

Kryteria wyboru metod pomiaru: jak rozpoznać rzetelne narzędzie?

prof. Agata Gąsiorowska¹

Czym są testy psychologiczne, jakie powinny być ich właściwości oraz jakie korzyści mogą z nich wynikać dla klientów? To kluczowe pytania, na które warto odpowiedzieć. Zgodnie z definicją Amerykańskiego Towarzystwa Psychologicznego (APA), test to narzędzie lub procedura służące ocenie. Ich istotą jest uzyskanie w określonych warunkach próbki zachowania osób badanych i następnie dokonanie oceny zgodnie z ustalonymi standardami. Testem psychologicznym nie jest więc każdy zbiór pytań czy zadań, a jedynie taki, który spełnia określone kryteria. Te kryteria są jasne: test musi być obiektywny i standaryzowany, musi charakteryzować się wysoką rzetelnością i trafnością oraz mieć odpowiednio opracowane normy (por. Anastasi i Urbina, 1999; Hornowska, 2005). Inaczej mówiąc, dobry test powinien zbierać reprezentatywną próbkę zachowania typowego dla przejawów określonego konstrukt psychologicznego (czego nie daje zbudowana ad hoc ankieta), dawać wynik niezależny od osoby prowadzącej i warunków zewnętrznych (czego nie daje obserwacja), powinien też pozwalać na porównanie wyniku konkretnej osoby badanej do populacji (czego nie daje wywiad). W dalszej części niniejszego rozdziału omówione zostały podstawowe kryteria jakości psychometrycznej testów psychologicznych, czyli formalne właściwości każdej metody pomiarowej, które decydują o wartości uzyskiwanych wyników, ze szczególnym uwzględnieniem testów samoopisowych stosowanych do diagnozy indywidualnej. Warto w tym miejscu wspomnieć, że kryteria te omawiane są na każdym kursie psychometrii w ramach studiów psychologicznych, jednak zasadniczy problem tkwi w rozumieniu konsekwencji stosowania metod, które nie spełniają tych warunków (por. Anastasi i Urbina, 1999; Hornowska, 2005).

¹ za: Wojtkowska, A., Gąsiorowska, A. (red.). *Profilaktyka, diagnoza i terapia e-uzależnień wśród dzieci i młodzieży: praktyczny podręcznik dla specjalistów pracy z dziećmi i ich rodzinami*. Wydawnictwo Naukowe Liberi Libri (2024, w druku).

3.2.1. Standaryzacja i obiektywność

Dobry test psychometryczny jest **obiektywny**, co oznacza niezależność wyników testowania od osoby, miejsca i czasu ich oceny. Test można uznać za obiektywny, jeżeli dwie różne osoby opracowujące jego wyniki dochodzą do tego samego rezultatu. Aby osiągnąć taki poziom obiektywności, niezbędna jest **standaryzacja**, czyli jednolitość warunków badania. Standaryzacja zapewnia, że testy są przeprowadzane w jednakowych warunkach, co minimalizuje wpływ zewnętrznych zmiennych zakłócających na wyniki testowania. Aby więc testowanie było skuteczne, konieczne jest dokładne i szczegółowe określenie sposobu prowadzenia badań testowych. Obejmuje ono między innymi procedury badania testem, które określają, jak test ma być przeprowadzany. Kolejnym krokiem są procedury obliczania wyników, które zapewniają jednolity sposób analizy i interpretacji danych uzyskanych z testu. Ostatecznie procedury interpretowania wyników pozwalają na prawidłowe zrozumienie i wykorzystanie danych. Wszystkie te procedury powinny być szczegółowo opisane w podręczniku, który stanowi podstawowe źródło informacji dla osób przeprowadzających testy. Podręcznik ten gwarantuje, że każdy diagnosta, niezależnie od swojego doświadczenia, będzie mógł przeprowadzić badanie zgodnie z ustalonymi standardami, zapewniając tym samym niezawodność i obiektywność wyników testowania. Elementy podręcznika w odniesieniu do standaryzacji badania testowego obejmują w szczególności:

1. kolejność czynności, które wykonuje osoba prowadząca badania, w tym dokładny tekst instrukcji podawanej osobie badanej, co umożliwi jednolite przekazywanie informacji;
2. arkusz odpowiedzi (jeśli jest przewidziany) przygotowany i stosowany w każdym badaniu w identycznym kształcie;

3. klucz, przy pomocy którego dokonuje się punktacji uzyskanych rezultatów badań;
4. normy oceny uzyskanych wyników surowych, które pozwalają na kontekstualizację wyników;
5. zasady, przy pomocy których interpretuje się uzyskane rezultaty badań testowych zgodnie z duchem teorii, w oparciu o którą test został opracowany, co umożliwia spójne i teoretycznie uzasadnione interpretowanie wyników (Anastasi i Urbina, 1999; Hornowska, 2005).

Dzięki tym elementom, podręcznik staje się niezastąpionym narzędziem w procesie testowania, zapewniającym zarówno standaryzację, jak i rzetelność badań. Warto w tym miejscu wspomnieć także, że każde naruszenie ww. procedur w procesie używania testu psychologicznego – np. stworzenie własnego arkusza, używanie w badaniach online testów, które zostały zwalidowane jedynie w badaniach papier-ołówek, wybranie części pytań zamiast posługiwania się całym zestawem pozycji testowych, używanie w sposób niekonsekwentny zasad interpretowania wyników – nie tylko naruszają standaryzację i obiektywność testowania, ale mają negatywne konsekwencje dla pozostałych kryteriów jakości psychometrycznej, w tym trafności interpretacji wyniku testowego i rzetelności pomiaru (Anastasi i Urbina, 1999; Hornowska, 2005).

3.2.2. Rzetelność

Dobry test jest **rzetelny**, co oznacza, że wyniki testowe osoby badanej są dokładne, a więc obarczone niewielkim błędem pomiaru. Im większa rzetelność, tym powtarzalność wyniku jest większa, a błąd pomiaru mniejszy, i odwrotnie. W pewnym stopniu wszystkie pomiary psychologiczne są nierzetelne, ponieważ wynik obserwowany (czyli otrzymany w teście) nigdy nie jest idealnym odzwierciedleniem rzeczywistej wartości mierzonej cechy – zawsze jest on obarczony pewnym błędem. Inaczej mówiąc, nie ma idealnego testu, tak jak

nie ma idealnego termometru czy centymetra. Jednak co istotne, rzeczywista wartość mierzonej cechy jest trudna do uchwycenia w pełni, dlatego każdy testowy wynik jest tylko przybliżeniem tej wartości – a diagnosta nie wie i nie może wiedzieć, jaki dokładnie błąd popełnia przy konkretnym pomiarze (Anastasi i Urbina, 1999; Hornowska, 2005).

Rzetelność i błąd pomiaru w pierwszej kolejności wynikają z samej konstrukcji testu – na przykład testy krótkie, o małej liczbie pozycji testowych, mają zwykle mniejszą rzetelność niż testy długie. Oczywiście, błąd pomiaru może wynikać z różnych źródeł pozatestowych, takich jak warunki badania, narzędzia pomiarowe czy subiektywność osoby przeprowadzającej test. Dlatego tak ważne jest dążenie do minimalizowania tych błędów poprzez standaryzację i dokładne przestrzeganie określenie procedur testowych (Anastasi i Urbina, 1999; Hornowska, 2005).

Skoro błąd pomiaru jest nieodzowną częścią testowania, każdy podręcznik do testu psychologicznego powinien zawierać zróżnicowane dane dotyczące rzetelności pomiaru. Najpopularniejszymi metodami określania rzetelności są takie parametry, jak alfa (α) Cronbacha lub omega (ω) McDonalda. Problem ze stosowaniem tych współczynników polega jednak na tym, że mierzą one bardzo specyficzny aspekt rzetelności, to jest spójność wewnętrzną pomiaru. Im bardziej pozycje testowe są podobne do siebie, i im jest ich więcej, tym wskaźniki te są wyższe, co niekoniecznie musi się przekładać na powtarzalność pomiaru. Z tego powodu dużo lepszym wskaźnikiem rzetelności pomiaru jest stabilność uzyskiwanych wyników w czasie, określana jako korelacje między co najmniej dwoma powtarzonymi pomiarami (test – retest). Niezależnie od rodzaju współczynników rzetelności, zwykle przyjmuje się, że ich satysfakcjonująca wartość powinna przekraczać 0,70–0,75. Współczynnik równy 0,80 oznacza, że 80% zmienności wyników w tym teście wynika ze zmienności wyników prawdziwych, a 20% jest związane z błędem pomiaru (Anastasi i Urbina, 1999; Hornowska, 2005).

Wykorzystanie danych o rzetelności w praktyce diagnozowania oznacza technicznie umiejętność zbudowania przedziału ufności dla wyniku prawdziwego badanej osoby. W podręczniku każdego profesjonalnego testu psychologicznego przeznaczonego do diagnozy indywidualnej powinny więc znajdować się dane o wielkości standardowego błędu pomiaru, które pozwalają na określenie granic przedziału, w którym z odpowiednim prawdopodobieństwem mieści się wynik prawdziwy badanej osoby, lub też bezpośrednio dane o szerokości przedziału ufności. Informacje te powinny być podane dla każdego wyniku czy wyników które podlegają interpretacji w danym teście psychologicznym. Jeśli więc test jest wielowymiarowy, autorzy powinni podać informacje o rzetelności i błędzie pomiaru dla każdego z analizowanych wymiarów (podskal narzędzia) (Anastasi i Urbina, 1999; Hornowska, 2005).

Do obowiązków diagnosty należy uwzględnienie wielkości tego błędu przy interpretowaniu wyników. Inaczej mówiąc, przy interpretacji wyniku testu należy uwzględnić wszystkie czynniki, które mogły wpłynąć na jego wartość, w tym potencjalne źródła błędu, ustrzegając się przed przywiązywaniem nadmiernej wagi do pojedynczego wyniku liczbowego. Podawanie wyniku testu w formie przedziału ufności o ufności przynajmniej na poziomie 85% sprzyja więc właściwej interpretacji wyników testowych (Anastasi i Urbina, 1999; Hornowska, 2005).

Dodatkowo warto zwrócić uwagę, na jakich próbach prowadzone były analizy dotyczące rzetelności testu. Próba używana w takiej analizie powinna spełniać trzy kluczowe kryteria: powinna być wystarczająco duża, zróżnicowana oraz reprezentatywna dla docelowej populacji. Duża próba zwiększa precyzję wyników i redukuje wpływ przypadkowych błędów. Zróżnicowana próba zapewnia uwzględnienie różnych czynników, które mogą wpływać na wyniki testu, co zwiększa jego uniwersalność. Reprezentatywność próby gwarantuje, że wyniki analizy będą miały zastosowanie do całej populacji, dla której

test jest przeznaczony. Na przykład jeśli test przeznaczony jest do użytku w badaniu dorosłych Polaków, próby te powinny być reprezentatywne dla dorosłej populacji Polski. Jeśli test przeznaczony jest dla dzieci w określonym wieku, analiza rzetelności powinna opierać się na odpowiednio zróżnicowanych próbach dzieci dokładnie w tym wieku. W przypadku badań dzieci i młodzieży może też być istotne, by przedstawić parametry rzetelności w poszczególnych podgrupach wyróżnionych ze względu na wiek (Anastasi i Urbina, 1999; Hornowska, 2005).

3.2.3. Trafność

Interpretacja wyniku testowego jest **trafna**, jeśli wynik w danym teście faktycznie można interpretować jako wskaźnik określonej cechy psychologicznej, którą ten test miał mierzyć, a nie np. cechy podobnej. Testy psychologiczne są zawsze stosowane w określonym celu (Anastasi i Urbina, 1999; Hornowska, 2005). Na przykład na podstawie wyników testów inteligencji możemy być zainteresowani przewidywaniem predyspozycji menedżerskich badanych kandydatów, a na podstawie testu stylów kierowania – oceną, czy dana osoba może być efektywnym kierownikiem. O możliwości wykorzystania testu w konkretny sposób decydują właśnie dane o trafności. Problem ten jest szczególnie istotny z perspektywy problematyki niniejszej monografii, ponieważ, jak wcześniej omówiono (por. rozdział 1.1.), granice definicyjne pomiędzy takimi pojęciami jak „nadużywanie e-mediów”, „problematiczne użytkowanie Internetu”, czy „uzależnienie cyfrowe” wcale nie są jasne, przez co w praktyce używane są narzędzia określane jako testy uzależnienia, a które de facto nie mierzą wszystkich aspektów uzależnienia wyróżnianych w aktualnie stosowanych teoriach naukowych.

Trafność testu to obszar jego zastosowania. Ważne jest, aby pamiętać, że trafność dotyczy zawsze konkretnego zastosowania testu (por. Anastasi i Urbina, 1999; Hornowska,

2005). Nie ma testów, które można stosować wszędzie i dla każdego celu. Dlatego dane dotyczące trafności powinny być analizowane szczególnie starannie, a wybór konkretnego testu powinien być poprzedzony dokładną analizą celu badania. Nie jest więc raczej możliwe, aby jednym testem zbadać wszystkie przejawy nadużywania mediów elektronicznych przez dzieci i młodzież.

W podręczniku testowym powinny więc znaleźć się informacje na temat tego, jak określano i weryfikowano trafność pomiaru oraz sposobów interpretacji wyniku testowego. Trafność ta zależy przede wszystkim od celnego doboru wskaźników, na podstawie których mierzona jest określona cecha. Każda pozycja testowa (ujęta w postaci pytania bądź stwierdzenia) ma więc celnie odnosić się do zdefiniowanego aspektu mierzonej cechy, zbierając dane o tym, jak przejawia się ona w obserwowalnych ludzkich zachowaniach. Zachowania, do jakich odnoszą się pytania lub stwierdzenia testu, powinny być jak najbardziej reprezentatywne w stosunku do ogółu zachowań, które wyrażają badaną cechę. Dobry test jest więc oparty na dowodach naukowych, badane nim cechy są jasno zdefiniowane, a uzyskane wyniki można interpretować w odniesieniu do obecnej wiedzy naukowej (teorii, modeli, wyników badań) i doświadczeń praktycznych (Anastasi i Urbina, 1999; Hornowska, 2005).

Określanie trafności danego testu, nazywane w psychometrii procesem **walidacji** testu (*validation*), polega na zbieraniu i ocenie danych świadczących o trafności interpretacji wyników testu. Im więcej badań przeprowadza się z udziałem danego testu, tym szerszy jest potencjalny obszar jego zastosowania. Procedura walidacji testu nie może więc skończyć się na podaniu jednego współczynnika trafności. Jest to proces ciągły, który obejmuje prowadzenie badań i akumulowanie zebranych informacji. Autorzy testu powinni więc dostarczyć w podręczniku zróżnicowanych informacji na temat tego, jak prowadzili jego walidację. Podobnie jak w przypadku rzetelności warto też zwrócić uwagę, na jakich

próbach prowadzone były analizy dotyczące trafności (Anastasi i Urbina, 1999; Hornowska, 2005).

Należy w tym miejscu także dodać, że na przestrzeni ostatnich 20 lat trafność przekształciła się z zamkniętej procedury w otwarty proces badawczy. Co więcej, nawet w jej definicji nastąpiło wyraźne przesunięcie punktu ciężkości z trafności testu na trafność interpretacji wyników danego testu. Ta zmiana perspektywy — od samego testu do konsekwencji jego stosowania — ma na celu zwrócenie uwagi użytkowników testów na to, że najważniejsze są efekty badania testami. Nawet najbardziej staranne opracowanie psychometryczne testu nie zastąpi refleksji psychologicznej.

3.2.4. Normalizacja

Dobry test psychologiczny służący do diagnozy powinien być normalizowany. Wiele narzędzi, zwłaszcza adaptowanych z innych kręgów językowych, ma zweryfikowane i satysfakcjonujące wskaźniki rzetelności i trafności pomiaru, ale nie zawiera norm dla populacji, w której mają być zastosowane. W takiej sytuacji narzędzia te mogą być przeznaczone wyłącznie do ilościowych badań naukowych, w których wystarczające jest posługiwanie się surowymi wynikami. Nie nadają się jednak do diagnozy indywidualnej, ponieważ nie umożliwiają kontekstualizacji oceny. Inaczej mówiąc, nie wiadomo, czy wynik konkretnej osoby badanej należy uznać za niski, przeciętny czy wysoki (Anastasi i Urbina, 1999; Hornowska, 2005).

Pojedynczy wynik otrzymany w teście psychologicznym nie ma znaczenia, dopóki nie można go porównać do precyzyjnego i jednolitego układu odniesienia. Istotą testów psychologicznych jest więc to, że podstawą interpretacji wyników jest zawsze określony układ odniesienia, najczęściej o charakterze statystycznym. Oznacza to, że wynik surowy danej osoby odnosi się do rozkładu wyników uzyskanego w próbie standaryzacyjnej, co

pozwała określić, w którym miejscu tego rozkładu mieści się badana osoba. Tym samym normalizacja to proces statystyczny, który polega na wyznaczeniu ram interpretacji surowych wyników osoby badanej na tle wyników uzyskiwanych przez jej grupę rówieśniczą lub ogólną populację. Zwykle do normalizacji używa się skal o mniejszym zróżnicowaniu wyników niż oryginalne zróżnicowanie wyników surowych, takich jak skala stenowa (od 1 do 10) czy staninowa (od 1 do 9). W ten sposób uzyskuje się informację, czy dany wynik świadczy o nasileniu niższym, wyższym czy podobnym do średniej uzyskiwanej w populacji odniesienia (Anastasi i Urbina, 1999; Hornowska, 2005).

Skoro podstawą formułowania wniosków o właściwościach psychologicznych badanych osób jest ocena tego, jak dana osoba wypada na tle grupy odniesienia, sposób wybór grupy odniesienia ma kluczowe znaczenie dla końcowych wniosków. Normy testowe mają charakter względny, zależą od tego, kto tworzy grupę odniesienia. W ten sposób wynik tej samej osoby może być raz zinterpretowany jako niski, raz jako przeciętny, a raz jako wysoki, w zależności od grupy, do której zostanie porównany (Anastasi i Urbina, 1999; Hornowska, 2005). Brak właściwych, opracowanych dla populacji polskiej norm, dyskwalifikuje metodę jako test psychologiczny. Nie można bowiem wyciągać sensownych wniosków z porównania polskiego nastolatka badanego w roku 2024 z normami opracowanymi dla populacji amerykańskiej 10 czy 15 lat temu.

Podsumowując, normalizacja wymaga zebrania licznych danych z wykorzystaniem normalizowanego testu, zwykle prowadzonych na reprezentatywnych grupach osób badanych, oraz zastosowania zaawansowanych metod obliczeniowych. Dlatego nie wszystkie narzędzia czy ich adaptacje są normalizowane. Dla diagnostów chcących przeprowadzić pomiar danej cechy u konkretnej jednostki, szczególnie istotne będzie wybieranie tych testów, które zostały znormalizowane. Pozostałe narzędzia, mimo dobrej

rzetelności i trafności, mogą być z powodzeniem wykorzystywane w badaniach naukowych, ale nie w diagnostyce indywidualnej (Anastasi i Urbina, 1999; Hornowska, 2005).

3.2.5. Adaptacja kulturowa

Pojęcie **adaptacji kulturowej** dotyczy przede wszystkim testów, które powstały poza granicami naszego kraju i zostały opracowane w innym języku niż polski, a które chcemy wprowadzić do użytku również w krajowej praktyce. Test taki musi być dostosowany do warunków populacji, na której ma być używany. Adaptacja definiowana jest więc jako proces przystosowania wersji pierwotnej do specyfiki kultury lokalnej. Co bardzo istotne, jeśli chcemy mieć pewność co do możliwości trafnej i rzetelnej interpretacji wyników testowych, w przypadku testów pochodzących z innego obszaru kulturowego trzeba przeprowadzić nie tylko procedurę tłumaczenia oryginalnych pozycji. Przenoszenie narzędzi diagnostycznych z jednej kultury do drugiej wiąże się bowiem z koniecznością zadania sobie pytania: jak sprawić, aby narzędzie trafne i rzetelne w jednej kulturze również dobrze diagnozowało w innej? Tym samym, adaptując narzędzie trzeba jeszcze wykazać, że metoda adaptowana mierzy ten sam konstrukt, co metoda wyjściowa, że robi to z podobną rzetelnością. Trzeba również przygotować normy dla populacji lokalnej. W przeciwnym razie może się okazać, że metoda adaptowana prowadzi do wyników o niskiej jakości psychometrycznej (Anastasi i Urbina, 1999; Hornowska, 2005).

Warto w tym miejscu wspomnieć jeszcze o innych aspektach adaptacji kulturowej, a więc po pierwsze o wpływie czasu na jakość psychometryczną testów. Szczególnie w przypadku e-uzależnień treść pozycji testowych nosi ślad epoki, w której powstawały. To oznacza, że konieczne jest regularne sprawdzanie, czy zadania mają dla obecnej populacji ten sam sens, co wcześniej. Jeśli nie, test powinien przejść procedurę adaptacji kulturowej. Podobnie, młodzież może podejmować inne zachowania jako przejawy e-uzależnienia niż

młodsze dzieci i odwrotnie, dlatego też przy używaniu testów przygotowanych dla jednej grupy wiekowej do badań innej grupy wiekowej istnieje duże ryzyko pomiaru o niskiej jakości. I w takiej sytuacji test powinien być zaadaptowany do nowych warunków, co oznacza również zebranie empirycznych danych potwierdzających trafność i rzetelność nowego zastosowania testu oraz danych normalizacyjnych.

W tabeli 19 zawarto kryteria dobroci testów psychologicznych i wskazówki pozwalające specjalistom sprawdzić, czy dane narzędzie spełnia podstawowe warunki psychometryczne i może zostać użyte do skutecznego pomiaru.

Tabela 19

Kryteria dobroci testu i sposoby ich weryfikacji - jak sprawdzić, czy dane narzędzie może zostać użyte do skutecznego pomiaru wskaźników e-uzależnienia?

Kryterium – podstawowe właściwości psychometrycz ne	TAK / NIE	Jak sprawdzić:
Obiektywność i standaryzacja		Czy dostępny jest podręcznik do testu? Czy podręcznik zawiera informacje dotyczące procedury badania testem, obliczania wyników i procedury ich interpretowania? Czy dostępny jest oryginalny arkusz do testu z instrukcją, kompletem pozycji testowych (pytań lub stwierdzeń) i kluczem odpowiedzi?
Rzetelność		Czy autorzy podają wyniki analizy rzetelności (najlepiej wyniki badania test – retest)? Czy analizy te zostały wykonane na podstawie danych zebranych w odpowiednich grupach badanych? Czy wyniki tych analiz są satysfakcjonujące?

	Czy autorzy podają standardowy błąd pomiaru i przedział ufności do interpretacji wyniku testowego?
Trafność	<p>Czy autorzy jasno określają, do jakich celów test może być stosowany i jak należy interpretować jego wyniki w oparciu o teorię, która stanowi podwaliny testu i o uzyskane wyniki analizy trafności?</p> <p>Czy autorzy podają wyniki zróżnicowanych analiz trafności?</p> <p>Czy analizy te zostały wykonane na podstawie danych zebranych w odpowiednich grupach badanych?</p> <p>Czy wyniki tych analiz są satysfakcjonujące (potwierdzają postawione hipotezy badawcze)?</p>
Normalizacja	<p>Czy test umożliwia porównanie wyniku surowego do rozkładu populacji odniesienia (zwłaszcza tej samej grupy wiekowej), to znaczy, czy dostępne są normy ilościowe dla populacji odniesienia?</p> <p>Czy normy te mają charakter standardowy (są oparte o rozkład normalny)?</p> <p>Jeśli obliczono normy w podgrupach (np. osobno dla płci lub badanych o różnym wieku), to czy uzasadniono ich wyróżnienie?</p>
Adaptacja kulturowa	<p>Jeśli test nie został skonstruowany w Polsce, to czy opisano procedurę tego tłumaczenia? Czy zaprezentowano dane dotyczące trafności oraz rzetelności i czy porównano je z odpowiednimi danymi dla wersji oryginalnej?</p> <p>Czy przeprowadzono normalizację zaadaptowanej wersji?</p>

Źródło: opracowanie własne na podstawie literatury (Anastasi i Urbina, 1999; Hornowska, 2005).