządzanie kontami użytkowników i dostępem do nich). Rozwiązania informatyczne stworzone w takiej architekturze (tzw. klient–serwer) mogą być wykorzystywane na niemal dowolnych urządzeniach, bez względu na ich konfigurację (jednakże pod warunkiem posiadania dostępu do internetu).

Trzeci problem dotyczy kwestii zapewniania wsparcia i dalszego rozwoju wytworzonych rozwiązań informatycznych. Znamiennym, a przy tym smutnym, przykładem jest tu wspomniany na początku tekstu TUNSS – test przygotowany

w formie samodzielnie działającej aplikacji na tablet z systemem Android. W wyniku jednej z kolejnych aktualizacji tego systemu operacyjnego pewne rozwiązania informatyczne wykorzystane w aplikacji TUNSS przestały działać, a w rezultacie niemożliwe stało się jej wykorzystanie na tabletach ze zaktualizowanym systemem. Aby rozwiązać problem, konieczne byłoby niewielkie zmodyfikowanie kodu aplikacji, ponieważ jednak projekt, w ramach którego przygotowano test, został zakończony, nie znalazły się na to środki finansowe. W konsekwencji niemożliwe stało się korzystanie z TUNSS[2]. Problem ten ma szczególne znaczenie w przypadku narzędzi, których rozwój finansowany jest – jak to często bywa na przestrzeni ostatniej dekady – ze środków europejskich przez instytucje publiczne, które zakładają późniejsze udostępnianie ich całkowicie za darmo. W przypadku wspomnianych wcześniej coraz częściej wykorzystywanych rozwiązań informatycznych działających w architekturze klient–serwer dodatkową komplikację stanowi konieczność ponoszenia kosztów utrzymania działania serwera. W takim przypadku twórcy narzędzia diagnostycznego udostępniają je użytkownikom nie w formie *produktu* (tj. czegoś, w czego posiadanie można wejść i następnie używać wielokrotnie), lecz jako *usługę* (która musi zostać każdorazowo dostarczona z udziałem twórców czy też zarządzających danym narzędziem). W związku z tym konieczne jest stosowanie przy tworzeniu narzędzi diagnostycznych w formie komputerowej odpowiednich modeli finansowania, uwzględniających ww. problemy.

Obliczanie wyników pomiaru – niewykorzystane dotąd możliwości

Normalizacja ekwikwantylową i jej ograniczenia

W odniesieniu do testów przygotowywanych w formie papierowej, stosowanych do celów diagnozy indywidualnej, do obliczania wyników pomiaru powszechnie wykorzystywana jest procedura oparta na metodzie normalizacji ekwikwantylowej sumy punktów (Brzeziński, 2019: 391:408; Żółtak, 2015: 21–22). W podejściu tym każda z odpowiedzi na zadanie/pytanie testu ma przypisaną określoną liczbę punktów: wynik pomiaru (określany niekiedy *wynikiem surowym*) w danym wymiarze oblicza się, sumując punktację odpowiadającą odpowiedziom udzielonym przez badanego na wszystkie zadania/pytania przypisane do danego wymiaru. Na podstawie wyników badania standaryzacyjnego na próbie losowej dobranej z referencyjnej populacji (najczęściej stosowane jest kryterium doboru osób określonej płci i w określonym – niezbyt zróżnicowanym – wieku) ustalany jest rozkład częstości takiej sumy punktów, który następnie porównywany jest z zakładanym rozkładem teoretycznym (co do zasady jest to rozkład normalny o zadanych parametrach). Na podstawie takich porównań ustalane jest przeliczenie sumy punktów na rangi centylowe – pokazujące, jak wiele osób w referencyjnej populacji osiągnęło wynik nie wyższy niż dany – albo na wynik wyrażony na jednej ze skal standardowych. W Polsce najczęściej jest to skala stenowa w psychologii lub staninowa w pomiarze edukacyjnym. Przeliczenia takie zapisywane są w formie tabel, z których mogą potem korzystać diagności.

[2] Chyba żeby specjalnie w tym celu utrzymywać na stanie instytucji tablety z odpowiednio starą wersją systemu operacyjnego.

Metoda ta ma pewne zalety. Przede wszystkim jest dosyć łatwa do zastosowania, zarówno przy określaniu postaci przekształcenia sumy punktów w wynik na skali standardowej na etapie badań standaryzacyjnych, jak i przy późniejszym stosowaniu przez diagnostów. Dla tych ostatnich relatywnie największym problemem jest występowanie w testach psychologicznych tzw. pytań odwróconych, tj. takich, w których te same odpowiedzi, które występują też w innych pytaniach, mają przypisaną inną (odwróconą) punktację. Możliwe jest też bardzo łatwe obliczanie wyniku łącznego w przypadku testu, który składa się z kilku powiązanych ze sobą wymiarów – wystarczy zastosować analogiczną procedurę w odniesieniu do sumy *wyników surowych* każdego z wymiarów.

Stosowanie normalizacji ekwikwantylowej jest jednak z kilku powodów problematyczne. Przede wszystkim należy zwrócić uwagę, że założenia tej metody nie korespondują z założeniami metod powszechnie wykorzystywanych do analizy trafności konstruowanych narzędzi. Spośród tych drugich podstawową jest analiza zróżnicowania siły związku pomiędzy odpowiedziami na poszczególne zadania/pytania a mierzoną cechą. Wykorzystuje się w tym celu albo wartości parametrów mocy różnicującej (KTT), albo parametry ładunków modelu czynnikowego (lub – równoważnie – parametry dyskryminacji modelu IRT). Analiza taka ujawnia zwykle istotne zróżnicowanie siły związków, które pozostaje (choć jest oczywiście redukowane) nawet pomimo usunięcia z testu pozycji szczególnie słabo powiązanych z mierzoną cechą w procesie dopracowywania struktury narzędzia diagnostycznego. Wykorzystanie prostej sumy punktów jako statystyki opisującej wynik pomiaru ignoruje istnienie takich różnic[3] (sprawia, że każde zadanie/pytanie *liczy się tak samo*), a tym samym przyczynia się do zmniejszenia trafności pomiaru. Aby lepiej mierzyć poziom interesującej nas cechy, należałoby brać pod uwagę, że niektóre zadania/pytania mierzą ją *lepiej*, a inne nieco *gorzej*.

Jeśli przygotowywane narzędzie mierzy kilka powiązanych ze sobą cech (określanych często w takim przypadku terminem *wymiary*), możliwe jest zwiększenie precyzji pomiaru każdej z nich poprzez uwzględnienie informacji o natężeniu pozostałych (oraz sile związków pomiędzy nimi). Normalizacja ekwikwantylowa, przeprowadzana oddzielnie dla wyników na każdym wymiarze, nie daje jednak takich możliwości.

Z drugiej strony, choć przy normalizacji ekwikwantylowej zwykle stosuje się oddzielne ustalenie norm dla kilku różnych grup badanych (np. chłopców i dziewcząt), to podczas analiz trafności narzędzia prowadzonych przy pomocy analizy czynnikowej (lub parametrów KTT zadań/pytań) z reguły nie uwzględnia się w żaden sposób podziału na te grupy. W tym kontekście należałoby postulować wykorzystania do analiz trafności odpowiednio wyspecyfikowanych wielogrupowych modeli czynnikowych (por. Kondratek i in., 2015: 66–69; Żółtak, 2015: 32), a następnie użycie tych modeli – z uwzględnieniem całej ich złożoności – do obliczania wyników pomiaru (szacowania natężenia mierzonych cech).

[3] Jest rzeczą interesującą, a mało znaną, że tę samą właściwość ma współczynnik alfa Cronbacha, powszechnie wykorzystywany do szacowania rzetelności testów. Jak wskazują badania, w sytuacji kiedy występuje zróżnicowanie siły związku pomiędzy mierzoną cechą a zadaniami/pytaniami, alfa Cronbacha będzie zaniżać oszacowania rzetelności (Green i Yang, 2009).

Alternatywne metody szacowania wyników pomiaru

Aby rozwiązać opisane problemy, konieczna jest zmiana metody obliczania wyników pomiaru. Spośród czterech najczęściej stosowanych (Embretson i Reise, 2000; Warm, 1989; por. Kondratek i Pokropek, 2015): ML (największej wiarygodności), WML (ważonej największej wiarygodności), MAP (wartości modalnej rozkładu *a posteriori*) i EAP (wartości oczekiwanej rozkładu *a posteriori*) na szczególną uwagę zasługują dwie ostatnie, gdyż w przypadku modeli wielowymiarowych pozwalają wykorzystywać przy szacowaniu poziomu natężenia mierzonych cech informacje o sile powiązań pomiędzy poszczególnymi wymiarami. Metody ML i WML, podobnie jak normalizacja ekwikwantylowa, muszą zaś być stosowane oddzielnie dla każdego z wymiarów.

Aby obliczyć wyniki pomiaru, korzystając z metody MAP lub EAP, konieczne jest wcześniejsze wyestymowanie modelu czynnikowego – czy to w formie konfirmacyjnej analizy czynnikowej na podstawie macierzy korelacji polichorycznych, czy to w formie modelu IRT – na próbie standaryzacyjnej. W zależności od struktury konstruowanego narzędzia diagnostycznego model taki może być dodatkowo wielowymiarowy (jeśli kilka powiązanych ze sobą cech jest przy pomocy danego narzędzia diagnostycznego badanych jednocześnie) i wielogrupowy (jeśli zakłada się tworzenie norm odrębnie dla kilku różnych grup). Należy zauważyć, że modele takie zwykle estymuje się w toku prac nad rozwojem narzędzia diagnostycznego, aby sprawdzić, czy jego struktura odpowiada założonej wcześniej strukturze teoretycznej[4]. Kiedy model taki jest już wyestymowany, jego parametry mogą zostać wykorzystane do oszacowania poziomu natężenia mierzonych cech dla poszczególnych badanych.

Zarówno metoda MAP, jak i EAP opierają się na rekonstruowaniu parametrów tzw. rozkładów *a posteriori* opisujących rozkład mierzonej cechy wśród hipotetycznej zbiorowości osób, które udzieliły określonego profilu odpowiedzi (tj. o określonych odpowiedziach na wszystkie zadania/pytania testu). Inaczej mówiąc, dąży się w nich do określenia, jak często – zgodnie z założeniami wyestymowanego wcześniej modelu czynnikowego – wśród wszystkich osób o danym profilu odpowiedzi powinny występować osoby o poszczególnych wartościach mierzonej cechy (por. Kondartek i Pokropek, 2015: 26). Następnie, w zależności od metody, szacuje się wartość modalną, tj. najczęściej występującą (MAP) lub oczekiwaną, czyli średnią (EAP) takiego rozkładu i przyjmuje ją jako oszacowanie natężenia danej cechy, które zostanie przypisane wszystkim osobom o takim profilu odpowiedzi. Oszacowania uzyskiwane oboma metodami zwykle nie różnią się od siebie znacząco[5].

Choć obie metody łączy odwołanie się przy szacowaniu mierzonych cech do własności rozkładów *a posteriori*, jednak różnią się one diametralnie, jeśli chodzi o podejście obliczeniowe. W metodzie EAP dokonuje się *rekonstrukcji* danego rozkładu *a posteriori*, aby móc następnie obliczyć jego wartość

[4] Choć w polskich warunkach często wykorzystuje się w tym celu mniej skomplikowane warianty wspomnianych wyżej metod analizy: estymację modelu czynnikowego na podstawie macierzy korelacji liniowych, brak wykorzystania wielowymiarowych modeli konfirmacyjnych, niestosowanie modeli wielogrupowych.

[5] Jako że omawiane rozkłady *a posteriori* są jednomodalne, a dla większości możliwych profilów odpowiedzi również niemal symetryczne.

oczekiwaną (i ew. inne parametry). Wymaga to zastosowania całkowania numerycznego, tj. obliczenia prawdopodobieństwa uzyskania danego profilu odpowiedzi dla wybranych kombinacji wartości mierzonych cech. Oprócz tego konieczne jest jeszcze użycie informacji o zakładanym prawdopodobieństwie[6] wystąpienia tych kombinacji mierzonych wartości w odpowiedniej grupie badanych. Wszystkie te wielkości można obliczyć na podstawie parametrów modelu czynnikowego, opisujących odpowiednio zadania/pytania (i ich powiązania z mierzonymi cechami) oraz wartości oczekiwane, a także wariancje i kowariancje mierzonych cech (Bock i Myslevy, 1982). Ważną cechą metody EAP jest to, że w sposób analogicznych do wartości oczekiwanej może zostać obliczone również odchylenie standardowe rozkładu *a posteriori*, które obrazuje błąd pomiaru oszacowania EAP[7]. Metoda MAP z kolei opiera się na wykorzystaniu technik optymalizacyjnych: na podstawie struktury modelu, metodami analitycznymi wyprowadzany jest wzór opisujący przebieg funkcji gęstości rozkładu *a posteriori* dla danego profilu odpowiedzi. Następnie poszukuje się takiej kombinacji wartości mierzonych cech, które maksymalizują wartość tej funkcji. Błąd pomiaru można dla estymatora MAP oszacować na podstawie wartości funkcji informacyjnej testu[8].

Aby otrzymać wyniki w formie analogicznej do tej, w jakiej są one uzyskiwane w wyniku zastosowania normalizacji ekwikwantylowej, należy jeszcze wystandaryzować oszacowania MAP lub EAP, wykorzystując wartości parametrów wartości oczekiwanej i odchylenia standardowego danej cechy w odpowiedniej grupie badanych. Ze względu na to, że oszacowania MAP i EAP charakteryzują się tzw. ściągnięciem w kierunku zera (tj. odchylenie standardowe tych oszacowań jest mniejsze niż wartość parametru modelu czynnikowego opisującego odchylenie standardowe mierzonej cechy), zamiast wartości parametru modelu wykorzystuje się przy tym niekiedy odchylenie standardowe oszacowań, obliczonych dla wszystkich badanych z danej grupy w badaniu standaryzacyjnym (por. Żółtak, 2015:27). Tak przekształcony wynik można traktować jako wyrażony na skali standardowej o średniej 0 i odchyleniu standardowym 1 oraz w łatwy sposób (przy pomocy przekształcenia liniowego i ew. zaokrągleń do liczb całkowitych) przeliczyć na wartość wyrażoną na dowolnej innej skali standardowej (np. stenowej, staninowej lub tenowej) albo – używając dystrybuanty rozkładu normalnego standaryzowanego – na rangę centylową.

Niewykorzystane możliwości testowania komputerowego

Żadna ze wspomnianych metod, wymagających odwołania się do wartości licznych parametrów modelu czynnikowego i wykonania skomplikowanych obliczeń, w oczywisty sposób nie mogła znaleźć zastosowania w samodzielnym

[6] Ściśle: wartości funkcji gęstości, gdyż rozkład mierzonych cech opisywany jest w modelu czynnikowym jako ciągła, a nie dyskretna zmienna losowa.

[7] W dużym uproszczeniu można myśleć o odchyleniu standardowym rozkładu *a posteriori* jako o analogu standardowego błędu pomiaru (SEM) w KTT. Trzeba mieć jednak na uwadze, że pomiędzy tymi dwoma zachodzi bardzo istotna różnica: o i ile SEM jest z założenia taki sam dla wszystkich badanych, o tyle odchylenie standardowe rozkładu *a posteriori* będzie różnić się dla osób o różnych profilach odpowiedzi.

[8] Należy przy tym zwrócić uwagę, że przy szacowaniu tego błędu nie odwołuje się do informacji o własnościach rozkładu *a posteriori* mierzonej cechy pod warunkiem uzyskania danego profilu odpowiedzi.

obliczaniu wyniku pomiaru przez diagnostę. Jednakże w sytuacji, kiedy obliczanie wyników implementowane jest w ramach aplikacji komputerowej, ich wykorzystanie staje się jak najbardziej możliwe. Co więcej, choć nawet pobieżny opis koniecznych do przeprowadzenia obliczeń może sprawiać wrażenie bardzo dużego stopnia komplikacji, to z użyciem współczesnych komputerów mogą być one bez problemu wykonane w czasie rzeczywistym lub – w przypadku wysoce złożonych modeli – zbliżonym do rzeczywistego, nie wymagając przy tym żadnych działań ze strony użytkownika aplikacji (diagnosty). Warto więc zadać pytanie, dlaczego metody te nie były dotychczas wykorzystywane.

Być może najważniejszy powód to utrwalenie się w polskiej pedagogice i psychologii standardu postępowania przy obliczaniu wyników testu, obejmującego właśnie wykorzystanie normalizacji ekwikwantylowej (por. Niemierko, 1999). Standard ten jest tym silniej przyjęty, że nauczanie bardziej zaawansowanej psychometrii, wykraczającej poza podstawy KTT, jest na polskich uczelniach raczej nieobecne. W konsekwencji w środowisku – zarówno diagnostów, jak i badaczy – brakuje świadomości co do ograniczeń stosowanej metody oraz dostępnych alternatyw. Warto przy tym zwrócić uwagę, że nawet w przypadku narzędzi diagnostycznych, które opracowywane są w Polsce w ramach dużych projektów przez zespoły badaczy o uznanej renomie, zakres analiz psychometrycznych przeprowadzanych w celu weryfikacji poprawności założonej struktury narzędzia i jego trafności jest często nader skromny. Spośród dwóch przygotowanych w ostatnich latach dużych narzędzi do diagnozy funkcjonowania społecznego i emocjonalnego dzieci: przeznaczonego dla młodszych dzieci, które opracowane zostało w IBE[9] (Czub, 2014) oraz – wspomnianego już wcześniej – narzędzia TROS-KA, opracowanego na zlecenie ORE (Domagała-Zyśk i in., 2017), w żadnym nie przeprowadzono analizy struktury diagnozowanych cech z wykorzystaniem wielowymiarowego modelu czynnikowego, ani nie stosowano modeli wielogrupowych.

Druga kwestia, która utrudnia wdrożenie nowych metod obliczania wyników w ramach komputerowych narzędzi do diagnozy, ma charakter technologiczny. W językach programowania, przy pomocy których tworzona jest większość aplikacji do diagnozy – PHP, Python czy Java – brak jest gotowych bibliotek, które dawałyby możliwość łatwego wykorzystania odpowiednich metod statystycznych. Metody te zostały co prawda zaimplementowane w zewnętrznych programach (np. Mplus) lub bibliotekach specjalistycznych języków programowania, dedykowanych do zadań związanych z analizą danych i statystyką, jak np. R, co jednak prowadzi do problemów z ich integracją. Twórcy systemów informatycznych mających obsługiwać narzędzia do diagnozy stoją więc obecnie przed wyborem pomiędzy koniecznością samodzielnej implementacji odpowiednich metod statystycznych w już wykorzystywanym środowisku technologicznym a zwiększaniem skomplikowania struktury tworzonego systemu informatycznego pociągającym za sobą konieczność utrzymywania wielu serwisów działających w różnych technologiach oraz interfejsów wymiany danych pomiędzy nimi. Choć są to poważne wyzwania, rozwiązanie opisane poniżej pokazuje, że z konfrontacji z nimi można wyjść zwycięsko.

[9] Narzędzie przygotowane zostało wyłącznie w klasycznej, papierowej formie, przy czym badanie przebiega poprzez przeprowadzenie przez diagnostę wywiadu z rodzicem dziecka (a nie z samym dzieckiem).

Nowe narzędzia

Implementacja nowych metod obliczania wyników pomiaru w ramach tworzonych aplikacji komputerowych przyjęta została jako jeden z celów projektu pt. „Opracowanie inteligentnych narzędzi do diagnozy psycho-społeczno-edukacyjnej dzieci i młodzieży", prowadzonego przez spółkę Diagmatic, a współfinansowanego przez Narodowe Centrum Nauki i Rozwoju ze środków Programu Operacyjnego Inteligentny Rozwój[10]. Założenia teoretyczne i struktura narzędzi diagnostycznych opracowanych i przebadanych w ramach tego projektu opisane zostały w odrębnym tekście (Wysocka, 2020), dlatego dalej skupię się wyłącznie na przedstawieniu, w jaki sposób w przygotowanym rozwiązaniu wykorzystane zostały nowe metody obliczania i prezentacji wyników pomiaru użytkownikom aplikacji. Sekcja dotycząca obliczania wyników pomiaru będzie interesująca zapewne głównie dla bardziej zaawansowanych badaczy, mających dobrze ugruntowaną wiedzę psychometryczną. Typowy użytkownik opracowanych narzędzi diagnostycznych (tj. diagnosta) nie styka się z opisanymi tam funkcjonalnościami – albo są one wykorzystywane jedynie przez autorów narzędzia diagnostycznego do *zdefiniowania go* w systemie, albo (samo obliczanie wyników na podstawie wyników badania konkretnej osoby) dokonują się całkowicie automatycznie. Następujący potem opis sposobu prezentacji wyników powinien zaś być przystępny dla wszystkich, nawet początkujących, diagnostów edukacyjnych i psychologicznych.

Obliczanie wyników pomiaru

W ramach przygotowanego rozwiązania zdecydowano się na samodzielną implementację obliczania oszacowań metodą EAP na podstawie wielowymiarowych, wielogrupowych modeli czynnikowych. Co istotne, wykorzystywane modele czynnikowe mogą mieć bardzo złożoną strukturę, w tym obejmować tzw. czynniki wyższego rzędu (por. Humenny i Grygiel, 2015: 139–142). W modelach takich wyniki pomiaru obliczane są zarówno dla poszczególnych wymiarów *niższego rzędu*, jak i dla czynnika *wyższego rzędu*, który stanowi analog wyniku łącznego (tj. obliczonego na podstawie sumy punktów z wszystkich wymiarów) w metodzie normalizacji ekwikwantylowej. Dopuszczalne jest też wykorzystanie modeli, w których występują tzw. ładunki krzyżowe, czyli powiązanie tego samego zadania/pytania z więcej niż jednym wymiarem (mierzoną cechą). W przypadku modeli wielogrupowych system nie nakłada żadnych ograniczeń na konieczność zachowywania przez nie inwariancji (niezmienniczości) pomiarowej (por. Grygiel i in., 2015; Kondratek i in., 2015), choć należy mieć na uwadze, że występowanie zróżnicowanego funkcjonowania pozycji testowych może skutkować koniecznością nieco innego interpretowania tych samych wymiarów (mierzonych cech) w różnych grupach.

Estymacja parametrów modelu czynnikowego (na podstawie wyników badania standaryzacyjnego) odbywa się w zewnętrznej aplikacji (dowolnie wybranej przez badacza), przy czym, aby model taki był kompatybilny z zaimplementowanym w systemie sposobem obliczania wyników pomiaru, musi on być estymowany na jeden z dwóch (formalnie równoważnych) sposobów:

[10] Sygnatura projektu: POIR.01.01.01-00-0402/18.

- jako model czynnikowy dla kategorialnych zmiennych obserwowanych, estymowany na podstawie macierzy korelacji tetra-/polichorycznych (konieczne jest również oszacowanie parametrów progów na etapie obliczania macierzy korelacji) z prezentacją wyników w tzw. parametryzacji *theta*[11];
- jako model IRT: dwuparametryczny lub trzyparametryczny (z dolną asymptotą) z probitową funkcją łączącą dla zmiennych binarnych lub model *graded response* z probitową funkcją łączącą dla zmiennych przyjmujących więcej niż dwie wartości[12] (patrz: Kondratek i Pokropek, 2015).

W ramach będącego częścią systemu informatycznego modułu do ankietowania możliwe jest określenie schematów punktowania odpowiedzi, w tym zarówno typowych dla pytań samoopisowych, w których każdej odpowiedzi przypisana jest inna wartość liczbowa (np. *zdecydowanie nie* – 0, *raczej nie* – 1, *raczej tak* – 2, *zdecydowanie tak* – 4), jak i charakterystycznych raczej dla pytań o wiedzę, kiedy tylko jedna odpowiedź traktowana jest jako *poprawna* (wartość 1), a wszystkie pozostałe jako *niepoprawne* (wartość 0). W odniesieniu do tych ostatnich pytań, jak zostało to już zasygnalizowane, możliwe jest wykorzystanie modelu psychometrycznego uwzględniającego parametr dolnej asymptoty, powszechnie znany jako *parametr zgadywania*.

Przygotowany system daje więc autorom narzędzi diagnostycznych bardzo daleko posuniętą swobodę, jeśli chodzi o formę (strukturę) wykorzystywanych modeli psychometrycznych. Wyniki estymacji modelu, zapisane w formie pliku CSV o ustalonej strukturze[13], importowane są do systemu informatycznego, który na jego podstawie odwzorowuje strukturę modelu (weryfikowana jest przy tym jej poprawność) oraz zapisuje w bazie danych wartości odpowiednich parametrów psychometrycznych. Dodatkowo do systemu wczytywane są też wartości tzw. *wariancji empirycznych*, opisujących zróżnicowanie oszacowań EAP w poszczególnych grupach badanych – są one konieczne do przeliczenia wyników pomiaru w taki sposób, aby usunąć sygnalizowane we wcześniejszej części tekstu ściągnięcie estymatora EAP. Przed udostępnieniem narzędzia diagnostycznego użytkownikom konieczne jest jeszcze uzupełnienie w systemie przez autorów narzędzia opisów zawierających interpretacje poszczególnych wymiarów (mierzonych cech) oraz opisy diagnostyczne odpowiadające wybranym zakresom ich natężenia (patrz: część poświęcona prezentacji wyników pomiaru).

Obliczanie wyników pomiaru dla konkretnego badanego odbywa się w czasie rzeczywistym, bezpośrednio po zakończeniu wypełniania przez niego kwestionariusza narzędzia. Najpierw na podstawie informacji zapisanych w systemie rekonstruowana jest macierz kowariancji pomiędzy poszczególnymi wymiarami (cechami mierzonymi w ramach danego narzędzia diagnostycznego). W kolejnym kroku na podstawie tej macierzy generowany jest zestaw kombinacji

[11] Np. w programie Mplus wymaga to zdefiniowania zmiennych obserwowanych jako kategorialne, wykorzystaniu estymatora WLS, WLSM lub WLSMV oraz zadeklarowania użycia parametryzacji *theta*.

[12] Np. w programie Mplus odpowiada to wykorzystaniu estymatora MLR przy zdefiniowaniu zmiennych obserwowanych jako kategorialnych i ustawieniu funkcji łączącej – opcja LINK – na *probit*.

[13] Ze względu na preferencje badaczy pracujących w projekcie odpowiada ona strukturze pliku, w jakim można zapisać wyniki estymacji przeprowadzonej w programie Mplus, przetworzone przy pomocy pakietu *MplusAutomation* języka programowania R.

wartości poszczególnych wymiarów, które zostaną użyte do przeprowadzenia całkowania numerycznego. Jeśli liczba wymiarów nie przekracza sześciu[14], kombinacje wartości wyznaczane są z wykorzystaniem regularnej siatki rozpiętej w przybliżeniu pomiędzy wartościami ±4 razy odchylenie standardowe danego wymiaru od wartości oczekiwanej danego wymiaru w danej grupie badanych (do której powinien zostać odniesiony wynik właśnie diagnozowanej osoby), z liczbą unikalnych wartości przypadających na wymiar zmniejszającą się od 41 w przypadku jednego wymiaru do 5 w przypadku sześciu wymiarów (co odpowiada wygenerowaniu od 41 do ponad 15 tys. różnych kombinacji wartości wymiarów). Jeśli liczba wymiarów przekracza sześć, stosowana jest tzw. procedura Monte Carlo, w ramach której losowanych jest 10 tys. różnych kombinacji wartości wymiarów spod wielowymiarowego rozkładu normalnego, w którym związki pomiędzy wymiarami opisywane są zrekonstruowaną w poprzednim kroku macierzą kowariancji oraz zapisanymi w bazie wartościami oczekiwanymi poszczególnych wymiarów w ramach danej grupy badanych (do której powinien zostać odniesiony wynik właśnie diagnozowanej osoby).

Następnie, zgodnie ze schematem opisanym we wcześniejszej części tekstu, dla każdej z wygenerowanych kombinacji wartości wymiarów, na podstawie informacji o wartościach parametrów zadań/pytań obliczane jest prawdopodobieństwo uzyskania takiego profilu odpowiedzi, jak profil właśnie badanej osoby. Przybliżenie częstości występowania danej kombinacji wartości wymiarów w rozkładzie *a posteriori* osiąga się poprzez pomnożenie obliczonych przed chwilą prawdopodobieństw przez wartości funkcji gęstości wielowymiarowego rozkładu normalnego[15] (o ustalonej wcześniej macierzy kowariancji pomiędzy wymiarami i wartościach oczekiwanych wymiarów) odpowiadające poszczególnym kombinacjom wartości wymiarów, a następnie podzielenie wyniku przez sumę tych iloczynów w zbiorowości wszystkich wygenerowanych kombinacji. W tym momencie można już łatwo obliczyć oszacowania EAP dla każdego z wymiarów jako sumę iloczynów wartości danego wymiaru w ramach danej kombinacji i przypisanej tej kombinacji w poprzednim kroku częstości występowania w rozkładzie *a posteriori*. Z wykorzystaniem tych samych danych obliczany jest też błąd standardowy oszacowania dla każdego z wymiarów (jako odchylenie standardowe rozkładu *a posteriori* danego wymiaru). Obliczony wynik pomiaru i jego błąd standardowy zapisywane są następnie w bazie danych systemu. Wyniki zapisywane są w postaci niewystandaryzowanej – opisana we wcześniejszej części tekstu operacja zrelatywizowania ich względem grupy badanych, do której powinien zostać odniesiony wynik właśnie zdiagnozowanej osoby, przeprowadzana jest każdorazowo (*w locie*) tuż przed zaprezentowaniem wyników użytkownikowi.

[14] W celu optymalizacji szybkości obliczeń system sprawdza, czy struktura modelu nie pozwala na przeprowadzenie obliczeń na pewnych wymiarach niezależnie od innych (jest to możliwe, jeśli w modelu założono ortogonalność niektórych wymiarów oraz pomiędzy ortogonalnymi wymiarami nie występują ładunki krzyżowe) – w takim przypadku kryterium sześciu wymiarów stosuje się oddzielnie do każdego zbioru wymiarów, dla którego obliczenia mogą zostać przeprowadzone niezależnie od innych.

[15] Jeśli do wygenerowania zestawu kombinacji wartości wymiarów wykorzystana została metoda Monte Carlo, zamiast wartości funkcji gęstości stosuje się wartość 1 dla wszystkich kombinacji – ich odpowiednie przeważenie zapewnia w takim wypadku procedura losowania będąca elementem metody Monte Carlo.

Przeprowadzone testy systemu informatycznego wykazały poprawność implementacji: uzyskiwane oszacowania były niemal identyczne jak te zwracane przez program Mplus (dla takiej samej specyfikacji modelu i profili odpowiedzi), zarówno dla prostych modeli jednowymiarowych, jak i dla modeli wielowymiarowych, zawierających ładunki krzyżowe, a także modeli z czynnikiem wyższego rzędu. Obliczanie wyniku nawet dla złożonych modeli może być dokonane w czasie zbliżonym do rzeczywistego, co gwarantuje uzyskanie dostępu do wyników diagnozy bezpośrednio po zakończeniu badania przez ucznia, bez konieczności podejmowania żadnych działań ze strony diagnosty.

Prezentacja wyników pomiaru

Zapewnienie dostępu do bardziej adekwatnych metod szacowania wyników pomiaru jest ważnym krokiem na drodze rozwoju komputerowych narzędzi diagnostycznych, niemniej z punktu widzenia diagnosty-użytkownika może to nie być argument rozstrzygający o chęci skorzystania z tego typu rozwiązania. Duże znaczenie wydaje się mieć w tym kontekście zapewnienie odpowiedniego sposobu prezentacji wyników, ułatwiającego ich interpretację i dalszą analizę oraz wspomagającego przejście od *suchych* liczbowych wyników pomiaru do diagnozy i określenia koniecznych do podjęcia działań. Aby wyjść naprzeciw tym potrzebom, w opracowanej przez Diagmatic *Aplikacji Diagnosty*, służącej użytkownikom do zarządzania prowadzonymi diagnozami indywidualnymi i analizy ich wyników, położono bardzo duży nacisk na zapewnienie użytkownikom dostępu do różnorodnych form prezentowania wyników pomiaru. Mogą one zostać podzielone na dwie grupy:

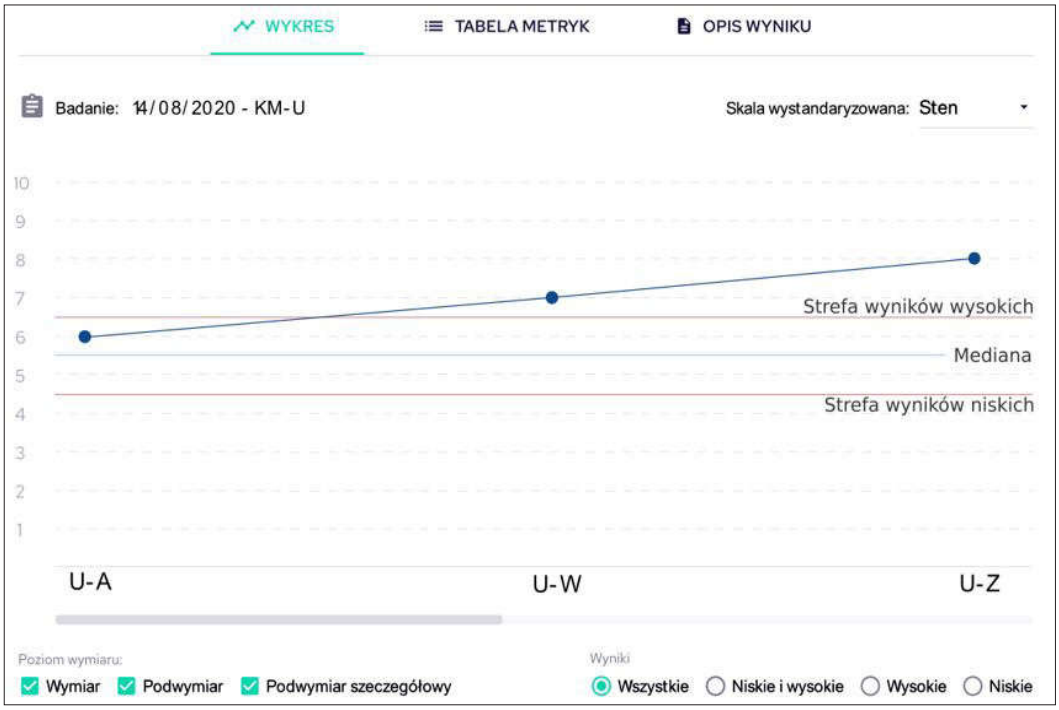
1. Formy tabelaryczne (tzw. *tabela metryk*, patrz: rysunek 1) i graficzne (wykresy – patrz: rysunek 2), które ułatwiają porównywanie ze sobą wyników osiągniętych na różnych wymiarach (w różnych obszarach podlegających diagnozie) lub – o ile dana osoba była badana co najmniej dwukrotnie – śledzenie zmian natężenia poszczególnych cech w czasie.
2. Formy opisowe, ułatwiające zrozumienie znaczenia uzyskanych wyników liczbowych i określenie dalszych działań, jakie należałoby podjąć względem zdiagnozowanej osoby (patrz: rysunek 3).

Zestawienie wyników w formie tabeli lub wykresu pozwala szybko i wygodnie przeanalizować wyniki testów przeprowadzonych dla danej osoby. W przypadku przygotowanych w ramach projektu narzędzi diagnostycznych zapewnienie wsparcia w tym zakresie jest o tyle istotne, że dostarczają one wyników w podziale na bardzo wiele wymiarów: łącznie do 24 w przypadku Baterii Kwestionariuszy Motywacji i Amotywacji i aż 51 w przypadku Baterii Kwestionariuszy Funkcjonowania Społecznego (por. Wysocka, 2020). Aby pomóc użytkownikowi przeanalizować interesujące go wyniki, zaprojektowany został interfejs, przy pomocy którego może on określić (w ramach danej baterii kwestionariuszy) zestaw prezentowanych na wykresie lub w tabeli wyników albo poprzez odwołanie się do struktury narzędzia diagnostycznego, tj. wskazanie interesujących go obszarów diagnozy (np. *sfera uczenia się* lub *sfera relacji rówieśniczych*), albo interesujących go wartości wyników pomiaru (pokazanie tylko tych wymiarów, w których diagnozowana osoba uzyskała wyniki szczególnie wysokie lub szczególnie niskie). W przypadku Baterii Kwestionariuszy

Funkcjonowania Społecznego, w którym to narzędziu wymiary zorganizowane są w sposób hierarchiczny: niektóre z nich opisują bardziej ogólne cechy, a inne (tzw. podwymiary) bardziej szczegółowe składowe tych pierwszych, możliwe jest też wybranie prezentowanych wymiarów poprzez określenie ich *poziomu* (wymiar/podwymiar).

Rysunek 1. Przykład prezentacji wyników pomiaru w formie tabeli w *Aplikacji Diagnosty* (prezentowane są wyniki na wymiarach amotywacji: U-A, motywacji wewnętrznej: U-W i motywacji zewnętrznej: U-Z w ramach *sfery uczenia się*)

Skala, na której prezentowane są wyniki pomiaru, może zostać wybrana przez użytkownika spośród czterech: stenowej, staninowej, tenowej i centylowej. Dwie pierwsze są szeroko znane i wykorzystywane, co czyni je odpowiednim dla mniej zaawansowanych, ale z racji niewielkiej liczby różnych możliwych wartości wyniku (odpowiednio 10 lub 9) są one mniej precyzyjne, a przez to nie pozwalają wykorzystać w pełni zalet zaimplementowanego w systemie sposobu obliczania wyników pomiaru. Z kolei skale tenowa i centylowa zapewniają większą precyzję wyrażenie wyników, ale dla niektórych mogą być one zbyt skomplikowane. Swoboda wyboru pozwala użytkownikowi wykorzystać skalę odpowiadającą jego umiejętnościom i potrzebom – wybór jednej z nich następuje w *Aplikacji Diagnosty* przy pomocy kontrolki umieszczonej przy wykresie lub tabeli, a wyniki przeliczane są automatycznie. Aby ułatwić odczytywanie informacji z wykresu (patrz: rysunek 2) graficznie wyróżnione zostały na nim poziomymi liniami mediana skali oraz granice oddzielające zakres wyników uznawanych za wysokie (powyżej 6. stena) i niskie (poniżej 5. stena) od wyników *średnich*.

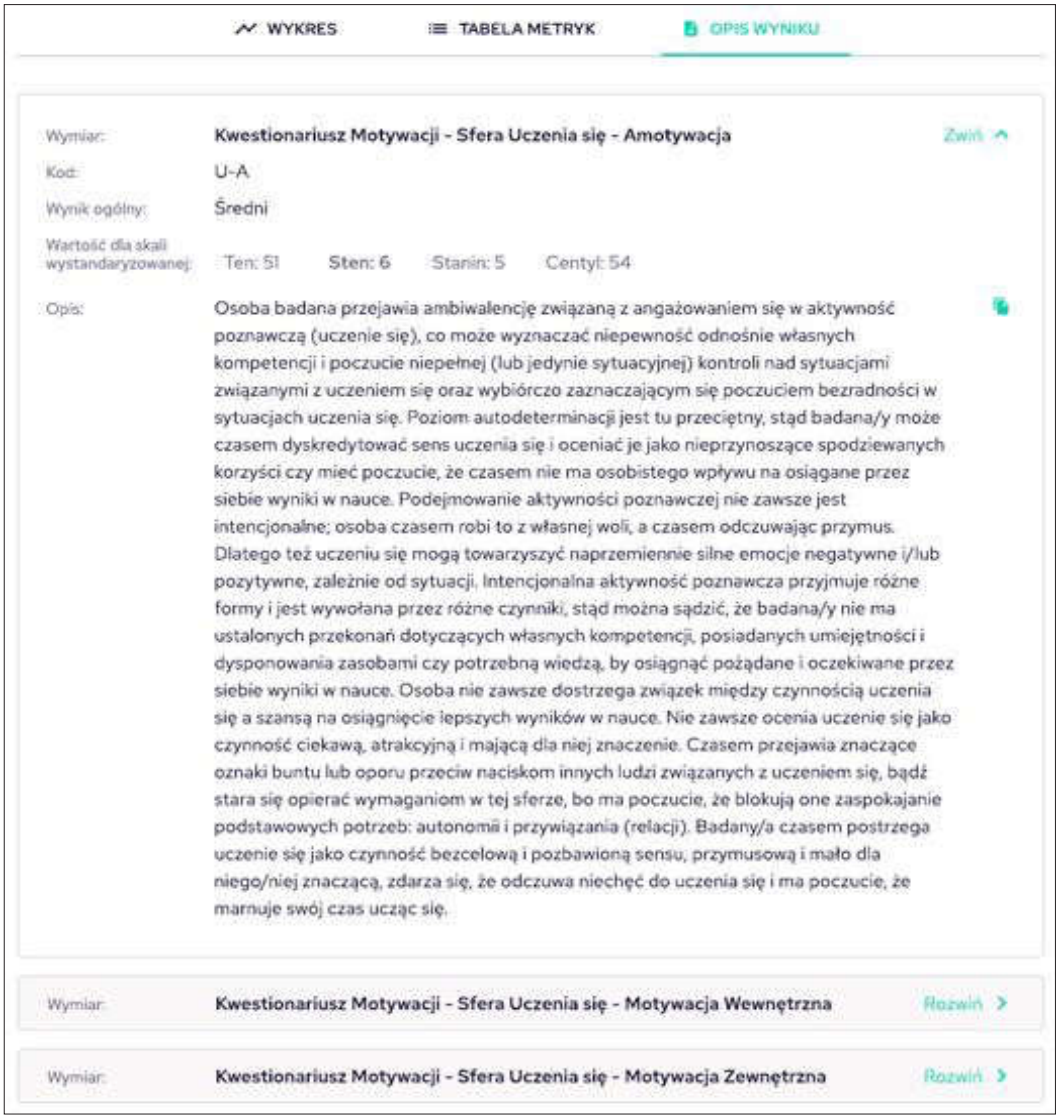


Rysunek 2. Przykład prezentacji wyników pomiaru na wykresie w *Aplikacji Diagnosty* (prezentowane są wyniki na wymiarach amotywacji: U-A, motywacji wewnętrznej: U-W i motywacji zewnętrznej: U-Z w ramach *sfery uczenia się*)

Oprócz możliwości przeprowadzenia analiz ilościowych niezwykle istotne wydaje się dla diagnostów przedstawienie wyników w sposób opisowy, ułatwiający interpretację i przede wszystkim określenie dalszych działań, jakie należy podjąć względem zdiagnozowanej osoby. Aby zapewnić użytkownikom wsparcie w tym zakresie, autorzy narzędzi diagnostycznych opracowanych w ramach projektu firmy Diagmatic przygotowali oddzielne opisy wyników dla każdego z wyróżnionych wymiarów (oraz podwymiarów), w podziale na trzy zakresy wyników: niskie (steny 1–4), średnie (steny 5–6) i wysokie (steny 7–10)[16]. Jest to warte podkreślenia, gdyż w dotychczas dostępnych narzędziach zwykle udostępniane są jedynie opisy poszczególnych wymiarów, które użytkownik samodzielnie musi zinterpretować w odniesieniu do wyniku pomiaru danej osoby. Przykładowy opis dla wymiaru amotywacji w ramach *sfery uczenia się* (dla zakresu wyników średnich – co odpowiada wynikowi danego badanego) przedstawiony został na rysunku 3. Przy pomocy przycisków *rozwiń* możliwe jest obejrzenie opisów również dla innych wymiarów tej samej baterii testów. Nad opisem prezentowane są również wyniki pomiaru wyrażone na trzech wspomnianych wcześniej skalach standardowych oraz na skali centylowej. Dodatkowo, poniżej opisów wyników dla poszczególnych wymiarów, prezentowane są w aplikacji (niewidoczne na rysunku 3) przygotowane przez autorów narzędzia zalecenia postdiagnostyczne, dotyczące sposobów przeciwdziałania przez nauczycieli negatywnym zjawiskom, do których diagnozy służy dane narzędzie.

[16] System informatyczny umożliwia dowolne zdefiniowanie takich zakresów przez autorów narzędzi diagnostycznych, ale w przypadku obu omawianych baterii kwestionariuszy zastosowano właśnie ten podział.

Wspomniane opisy zawarte są też w podręcznikach do opracowanych narzędzi diagnostycznych (baterii kwestionariuszy). *Aplikacja Diagnosty* zapewnia jednak użytkownikom znacznie wygodniejszy dostęp do tych informacji niż dokument papierowy lub jego wersja elektroniczna (np. plik PDF), w których użytkownik musiałby samodzielnie wyszukiwać odpowiednie fragmenty z interesującymi go informacjami. Niezależnie od aspektu technologicznego podkreślić należy przede wszystkim bardzo szeroki i szczegółowy zakres opisów (interpretacji) przygotowanych w ramach projektu przez autorów narzędzi. Możliwość skorzystania z nich będzie niewątpliwie dużym ułatwieniem dla użytkowników, zwłaszcza tych posiadających ograniczone doświadczenie w interpretacji wyników analiz ilościowych.



Rysunek 3. Przykład prezentacji wyniku w formie opisu w *Aplikacji Diagnosty*

Podsumowanie

Rozwój technologii informatycznych otwiera nowe możliwości również w dziedzinie diagnostyki edukacyjnej i psychologicznej. Wykorzystanie komputerowych narzędzi do diagnozy pozwala uczynić pracę diagnosty łatwiejszą i szybszą, pozostawiając więcej czasu na bezpośredni kontakt z badanymi i wypracowywanie rozwiązań mających na celu wsparcie ich w rozwoju. Ważne jest jednak przy tym, aby tworząc nowe rozwiązania, nie ograniczać się tylko do implementacji w formie aplikacji komputerowych dokładnie tych samych procedur, które stosowane są w badaniach prowadzonych w klasycznej, papierowej formie, lecz aby wdrażać również metody, które wcześniej, z racji ograniczeń technicznych, były niedostępne. W szczególności dotyczy to sposobów obliczania wyników pomiaru oraz formy ich prezentacji. W tekście zilustrowane zostały możliwości zastosowania nowych metod w tej dziedzinie, na przykładzie rozwiązań przygotowanych przez firmę Diagmatic w ramach projektu „Opracowanie inteligentnych narzędzi do diagnozy psycho-społeczno-edukacyjnej dzieci i młodzieży". Obejmują one wykorzystania estymatora EAP, wykorzystującego parametry wielowymiarowego i wielogrupowego modelu czynnikowego do obliczania wyników pomiaru oraz zapewnienie diagnoście dostępu do różnych form prezentacji wyników, zarówno w formie liczbowej, jak i opisowej. Aby wykorzystać potencjał, który dają te rozwiązania technologiczne, konieczne jest jednak również dysponowanie dostosowanymi do wykorzystania razem z nimi nowoczesnymi, wysokiej jakości narzędziami diagnostycznymi. W ramach wspomnianego projektu opracowane zostały dwa takie narzędzia: Bateria Kwestionariuszy Motywacji i Amotywacji oraz Bateria Kwestionariuszy Funkcjonowania Społecznego (patrz: Wysocka, 2020). Szeroki zakres diagnozowanych cech, zachowanie najwyższych standardów w procesie opracowania kwestionariuszy[17], zapewnienie szerokiego zakresu opisów pomagających w interpretacji wyników diagnozy oraz integracja z omówionymi wcześniej rozwiązaniami technologicznymi stanowią gwarancję ich wysokiej użyteczności w pracy diagnostycznej.

Bibliografia

American Psychological Association (APA), *The Standards for Educational and Psychological Testing*, AERA 2014.

Bock, R. D., Mislevy, R. J., *Adaptive EAP estimation of ability in a microcomputer environment* [w:] „Applied Psychological Measurement" 1982 nr 6, s. 431–444.

Brzeziński, J. M., *Metodologia badań psychologicznych*, Wydawnictwo Naukowe PWN, Warszawa 2019.

Czub, M. (red.) *Diagnoza funkcjonowania społeczno-emocjonalnego dziecka w wieku od 1,5 do 5,5 lat*, Instytut Badań Edukacyjnych, Warszawa 2014.

[17] Procedura opracowania narzędzi przeprowadzona została zgodnie z wytycznymi American Psychological Association (2014), European Federation of Psychologists' Asscociations (2013) i International Test Commision (2000).

Domagała-Zyśk, E., Knopik, T., Oszwa, U., *Diagnoza funkcjonalna rozwoju społeczno-emocjonalnego uczniów w wieku 9–13 lat*, Ośrodek Rozwoju Edukacji, Lublin 2017.
Embretson, S. E., Reise, S. P., *Item Response Theory for Psychologists*. Erlbaum, Mahwah 2000.
European Federation of Psychologists' Asscociations (EFPA), *EFPA Review Model for the Description and Evaluation of Psychological and Educational Tests* version 4.2.6, 2013, http://assessment.efpa.eu/download/650d0d4ecd407a51139ca44ee704fda4
Green, S. B., Yang, Y., *Commentary on Coefficient Alpha: A Cautionary Tale* [w:] „Psychometrika" 2009 nr 74(1), s. 121–135.
Grygiel, P., Switaj, P., Humenny, G., *Zróżnicowane funkcjonowanie pozycji testowych skali stygmatyzacji z Kwestionariusza piętna i dyskryminacji* [w:] Pokropek, A. (red.), *Modele cech ukrytych w badaniach edukacyjnych, psychologicznych i socjologii. Teoria i zastosowania*, Instytut Badań Edukacyjnych, Warszawa 2015.
Humenny, G., Grygiel, P., *Wielowymiarowa struktura latentna w perspektywie analizy czynnikowej* [w:] Pokropek, A. (red.), *Modele cech ukrytych w badaniach edukacyjnych, psychologicznych i socjologii. Teoria i zastosowania*, Instytut Badań Edukacyjnych, Warszawa 2015.
International Test Commission (ITC), *International Guidelines for Test Use., 2000,* www.intestcom.org
Kaczan, R., Rycielski, P., *Diagnoza umiejętności dzieci 5-, 6- i 7-letnich za pomocą Testu Umiejętności na Starcie Szkolnym TUnSS* [w:] Niemierko, B. i Szmigel, M. K. (red.), *Regionalne i lokalne diagnozy edukacyjne. Materiały XVIII Krajowej Konferencji Diagnostyki Edukacyjnej*, Grupa TOMAMI, Kraków 2012.
Karwowski, M., Dziedziewicz, D., *Test Umiejętności na Starcie Szkoły (TUNSS). Podręcznik*. Instytut Badań Edukacyjnych, Warszawa 2012.
Kondratek, B., Pokropek, A., *Teoria odpowiedzi na pozycje testowe: jednowymiarowe modele dla cech ukrytych o charakterze ciągłym* [w:] Pokropek, A. (red.), *Modele cech ukrytych w badaniach edukacyjnych, psychologicznych i socjologii. Teoria i zastosowania*, Instytut Badań Edukacyjnych, Warszawa 2015.
Kondratek, B., Skórska, P, Świst, K., *Wprowadzenie do zróżnicowanego funkcjonowania pozycji testowej* [w:] Pokropek, A. (red.), *Modele cech ukrytych w badaniach edukacyjnych, psychologicznych i socjologii. Teoria i zastosowania*, Instytut Badań Edukacyjnych, Warszawa 2015.
Niemierko, B., *Pomiar wyników kształcenia*, WSiP, Warszawa 1999.
Warm, T. A., *Weighted likelihood estimation of ability in item response theory* [w:] „Psychometrika" 1989 nr 54, s. 427–450.
Wysocka, E., *Baterie Kwestionariuszy Motywacji (B-KMiA) i Funkcjonowania Społecznego (B-KFS) – założenia teoretyczne i zastosowanie w poradnictwie psychologiczno-pedagogicznym* [w:] Niemierko, B. i Szmigel, M. K. (red.), *Rola społeczna diagnostyki edukacyjnej. Materiały XXVI Krajowej Konferencji Diagnostyki Edukacyjnej*, Grupa TOMAMI, Kraków 2020.
Żółtak, T., *Statystyczne modelowanie wskaźników Edukacyjnej Wartości Dodanej. Podsumowanie polskich doświadczeń 2005–2015*, Instytut Badań Edukacyjnych, Warszawa 2015.