Psychometria - notatki do egzaminu

Psychometria (SWPS Uniwersytet Humanistycznospołeczny)

Psychometria

Metody badawcze w psychologii akademickiej
• Metody eksperymentalne
• Metody korelacyjne
W obu metodach konieczne jest trafne i rzetelne mierzenie zmiennych.

Badania psychologiczne stosowane
• Metody eksperymentalne
• Metody korelacyjne
• Metody diagnostyczne w diagnostyce

Narzędzia diagnozy psychologicznej
Klasyfikacja ze względu na stopień formalizacji procedury:
• Diagnoza swobodna (intuicyjna, "das Praecox Gefühl")
• Testy projekcyjne -zachowania znakami
• Testy standaryzowane -testy próbkami zachowania

Diagnoza kliniczna: „procedura stosowana przez lekarzy i psychologów praktyków, w której diagnosta zestawia razem różne dane i wyciąga wnioski, stosując metody nieformalne, subiektywne.

Diagnoza algorytmiczna: Sformalizowana procedura, w której wnioski wyprowadza się z danych za pomocą ścisłego algorytmu”.

Funkcje testów
Testy są metodami, służącymi do zbierania informacji o człowieku, niezbędnej do diagnozy psychologicznej.
Diagnoza psychologiczna jest procesem aktywnego poszukiwania danych potrzebnych do podjęcia działań, których celem jest zmiana aktualnego stanu psychospołecznego ludzi (np. terapii, porady, interwencji, decyzji administracyjnej, itp.).

Co testy mierzą?
Testy służą do rejestracji danych psychologicznych, tj. informacji o zachowaniach osób: przeszłych lub aktualnych, typowym (stałym lub najczęstszym -cechy) lub chwilowym (stany).
Testy są metodami rejestrującymi różnice w zachowaniu ludzi oraz stałość czasową i między sytuacyjną zachowań pojedynczych osób.

Czym jest test?
Test= ścisła procedura diagnozowania.
Zbiór zadań lub pytań, które -w standardowych warunkach (identycznych dla każdego testowanego dla wszystkich) -wywołują określone rodzaje zachowań, które można zebrać w postać wskaźnika liczbowego o pożądanych własnościach psychometrycznych, tj. posiadających wysoką rzetelność i wysoką trafność.
SKALA DEPRESJI BECKA - Ocena dotyczy ostatniej doby

Elementarne składniki testu

1) Zadania testowe.

2) Procedura.

Test psychologiczny to standardowe bodźce,e na które reaguje diagnozowany w obecności osoby badającej, tak więc test psychologiczny to także forma interakcji między osobą badającą a osobą badaną, interakcja ta zachodzi w określonych warunkach czasowych, fizycznych i psychologicznych (procedura i materiał są ze sobą nierozerwalnie związane).

3) Skala (skale) standardowe pozwalające ocenić wywołane przez bodźce (zadania) reakcje (zachowania).

Rejestrowane zachowania:

• są egzemplifikacją cechy,

• są typowe dla bodźców występujących w teście (np. typowe jest układanie puzzli, ale nie ich zjadanie)

• podają się ocenie ilościowej, a nie jakościowej (choć tu problem ze skalami nominalnymi dwukategorialnymu; 0-1; tak-nie; ma-nie ma; rozwiązał-nie rozwiązał itd., bo)

4) Reguł dochodzenia do wyniku testu oraz reguły oceny tego wiarygodności tego wyniku.

Tak więc reguły te regulują wnioskowanie o wyniku testu w aspekcie:

• indywidualnym -o nasileniu danej cechy u danej jednostki;

• pomiarowym -o "dobroci" pomiaru testem.

Kardynalne cechy dobrego testu:

Testy muszą być tak skonstruowane, aby ich zastosowanie do pomiaru cech człowieka, charakteryzowało się odpowiednio wysoką:

rzetelnością, trafnością, obiektywnością, standaryzacją i znormalizowaniem

Test musi spełniać wszystkie te wymagania, by mógł być stosowany, i żeby mógł być nazwany testem.

Kryterium normalizacji

Normalizacja oznacza, że na potrzeby testu opracowano procedury przekształcania wyników liczbowych (surowych) w wyniki różnicowe (standaryzowane), umożliwiające ocenę wyniku danej osoby na tle populacji, z której ona pochodzi.

Kryterium standaryzacji

Standaryzacja oznacza, że na potrzeby testu opracowano ścisłą procedurę jego użycia. Procedura ta musi być na tyle szczegółowa i konkretna, żeby wykonywano go za każdym razem identyczne -niezależnie od osoby badającej i osoby badanej;.

Celem standaryzacji jest zminimalizowanie wpływu czynników zakłócających na wyniki testu.

Kryterium obiektywności

Obiektywność oznacza, że na potrzeby testu opracowano ścisłą i szczegółową procedurę interpretacji uzyskanych w badaniu nim wyników (zachowań badanego). Interpretacja ta musi być za każdym razem taka sama niezależnie od osoby badającej i/lub interpretującej

wyniki.
Celem obiektywności jest zminimalizowanie wpływu osoby badającej/interpretującej na wyniki testowe.

Kryterium trafności (wiarygodności)
Trafność oznacza, że na potrzeby testu opracowano materiały testowe (zadania) i procedury badania, które maksymalizują stopień, w jakim test mierzy daną (docelowo mierzoną) cechę. Trafność ma szereg ważnych aspektów, o których mówić będziemy na stosownych wykładach.

Kryterium rzetelności (dokładności)
Rzetelność oznacza, że na potrzeby testu opracowano materiały testowe (zadania) i procedury badania, które maksymalizują dokładność pomiaru cechy badanej przez test

• Test=Miara -jedno- lub wielopozycyjna skala lub wskaźnik (najogólniejsze określenie)
• Skala -opublikowana, posiadająca swoją nazwę własną miara lub zbiór miar
• Bateria - kolekcja miar różnego pochodzenia
• Kwestionariusz-narzędzie samoopisowe
• Ankieta= Kwestionariusz

Przykład baterii
• Bateria Testów APIS-Z
• APIS-Z jest wielowymiarową baterią służącą do pomiaru inteligencji ogólnej.

Definicja testu psychologicznego
• Test psychologiczny jest to narzędzie
– pozwalające na uzyskanie takiej reprezentatywnej próbki zachowań, o których można przyjąć (np. na podstawie założeń teoretycznych lub związków empirycznych), że są one wskaźnikami interesującej nas cechy psychologicznej;
– jest to narzędzie obiektywne, wystandaryzowane, trafne, rzetelne i znormalizowane;
–Jest to narzędzie wyposażone w reguły obliczania wartości mierzonej cechy psychologicznej oraz jasno określające zakres i rodzaj dopuszczalnych zachowań ze strony diagnosty.
– Nadto, badanie testowe to taka sytuacja, w której osoba badana uczestniczy dobrowolnie, świadoma celu, jakim jest jej ocena.

Rodzaje testów
1) Standaryzowane vs. niestandaryzowane
• Testy standaryzowane to takie testy, które posiadają dokładnie sprecyzowane zasady ich stosowania oraz zostały wyposażone w normy.
• Testy niestandaryzowane - np. testy wiadomości budowane przez nauczycieli na ich własny użytek, dopuszcza się w nich możliwość modyfikowania sytuacji badania. Często stosowane tylko jeden raz.
2) Indywidualne vs. grupowe
3) Szybkości vs. mocy -kryterium czasowe, test mocy
• Test szybkości -czas rozwiązania jest ograniczony i -z założenia -żadna osoba badana nie powinna go ukończyć w wyznaczonym czasie. Składa się z zadań średnio trudnych, tj. takich, które przy braku presji czasu rozwiązuje większość osób, dla których przeznaczony

jest test.
• Test mocy (t. zdolności potencjalnych) stwarza każdej osobie badanej szansę na podjęcie próby rozwiązania wszystkich pozycji testu. Trudność zadań stopniowo rośnie. Na końcu zadania skrajnie trudne, które bardzo nieliczne osoby rozwiązują. Jedynie niewielka liczba osób otrzymuje w takim teście maksymalny wynik.

4) Obiektywne vs. nieobiektywne -kryterium klasyfikacji jest sposób obliczania wyników w teście.

• Testy obiektywne posiadają starannie opracowane, standardowe metody obliczania wyników, wynik testu słabo (lub wcale) zależy od umiejętności diagnosty (najczęściej procedura obliczania wyników polega na przyłożeniu szablonu do arkusza odpowiedzi).
• Testy nieobiektywne: ocena odpowiedzi osoby diagnozowanej wymaga dużej znajomości przedmiotu pomiaru i otrzymany wynik często odzwierciedla subiektywne umiejętności psychologa.

5) Słowne vs. bezsłowne -klasyfikowane ze względu na rodzaj zadań

• Testy słowne na wynik testu wpływa sprawność językowa osoby badanej.
• Testy bezsłowne na wynik testu nie wpływa sprawność językowa osoby badanej.
Wykonywanie testu = wykonanie pewnych czynności. Słabe opanowanie języka nie jest czynnikiem wpływającym istotnie na otrzymywane wyniki.

6) Właściwości poznawcze vs. właściwości afektywne

•Testy właściwości poznawczych testy mierzące wytwory procesów poznawczych (np. testy zdolności, uwagi, pojemności pamięci).
•Testy właściwości afektywnych testy mierzące emocje, postawy, wartości, zainteresowania.
•Testy cech osobowości (czyli niemal zachowań)

7) Zorientowane na normy vs. zorientowane na kryterium (kryterium podziału stanowi sposób interpretacji wyników).

Testy zorientowane na normy punktem odniesienia jest przeciętny poziom wykonania testu w określonej grupie odniesienia; tym samym normatywna interpretacja wyniku testowego jest interpretacją relatywną i zależy od tego, kto wchodził w skład badanej grupy osób.
Testy zorientowane na kryterium punktem odniesienia jest konkretny zakres wiedzy. (np.. 50% poprawnych odpowiedzi daje ocenę dostateczną)

Rodzaje testów

Z uwagi na sposób zbierania informacji (pomiaru) rozróżnia się:

1) Testy wykonania.

Test, w którym osoba diagnozowana wykonuje pewne zadanie. Wynik testu jest tym wyższy im lepiej osoba radzi sobie z zadaniem. Test taki wymaga maksymalnego (np. testy inteligencji, zdolności, wiadomości, sprawności psychomotorycznej).

2) Kwestionariusze=miary samoopisowe.

Test, w którym osoba diagnozowana sama opisuje swoje zachowanie, reakcje lub preferencje -rejestrowany jest samopis typowego zachowania (np.. kwestionariusze, inwentarze, ankiety, arkusze biograficzne). Wynik testu jest tym wyższy im diagnozowany silniej identyfikuje u siebie diagnozowaną cechę.

3) Testy zachowania.

Test, w którym zgodnie w procedurą obserwuje się osobę diagnozowaną w standardowych okolicznościach. Rejestrowane są dane o typowych zachowaniach i/lub poziomie wykonania w naturalnych warunkach (próbki pracy zawodowej, arkusze obserwacyjne, arkusze ocen,

arkusze szacowania cech, listy diagnostyczne/kontrolne (checklists).

Testy jedno- i wielowymiarowe
1) Testy jednowymiarowe –interpretacja w terminach intensywności cechy (diagnoza ilościowa) albo interpretacja typologiczna (diagnoza jakościowa).
2) Testy wielowymiarowe – interpretacja profilu testowego (model cech równorzędnych lub model hierarchiczny), który może być wykorzystany w diagnozie typologicznej lub intraindywidualnej (różnicowej).

Intraindywidualna diagnoza różnicowa
• Porównania wyników poszczególnych testów z normami oraz porównania wyników poszczególnych testów ze sobą nawzajem np. werbalny i niewerbalny I.I.

Jakie są zastosowania testów?
Testy są stosowane w badaniach naukowych i praktycznych:
a) przydatności zawodowej, która wiąże się z:
-doborem zawodowym,
-poradnictwem zawodowym;
b) diagnostyce klinicznej;
c) diagnostyce zdolności specjalnych.

Zalety i wady testów
Testy „reprezentują najbardziej wartościową i sprawiedliwą technologię, umożliwiającą podejmowanie wielu ważnych decyzji o ludziach", ale jednocześnie „testowanie psychologiczne jest bardzo kontrowersyjne" (Murphy i Davidshofer, 1989, s. 2).

Zastosowania testów
• badania naukowe
• badania przydatności zawodowej
-dobór zawodowy,
-poradnictwo zawodowe;
• diagnostyka dojrzałości szkolnej
• diagnostyka kliniczna;
• diagnostyka zdolności (cech) specjalnych

Uwarunkowania diagnostyki testowej (algorytmicznej)
Diagnostyka psychologiczna za pomocą testów standaryzowanych uwarunkowana jest kontekstami:
• psychologicznym
• metodologicznym
• psychometrycznym
• etycznym

Kontekst psychologiczny diagnozy
Związany ze znaczeniem jaki psychologia jako nauka nadaje konkretnemu pomiarowi (bez teorii pomiary niemal nic "nie znaczą") i są to:
1) problem badawczy wyznaczonego przez cel diagnozy;

2) model (teoria) psychologiczny, odnoszący się do cech mierzonych testem;
3) wnioski formułowane na podstawie wyników pomiaru testowego.

Test psychologiczny jest tego rodzaju narzędziem, które wymaga nie tylko znajomości samej procedury stosowania, ale również znajomości leżącej u jego podstaw teorii psychologicznej oraz teorii psychometrycznej, decydującej o sposobie ilościowej interpretacji wyników tego testu.

Kontekst psychometryczny diagnozy
Określa zasady poprawności wnioskowania o wartości diagnostycznej dokonanego pomiaru.

poziom zmierzony < > poziom prawdziwy

Kontekst metodologiczny diagnozy
Określa formalne zasady stosowania testu:
• procedury badania,
• obliczania wyników
• standaryzacji
• interpretacji uzyskanych danych
Określa zasady poprawności wnioskowania diagnostycznego.

Kontekst etyczny diagnozy
Fizyczne, psychologiczne i społeczne konsekwencje pomiaru testem dotykające osoby diagnozowanej.
Wyniki testów psychologicznych są podstawą decyzji o ważnych społecznie konsekwencjach.
• Wyniki testów nie tylko służą teoretycznym dyskusjom psychologów, ale są również wykorzystywane do kształtowania polityki społecznej, i w ten sposób mogą wpływać istotnie na losy ludzi.

W potocznym odbiorze testy zatraciły opinię obiektywnych miar i często traktuje się je jako niebezpieczne i niesprawiedliwe narzędzie uzyskiwania przewagi przez wtajemniczonych profesjonalistów (selekcjonujących ludzi i działających bez społecznego przyzwolenia) nad zwykłymi obywatelami (uczniami, kandydatami do pracy, pacjentami, klientami sądów, itd.)
Użyteczność testów (ilorazu inteligencji, ale również testów na koniec nauki i w charakterze egzaminów wstępnych) w przewidywaniu powodzenia w nauce szkolnej jest jedną z najważniejszych przyczyn ich popularności.
Jednak testy takie są zwykle kulturowo obciążone i nie uwzględniają specyficznego pochodzenia kulturowego osób należących do mniejszości etnicznych i o niskim statusie społeczno-ekonomicznym. Co gorsza istnieją środowiska przywiązane do interpretacji, że gorsze wyniki testowania są rezultatem gorszego wyposażenia genetycznego.

Co jest w takim razie rozwiązaniem?
Odp: Unikanie błędów i niekorzystnych następstw testowania. Wg Hornowskiej, bezrefeleksyjne stosowanie testów może prowadzić do wielu niekorzystnych zjawisk społecznych:

1) Rozumienia inteligencji jako jedynej lub głównej cechy warunkującej powodzenie w

bardzo wąsko definiowanych zadaniach. Współcześnie zwraca się uwagę np. na znaczenie inteligencji emocjonalnej, interpersonalnej, społecznej, twórczej czy praktycznej.
Wg Hornowskiej, bezrefeleksyjne stosowanie testów może prowadzić do wielu niekorzystnych zjawisk społecznych:

2) Etykietowanie i stygmatyzowanie w zakresie statusu intelektualnego. Psycholog stygmatyzuje ludzi, jeśli „w stawianych przez siebie diagnozach przypisuje im pewne etykiety, jeśli naznacza ich jakimiś społecznie pejoratywnymi właściwościami i naraża na szwank ich poczucie własnej wartości i godności. W diagnozach tych w sposób jawny lub ukryty występuje element wartościowania jednostek i grup społecznych, jeśli wskazuje, w jakim stopniu i pod jakim względem ich społeczne zachowania są niepożądane, szkodliwe, nienormalne, słowem: zakazane" (Poznaniak, 1994, s. 73).

3) Przypisywania psychologom roli osób kontrolujących i determinujących losy życiowe badanych osób.
•Podstawą tego społecznego zwyczaju jest przypisywanie narzędziom stosowanym przez psychologów cechy bezwarunkowego obiektywizmu. Dodatkowym uzasadnieniem jest też to, iż wyniki badań psychologicznych podawane są w liczbach: stwierdzenie „wysoka inteligencja" czy „wysoki poziom niepokoju" wydają się ludziom o wiele mniej precyzyjne niż I.I. = 118 .

4) Biurokratyczne podejmowanie decyzji dotyczących oceny badanych osób.
•Umiejętności psychologicznych nie należy ograniczać do automatycznego odczytywania norm dostępnych w podręcznikach testowych.

Prawa osób diagnozowanych
1) Prawo do wyrażenia świadomej zgody na badanie testem.
Osoby badane mają prawo wiedzieć:
–dlaczego są testowane,
–jakie informacje o wynikach testowania i komu zostaną następnie udostępnione,
–w jaki sposób będą wykorzystane wyniki. Informacje takie należy przekazywać w sposób zrozumiały dla osób badanych i na tej podstawie uzyskiwać zgodę na badanie testowe. Jednak nie należy wcześniej pokazywać badanemu pozycji testowych ani informować go, jak będą oceniane określone odpowiedzi. Udzielenie tego rodzaju informacji unieważnia zazwyczaj test.
2) Prawo do informacji o wynikach testowania.
1. Sposób przekazania informacji o wynikach uzyskanych w teście musi być dostosowany do możliwości osób badanych. Informacje takie nie powinny być przekazywane rutynowo, a powinny dostarczać zindywidualizowanych wyjaśnień interpretacyjnych.
2. Przekazywanie informacji o wynikach testowych osobom trzecim lub instytucjom powinno mieć miejsce tylko wtedy, kiedy stoją za tym racje merytoryczne.
3. Wyniki testowe powinny być one przekazywane jedynie osobom, które mają wystarczające kwalifikacje, aby je zinterpretować. Należy zadbać także o to, aby przekazywać informacje w taki sposób, który nie będzie prowadził do błędnych interpretacji.
3) Prawo do minimalizowania skutków etykietowania.
1. Zgodnie ze Standardami... (1985a, s. 86), opisując wynik osoby badanej, należy posługiwać się takimi określeniami, które w minimalnie możliwym stopniu etykietyzują osobę badaną.

2. Przygotowując orzeczenie psychologiczne, należy unikać stosowania skrótowych etykietek. Stosowanie takich określeń zawsze wiąże się z wartościowaniem. Dlatego osoby przygotowujące interpretacje wyników testowych powinny starannie określać znaczenie stosowanych terminów i dbać o to, by ci, do których trafi taka interpretacja, nie nadawali jej innego znaczenia.

4) Prawo do zachowania tajemnicy o wynikach testowania.

1. Dane z badań mogą być udostępniane innym osobom tylko po świadomym wyrażeniu na to zgody.
2. Wg Anastasi i Urbina (1999, s. 681): „podstawowa zasada głosi, że protokołu (z badań psychologicznych) nie należy ujawniać bez wiedzy i zgody badanego, chyba że jest to z uzasadnionych powodów wymagane lub dopuszczane prawem".
3. Prawo do zachowania tajemnicy o wynikach testowania oznacza również obowiązek odpowiedniego zabezpieczenia danychprzez osoby stosujące testy. Dotyczy to zarówno danych przechowywanych w postaci fizycznej (np. papierowych protokołów), jak i w formie elektronicznej.

5) Prawo do prywatności.

1. Wg Poznaniaka (2000), każdy psycholog powinien zdawać sobie sprawę, że zadawane przez niego pytania mogą naruszyć sferę prywatności i że musi się on dobrze zastanowić, zanim zacznie je zadawać, a klient (osoba badana) ma prawo do odmowy odpowiedzi na pytania zadane mu przez psychologa.

Testy w postępowaniu sądowym

• Uwrażliwienie na fakt, że różnice w wynikach testowych nie odzwierciedlają w pełni rzeczywistych różnic w poziomie mierzonej cechy.
• Opinia publiczna oczekuje gwarancji, że decyzje podejmowane na podstawie wyników testowych są „uczciwe". Ponieważ takie gwarancje nigdy nie będą bezwarunkowe, testy i testowanie nie budzą społecznego zaufania.
• Sprawy sądowe, w których oskarżano testy, toczyły się wielokrotnie w stosunku do testów I.I. I testowej rekrutacji pracowników. Najczęstszym wówczas zarzutem był zarzut dyskryminacji rasowej, której szukano w wynikach testowych.
• Nieaktywna Ustawa o zawodzie psychologa i samorządzie zawodowym psychologów regulowała ten problem. Prawo do stosowania testów psychologicznych i do orzekania na podstawie ich wyników mieli dyplomowani psychologowie. Miało to wyeliminować z rynku nieprofesjonalistów, stosujących bez zastanowienia testy psychologiczne przy każdej okazji.

Testy– potrzeby rynku, oczekiwania klientów Wg Sternberga (1992) i Morelanda (1995) oraz Standardów APA (1985), współczesny klient

- przekonany o społecznej zasadności testowania
-chciałby, aby testy psychologiczne gwarantowały:

• przewidywanie osiągnięć, • stabilność wyników,
• właściwą normalizację i standaryzację,
• łatwość stosowania,
• łatwość interpretacji,
• obiektywną punktację,
• brak stronniczości,
• uzasadnione koszty stosowania,
• ochronę wyników,
• sądową kontrolę decyzji administracyjnych

-wyniki testów psychologicznych muszą się dawać obronić, gdyby decyzje podjęte na ich podstawie trafiły do sądów.

Pomiar
• Pomiar pewnej cechy będącej właściwością obiektów z określonego zbioru, to nic innego, jak przyporządkowanie tym obiektom wartości liczbowych w taki sposób, żeby odpowiednie relacje zachodzące między liczbami odzwierciedlały interesujące badacza relacje między obiektami, wynikające z posiadanej przez nie cechy.

Pomiar różnicowego w psychologii
W psychologii (a pewnie i w ogóle) pomiar możliwy jest tylko poprzez porównanie mierzonego obiektu ze standardem. W psychologii jedynym dostępnym standardem są inni ludzie a więc możliwy jest tylko pomiar różnicowy, który pozwala na lokowanie człowieka (pod względem jakiej cechy) na kontinuum tej cechy tle innych osób.

Pojęcie cechy
Cecha – zmienna osobowa, która wykazuje międzyosobniczą zmienność i wewnątrzosobniczą stałość (czasową i sytuacyjną) oraz spójność wskaźników.
Dla cech możliwy pomiar w skali co najwyżej przedziałowej.

Ze względu na rodzaj skali, w której zmienna jest mierzona po operacjonalizacji (tu już świat empirii) zmienne można podzielić na :
•zmienne nominalne
•zmienne porządkowe: A>B>C
•zmienne interwałowe: A=B-2=C-4; punkt zerowy jeśli jest, to jest umowny (stopnie C)
•zmienne ilorazowe: A=B/2=C/4; punkt zerowy skali w wyznaczonym przez Naturę miejscu (stopnie K)
•Ponieważ w psychologii każdy pomiar jest pomiarem różnicowym maksymalnym dostępnym poziomem pomiaru jest skala **interwałowa.**

1. **Skala nominalna.** Jeżeli dwa obiekty różnią się wartością cechy, to reprezentacja liczbowa ich cechy poprzez pomiar da dwie różne liczby (np. 1 i 2; 3 i 6).
2. **Skala porządkowa.** Jeżeli jeden obiekt ma większe natężenie danej cechy od drugiego, to pomiar dostarczy odpowiednio liczb: większej oraz mniejszej.
3. **Skala przedziałowa.** Jeżeli możliwe jest porównywanie różnicy między natężeniem cechy dwóch obiektów, to pary obiektów o tej samej różnicy natężenia cechy, za pomocą pomiaru zostają odzwierciedlone w pary liczb różnice się między sobą o taką samą wartość.
4. **Skala ilorazowa.** Jeżeli można stwierdzić, że jeden obiekt ma k-krotnie większe natężenie cechy od drugiego, to wartości liczbowe dostarczone pomiarem powinny także być związane tą relacją.

Założenie o normalności rozkładu cech podstawą klasycznej teorii testów
Zakłada się, że cechy psychologiczne mają rozkład normalny w populacji. Na podstawie tego założenia „krzywa normalna” jest traktowana jako model rozkładu wyników testu.

Rozkładu normalny
Rozkład symetryczny, wykazujący odpowiednie zagęszczenie wyników wokół średniej (skośność miara asymetrii oraz kurtoza -miara zagęszczenia)

Jak konstruuje się psychometryczną „linijkę”?
1. Decyzja o potrzebnym poziomie pomiaru.
•Czasami potrzebujemy tylko odpowiedzi na pytanie, czy osoby są identyczne pod względem cechy lub kto jest lepszy

Ale
•najczęściej jednak chcielibyśmy odpowiedzi na pytania o dystans dzielący osoby i wtedy potrzebujemy pomiaru interwałowego.

Jak konstruuje się idealną interwałową psychometryczną „linijkę”? Test mocy.
1. Przyjęcie założenia, że cecha w populacji ma rozkład normalny.
2. Przyjęcie założenia o liniowym (a w każdym razie monotonicznym) związku wyników w teście z wartościami prawdziwymi „wynikami prawdziwymi”.
3. Przyjęcie założenia, że wszystkie tworzone zadania testowe mierzą dokładnie tą samą cechę (np. zdolności matematyczne). 3. Przygotowanie dużej puli takich pytań różniących się trudnością.
4. Pobranie dużej próbki losowej osób z populacji. I danie tej próbce zadań do wykonania.
5. Powtórzenie pkt. 4 dla wielu zadań o różnej trudności. Za każdym razem zadania dzielą próbkę na 2 grupy osób:
•tych którzy je rozwiązali (było dla nich łatwe, mieli nad nim zadanie)
•tych którzy nie dali rady go rozwiązać (było dla nich za trudne)
•Jeśli podzielimy liczebności tych grup przez liczebność próbki otrzymamy dla każdego testu pi i qi=(1-pi)
•pi = proporcja osób, które rozwiązały zadanie
•qi=(1-pi) = proporcja osób, które nie rozwiązały zadania

Dla każdego zadania empirycznie zostaje wyznaczona jego trudność pi, która jednocześnie określa położenie osób testowanych na wymiarze zdolności.
•Ci którzy „biją” takie zadanie mają wyższe od potrzebnych do jego rozwiązania zdolności.

Jeśli dobierzemy zadania tak, by ich wyniki standaryzowane oddalone były wzajemnie od siebie o równe interwały otrzymamy test mierzący cechę (tu zdolność) w skali interwałowej.

Jak konstruuje się idealną interwałową psychometryczną „linijkę”? Test szybkości i skale likertowskie.
Dla testów szybkości i testów cech osobowości i postaw opartych o szacowanie (w tym likertowskie) nie ma tak spójnej i osadzonej w teorii pomiaru, jak przedstawiona dla testów mocy metody budowania pomiaru w skali interwałowej. W tego typu testach unika się pozycji neutralnych (używa się pozycji łatwych (testy szybkości) lub wyraźnie nacechowanych (testy osobowości i miary postaw). Zakłada się również, że pozycje te są w jednakowym stopniu trudne (lub nacechowane) i w identycznym stopniu powiązane z mierzoną cechą.

Następnie zakłada się, że osoby mające wyraźną przewagę (dominujące) nad pozycjami radzą sobie z nimi lepiej.
•W testach szybkości rozwiązuje je szybciej.
•W testach z szacowaniem (likertowskim) dają bardziej spójne (i skrajne) odpowiedzi.

W konsekwencji zsumowanie odpowiedzi z wielu takich pozycji daje dla każdej osoby jej pozycję na kontinuum a odległości pomiędzy wynikami osób są mierzone **przedziałowo** a jednostki pomiaru to:
•wykonane zadanie (testy szybkości)
•odpowiedź diagnostyczna (testy z odpowiedziami tak/nie)
•odległość pomiędzy punktami skali szacunkowej (testy oparte o szacowanie likiertowskie).

Jednak powyższy pomiar jest możliwy tylko o tyle o ile:
•cecha rozkłada się w populacji
zgodnie z rozkładem normalnym. A więc i w konsekwencji wyniki uzyskane (testowe) też (dla dobrze zbudowanego testu) mają rozkład normalny,
•pozycje narzędzia są jednakowo trudne i w jednakowym stopniu związane z mierzoną cechą,
•w skalach likertowskich opcje odpowiedzi są od siebie oddalone o równe interwały,

Wniosek
Do pomiaru zmiennych psychologicznych w skali przedziałowej konieczne jest:
1. Przyjęcie założenia o normalnym rozkładzie wartości tej zmiennej wśród osób z populacji, która podlega opracowaniu (m. In. z tego powodu testy da zawsze dla jakiejś populacji).
2. Wykorzystanie zmienności międzyosobniczej jako jednostki miary.
3. Przygotowanie wielu pozycji (testów).

Aby test dostarczał wyników ilościowych, tzn. informacji o natężeniu mierzonej właściwości (cechy) konieczne jest wprowadzenie wielu zadań do testu.
Zabieg ten umożliwia:
-pomiar interwałowy (precyzyjne różnicowanie wyników osób badanych)
-kontrolowanie błędów losowych
-pomiaru diagnozę cech definiowanych jako źródło współwystępowania zachowań
-ogólność psychologiczną pomiaru (mimo konkretności pozycji)

Pomiar ilościowy
Mimo, że najczęściej pozycje testowe dostarczają danych mierzonych w skali nominalnej, to zsumowanie wyników wszystkich pozycji daje wynik ogólny testu, który (gdy test jest dobrze skonstruowany) daje pomiar w skali przedziałowej.

Błąd pomiaru
Prawidłowa (diagnostyczna) odpowiedź na pojedynczą pozycję może być odgadnięta lub zupełnie przypadkowa. Im więcej pozycji ma test tym mniejsza szansa (prawdopodobieństwo) na to, że osoba testowana uzyska w nim wysoki wynik wyłącznie dzięki zgadywaniu lub przypadkowi (prawdopodobieństwo prawidłowego odgadnięcia rozwiązania w pojedynczym teście 0-1, to oczywiście 50%, ale prawidłowe odgadnięcie wszystkich

Współwystępowanie zachowań
Cecha jest własnością funkcjonowania człowieka, która przejawia się w różnych zachowaniach.
Cechy definiuje się jako właściwości determinujące współwystępowanie pewnego zbioru zachowań. Konkretne pojedyncze zachowanie bywa egzemplifikacją dla wielu cech, ale kombinacja (zespół współwystępujących) wielu zachowań jednoznacznie określa poziom

nasilenia danej cechy.

Rozwiązania pozycji a wynik testu
Wynik testu jest kombinacją wyników wszystkich pozycji testowych. Najczęściej jest to suma wyników pozycji:
• nieważona (wszystkie pozycje równoważne), np. testy szybkości
• ważona (trudne lub wyrażające silniejsze natężenie cechy pozycje z wyższymi wagami), np. testy mocy. Tak więc własności testu jako całości zależą bezpośrednio od wyników poszczególnych pozycji (oraz ich interkorelacji).

Średnia testu a średnia pozycji
Wynik testu jako suma wyników poszczególnych pozycji
(wyniki 0-1: rozwiązane błędnie - dobrze)

• średnia wyników testu równa jest sumie średnich pozycji.
• pozycja dodana do testu powoduje wzrost średniej wyników testu

Wariancja testu a wariancja pozycji
Wynik testu jako suma wyników pozycji (wyniki 0-1)

Współczynnik korelacji r-Pearsona
Korelacja jest miarą współzmienności (związku dwóch zmiennych)

Uwaga
Wariancja całkowita testu jest równa sumie wariancji pozycji oraz ich podwojonych kowariancji.
Nowa pozycja dodana do testu zwiększa wariancję całkowitą tylko wtedy, gdy wariancja pozycji nie jest równa zero. Nowa pozycja dodana do testu zwiększa znacznie wariację całkowitą, jeśli kowariancje nowej pozycji z innymi pozycjami są dodatnie. Tylko pozycje z niezerowymi wariancjami oraz dużymi pozytywnymi kowariancjami powinny być dodawane do testu, bo:
• uzyskanie dużego zróżnicowania wyników testu jest celem pomiaru różnicowego
• dodatnie kowariancje pozwalają przyjąć, że pozycje mierzą jedną i tę samą cechę

Rzetelność pomiaru testowego. Podstawy teorii rzetelności testów psychologicznych
W języku potocznym, rzetelność oznacza niezawodność (dokładność) pomiaru.
W psychometrii rzetelność odnosi się do powtarzalności (stałości) wyników otrzymywanych tym samym testem w ujednoliconych (standaryzowanych) warunkach warunkach badania.

Błąd pomiaru testem
Pojęcie rzetelności jest powiązane z pojęciem błędu pomiaru:
1. nie możliwy jest pomiar bez błędny,
2. im większy jest błąd, tym mniejsza jest rzetelność pomiaru danym narzędziem.
Błąd obniża precyzję pomiaru, bo jest zniekształcany przez błąd, zamiast wyniku prawdziwego uzyskuje się wynik zaniżony lub zawyżony.

Źródła błędów pomiaru

• Ogólne cechy osoby badanej (np. ogólny styl rozwiązywania testów lub jej zdolność rozumienia instrukcji)
• Specyficzne cechy osoby badanej dotyczące danego testu jako całości (np. umiejętności specyficzne dla danego testu lub typu pozycji nań się składających).
• Ogólne właściwości osoby badanej przypadkowo zbieżne z testowaniem (np. zdrowie, zmęczenie, motywacja, lęk).
• Specyficzne właściwości osoby badanej związane z badaniem testowym (np. techniki radzenia sobie z zadaniami, rozumienie pewnych typów zadań, poziom wyćwiczenia specyficznych umiejętności).
• Specyficzne właściwości osoby badanej związane z pozycjami testowymi (np. wahania w koncentracji uwagi).

Systematyczne i incydentalne okoliczności związane z testowaniem:

• warunki testowania (np. obecność dystraktorów, jasność instrukcji, łatwość dostosowania się do limitu czasu, UWAGA: te źródła błędów powinno się wyeliminować stosując standardową procedurę testowania),
• interakcja płci, osobowości osoby badanej i badającej, itp., zniekształcenia w ocenie zachowania oraz
• zdarzenia czysto losowe (np. zgadywanie).

Błąd losowy
• Po wyeliminowaniu ewidentnych źródeł błędów systematycznych w badaniu testowym na jego wynik poza prawdziwym poziomem mierzonej zmiennej wpływa wielu niekontrolowanych i nieprzewidywalnych czynników (wewnętrznych i zewnętrznych), które powodują, że reakcje osoby badanej na pozycje testowe stają się częściowo nieprzewidywalne i niespójne.
• Mnogość tych zakłócających czynników powoduje, że przyjmuje się, iż powodowany przez nie błąd ma charakter losowy (jest losowy).

Teorie rzetelności pomiaru testowego
Istnieją dwa psychometryczne modele opisujące i wyjaśniające błąd i rzetelność pomiaru:
• Klasyczna Teoria Testów (KTT) (Gulliksen, 1950 oraz Lord i Novick, 1968)
-Teoria Wyników Generycznych - współczesna forma KTT oraz
• Teoria Odpowiadania na Pozycje Testowe.

Podstawowe założenia Klasycznej Teorii Testów Wynik testowy jest efektem działania dwóch grup czynników:
• Czynnika, który wpływa na spójność reagowania na zadanie testowe – mierzonej cechy.
• Czynników, które wpływają na niespójność reagowania na zadania testowe – zmienne te determinują reakcje osoby badanej, ale nie są w ogóle związane z badaną cechą.

Tak więc:
• dla wyniku testowego: obserwowany (otrzymany)
wynik testu= wynik prawdziwy + błąd pomiaru
• dla wariancji wyników testu:
wariancja wyników otrzymanych=wariancja wyników prawdziwych + wariancja błędu

Twierdzenie 1
Wynik testu składa się z wyniku prawdziwego i błędu pomiaru

Twierdzenie 2
Średnia wyników otrzymanych jest równa średniej wyników prawdziwych testu a błędy się znoszą

Twierdzenie 3
Wariancja wyników otrzymanych jest równa sumie wariancji wyników prawdziwych oraz wariancji błędu. Kowariancja pomiędzy wynikami prawdziwymi oraz błędem równa się zero, ale wariancja wyników otrzymanych jest zwiększana przez wariancję błędu.

Twierdzenie 4
Rzetelność pomiaru testem

Twierdzenie 5
Błąd standardowy pomiaru

Interpretacja współczynnika rzetelności
• Współczynnik rzetelności jest ilorazem wariancji wyników prawdziwych do wariancji wyników otrzymanych lub ilorazem wariancji błędu do wariancji wyników otrzymanych, odejmowanej od jedności.
• 1-współczynnik rzetelności wskazuje jaka część wariancji wyników otrzymanych wynika z błędu (niespójności odpowiedzi testowych).

Interpretacja błędu standardowego pomiaru
• Błąd standardowy pomiaru to odchylenie standardowe rozkładu wyników badania danej osoby nieskończenie wiele razy lub badania danej osoby nieskończoną liczbą testów równoległych.
• Średnia takiego rozkładu odpowiada wynikowi prawdziwemu, a odchylenie standardowe jest standardowy błędem pomiaru.
• Ponieważ wielokrotne (nieskończenie wielokrotne) badanie jednej osoby jest nie możliwe błąd pomiaru estymowany jest na podstawie rozkładu błędów pomiaru licznych (ale różnych) osób badanych.
• Ponieważ zakłada się, że błędy pomiaru są nieskorelowane, to nie ma istotnej różnicy pomiędzy efektami losowymi w wynikach licznej grupy osób badanych jednorazowo a w wynikach wielokrotnego badaniem jednej osoby.
• W ramach KTT standardowy błąd pomiaru definiuje zakres wyników, w obrębie którego z określonym prawdopodobieństwem znajduje się wynik prawdziwy osoby badanej.
• W ramach KTT standardowy błąd pomiaru jest identyczny dla wszystkich osób badanych i dla wszystkich wartości wyników otrzymanych (niezależny od wyniku otrzymanego).

Standardowy błąd pomiaru pozwala wyznaczyć przedział ufności dla wyniku prawdziwego z określoną pewnością (najczęściej 0,99 lub 0,95). W tym celu oszacowany dla danego narzędzia SET mnoży się przez stosowną wartość statystyki z rozkładu normalnego (2,58 dla przedziału 99% lub1,96 dla przedziału 95%).

Pojęcie testów równoległych

• Testy równoległe mierzą tę samą cechę z taką samą dokładnością M1 = M2, S12 = S22
• Testy równoważne mierzą tę samą cechę, ale nie tak samo dokładnie M1 = M2
• Testy quasi-równoważne mierzą tę samą cechę wraz z dodatkowym czynnikiem
M1 = M2 + c

Zastosowania koncepcji testów równoległych
• Koncepcja testów równoległych lub pomiarów równoległych jest stosowana w większości metod oceny rzetelności pomiaru testem
• Koncepcja testów równoległych była punktem wyjścia dla teorii wyników generycznych (teoria uniwersalizacji).

Teoria uniwersalizacji
Teoria uniwersalizacji (wyników generycznych) wykorzystuje koncepcję testów równoległych dla stworzenia ram teoretycznych dla myślenia o rzetelności testów w których błędy pomiarowe mogą być skorelowane i może to być empirycznie stwierdzone) (np. testy egzaminacyjne na prawo jazdy). Teoria uniwersalizacji pozwala zrezygnować z nietestowalnych założeń oryginalnej KTT.

Metody szacowania rzetelności pomiaru testem
• Oparte o szacowanie zgodności wewnętrznej
• Oparte o szacowanie stabilności czasowej
• Metoda testów równoległych

Zgodność wewnętrzna
• Metoda zgodności połówkowej (założenie o równoległości połówek testu).
• Metoda zgodności wewnętrznej przy podziale testu na wiele części (założenie o równoległości tych części testu).
• Metoda zgodności wewnętrznej oparta o założenia analizy wariancji.

Metoda zgodności wielu części testu -analiza wariancji
• Istnieje wiele różnych propozycji szacowania rzetelności w ramach modelu analizy wariancji, najprzystępniejszą z nich jest metoda Hoyta

Podstawowe statystyki połówek testowych
Test losowy:
Średnia korelacji pozycji r1-4 = 0,03
Średnie odchylenie standardowe SD1-4 = 0,49
Korelacja połówek r12 = -0,33

Test psychologiczny:
Średnia korelacji pozycji r1-4 = 0,83
Średnie odchylenie standardowe SD1-4 = 0,49
Korelacja połówek r12 = 0,91

Stabilność czasowa
1. Stabilność bezwzględna(powtórny odroczony pomiar tym samym testem
-tymi samymi pozycjami)
-test-retest (pomiar tym samym testem raz po razie)
2. Stabilność względna (powtórny pomiar równoległą wersją testu)

Estymacja rzetelności wyników testów grupowych:
Badanie rzetelności metodą dwukrotnego badania tej samej grupy osób (technika test-retest) .
Współczynnik rzetelności szacowany tą metodą, nazywany jest też współczynnikiem stabilności bezwzględnej. Jest on miarą stałości wyników testowych mimo przypadkowych zmian, dotyczących zarówno osoby badanej, jak i warunków badania.

Estymacja rzetelności wyników testów grupowych
•Długość przerwy między pierwszym a drugim testowaniem staje się jest istotnym czynnikiem wpływającym na wielkość otrzymanego współczynnika rzetelności.
•Określając optymalną długość przerwy między kolejnymi badaniami tym samym testem, warto zadbać, by:
– przerwa ta była na tyle długa, aby osoby badane zapomniały swoje poprzednie odpowiedzi w teście;
– przerwa ta była na tyle krótka, aby w trakcie jej trwania nie doszło do zmian w wyniku procesów rozwojowych czy uczenia się;

•Określając optymalną długość przerwy między kolejnymi badaniami tym samym testem, warto zadbać, by:
– wyznaczając długość przerwy uwzględnić cel testowania (np. osobowość czy emocjonalność), jak również to, dla kogo test jest przeznaczony (dzieci czy dorośli);
– Zazwyczaj czas przerwy waha się od kilku tygodni do kilku miesięcy. Wszelkie zmiany, które pojawiają się w okresie dłuższym niż kilka miesięcy, raczej mają charakter zmian progresywnych niż zmian losowych.

•Specyficzną odmianą techniki test-retest jest dwukrotne badanie lej samej grupy osób tym samym testem bez żadnej przerwy czasowej. Z punktu widzenia osoby badanej jest to jedno badanie, w którym dwukrotnie powtarzają się te same pozycje.
•Współczynnik korelacji między wynikami pierwszego i drugiego testu jest opisywany jako współczynnik wiarygodności testu. W technice tej maksymalizowany jest czynnik zapamiętywania, zaś minimalizowany jest czynnik uczenia się.

•Technika szacowania rzetelności metodą dwukrotnego testowania tej samej grupy osób, mimo jej intuicyjnej prostoty, budzi jednak wiele wątpliwości. W Standardach dla testów stosowanych w psychologii i pedagogice (1985a, s. 58) wyraźnie podkreśla się, że „(...) nie jest to pożądana technika badania rzetelności„.
•Technika ta daje się zaakceptować w wypadku testów motorycznych czy różnicowania sensorycznego (tj. takich testów, w których zakłada się, że powtarzanie badania nie wpływa w sposób istotny na wyniki testowania).

Czynniki wpływające na stabilność czasową skal osobowości

• Zgodność wewnętrzna skal (wyższa stabilność dla bardziej rzetelnych skal);
• Liczba pozycji w skali (wyższa stabilność dla dłuższych skal);
• Długość przerwy (wyższa stabilność przy krótszej przerwie);

Wiek osób badanych a stabilność czasowa skal osobowości
• Wiek osób badanych podczas pierwszego badania (wyższa stabilność dla starszych osób).
współczynniki stabilności:
0,31 -dzieci, 0,54 -młodzież i młodzi dorośli,
0,64 -dorośli po ukończeniu 30 lat,
0,74 –dorośli po ukończeniu 50 lat.

Metoda testów równoległych
Metoda testów równoległych wymaga dwóch odrębnych testów
- jest metodą uogólnioną, łącząca zgodność wewnętrzną oraz test-retest. W metodzie tej współczynnik korelacji Pearsona jest oszacowaniem rzetelności testu
-rzetelność pomiaru jest równa współczynnikowi korelacji obu testów równoległych (wielkości kowariancji obu testów).

Uwarunkowania rzetelności testu
Zakres (zmienność) wyników w badanej próbie – współczynniki są niższe, gdy zmienność w próbie jest mniejsza (osoby badane mają zbliżone nasilenie cechy).
Gdy nie ma zmienności, rzetelność równa się 0.
To oczywiste, gdyż współczynnik rzetelności odnosi się do rzetelności pomiaru różnicowego.

Uwarunkowania rzetelności testu
Długość testu –współczynniki są wyższe gdy test zawiera dużo pozycji (z uwagi na dużą liczbę kowariancji).

Uwarunkowania rzetelności testu
Metoda estymacji rzetelności testu –współczynniki zgodności wewnętrznej dają wyższe oszacowanie rzetelności niż współczynniki stabilności
Testy o wyższej zgodności wewnętrznej zwykle wykazują też wyższą stabilność czasową (wyjątkiem testy stanów psychologicznych).

Umowne kryteria akceptacji rzetelności pomiaru testem Zgodność wewnętrzna:
1.Testy przeznaczone do diagnozy indywidualnej rzetelność minimalna 0,80, pożądana – ponad 0,90.
2.Testy przeznaczone do badań naukowych -rzetelność minimalna 0,70, pożądana –ponad 0,80.
3.Bezwzględnie minimalna wartość współczynnika rzetelności -0,50, czyli połowa wariancji pomiaru jest wynikiem działania czynników losowych (błędu).
4. Testy przeznaczone do diagnozy indywidualnej rzetelność minimalna 0,80, pożądana – ponad 0,90.
5. Testy przeznaczone do badań naukowych -rzetelność minimalna 0,70, pożądana –ponad 0,80.
6. Bezwzględnie minimalna wartość współczynnika rzetelności -0,50, czyli połowa wariancji pomiaru jest wynikiem działania czynników losowych (błędu).

Umowne kryteria akceptacji rzetelności pomiaru testem Stabilność czasowa:
Wartość minimalna = 0,50 (połowa zmienności wynika ze zgodności wyników obu pomiarów).
Testy równoległe: Wartość minimalna = 0,50 (połowa zmienności wynika ze zgodności wyników obu testów).
Uwaga: Wartości dotyczą miar stałych cech a nie stanów.

Rodzaj testu a wybór metody szacowania rzetelności
• Testy zdolności (mocy) – metody połówkowe (z uwagi na różną trudność pozycji, które nie są równoległe).
• Inwentarze osobowości –alfa Cronbacha lub KR-20 (zgodność wewnętrzna na poziomie pozycji).
• Testy szybkości –metoda test-retest lub metoda testów równoległych.

Wykorzystanie oszacowania rzetelności testu
Współczynnik rzetelności testu pozwala estymować standardowy błąd pomiaru wyników otrzymanych oraz standardowy błąd estymacji wyniku prawdziwego

Wyznaczanie przedział ufności dla wyniku prawdziwego
Standardowy błąd pomiaru pozwala wyznaczyć przedział ufności dla wyniku prawdziwego z określoną pewnością (najczęściej 0,99 lub 0,95).
W tym celu oszacowany dla danego narzędzia SET mnoży się przez stosowną wartość statystyki z rozkładu normalnego (2,58 dla przedziału 99% lub 1,96 dla przedziału 95%).

Wyznaczanie przedziałów ufności
Aby wyznaczyć przedział ufności trzeba wyznaczyć półprzedział, tzn. Sbp lub Sbe przemnożyć przez wartość 2,58 (99% pewność), 1,96 (95% pewność) a następnie dodać i odjąć od wyniku:
• otrzymanego lub
• oszacowanego wyniku prawdziwego.
W ten sposób wyznaczane są granice przedziału ufności dla wyników: • otrzymanego lub • oszacowanego wyniku prawdziwego.

Zastosowanie standardowych błędów pomiaru
Błędy pomiaru służą do:
1) Wyznaczenia granic przedziału ufności wyniku otrzymanego i przedziałowej estymacji wyniku prawdziwego (w zakresie którego mieści się –z określoną pewnością wynik prawdziwy osoby badanej).
2) Porównania wyniku danej osoby z normą (średnią w grupie) czy inną wartością (np. wynikiem progowym).

Błędy pomiaru służą do:
3) Porównania wyników różnych osób badanych tym samym testem (sprawdzenie czy różnica jest realna –wynika z różnic w natężeniu cechy czy jest spowodowana przez błąd pomiaru).
4) Porównania wyników danej osoby testowanej dwoma różnymi testami (sprawdzenie czy

różnica jest realna – wynika z różnic w natężeniu cechy czy jest spowodowana przez błąd pomiaru).

Pojęcie trafności
• Trafność = dokładność z jaką test realizuje założone cele przez jego autorów.
• Trafność interpretacji wyników danego testu w aspektach:
“Co test mierzy i jak dobrze to robi?“
“Jaki jest obszar zastosowania danego testu?"
“Czy dany test odpowiada celom jego użytkownika?".

Wg Standardów dla testów ...(1985):
“pojęcie trafności dotyczy poprawności wniosków wyprowadzanych na podstawie wyników testowych lub innych form badania", “(...) trafność jest pewnym wnioskiem, a nie pomiarem. W podręczniku testowym można przedstawić jedynie współczynniki trafności. To na ich podstawie wyciąga się wnioski o trafności konkretnego zastosowania testu (...)".

Pojęcie trafności
Oszacowywanie trafności testu, nazywane w psychometrii procesem walidacji testu (ang. validation), jest procesem zbierania i oceniania danych świadczących o trafności określonej interpretacji wyników testu. Im więcej przeprowadza się badań z udziałem danego testu, tym szerszy jest potencjalny obszar jego zastosowania.
Procedura walidacji testu nie kończy się zatem na podaniu jednego współczynnika trafności, a polega na prowadzeniu ciągłych badań i gromadzeniu informacji.

Trafność w KTT
Pojęcie trafności odwołuje się do założeń Klasycznej Teorii Testów, że
• wyniki prawdziwe i błędy pomiaru są nieskorelowane (rtb = 0), ,
• błędy są nieskorelowane (rbb = 0),
Co pozwala przyjąć, że obserwowane korelacje (między pozycjami, testem i innymi testami oraz testem a kryteriami) są korelacjami wyników prawdziwych.

Trafność a rzetelność pomiaru
Rzetelność jest koniecznym , ale niewystaczającym warunkiem trafności pomiaru.
Test może być rzetelny i nietrafny, ale test nierzetelny jest nietrafny jednocześnie.
Rzetelność jest górnym kresem trafności, ponieważ wariancja prawdziwa jest podstawą estymacji rzetelności a także trafności.

Źródła wariancji przy analizie rzetelności i trafności
(A) Wariancja wspólna z innymi testami.
(B) Wariancja specyficzna dla danego testu.
(C) Wariancja błędu (losowa).

Trafność a rzetelność pomiaru
Rzetelność i trafność są parametrami psychometrycznymi pomiaru testem i są wyznaczane

przez podobne czynniki zmiany w procedurze standaryzacji (alternatywne zastosowanie testu).
• zmiany w demograficznym składzie próby -ograniczona zmienność wyników testu lub wyników kryterialnych w grupie;
- wpływają i na rzetelność i na trafność, najczęściej je obniżając.

RODZAJE TRAFNOŚCI

W psychometrii na ogół wyodrębnia się trzy rodzaje trafności z uwagi na rodzaj źródła informacji o trafności testu
• trafność treściową,
• trafność kryterialną oraz
• trafność teoretyczną
Zdaniem Cronbacha (1990) każde badanie trafności testu powinno integrować informacje z wszystkich tych źródeł.

Trafność treściowa.

Trafność treściowa, nazywana też jest czasami trafnością wewnętrzną lub logiczną (content validity)
= zakres, w jakim pozycje testowe właściwie reprezentują uniwersum pozycji testowych operacjonalizujących mierzony konstrukt = jako zakres, w jakim treść testu stanowi reprezentatywną próbę dziedziny, która ma być przedmiotem pomiaru Trafność treściową odróżnia się od tzw. trafności fasadowej (face validity), która odnosi się do trafności w sensie definicyjnym i dotyczy tego, co wg osób badanych dany test wydaje się mierzyć. Trafność fasadowa nie charakteryzuje dobroci testu.

Trafność kryterialna (criterion-related validity)

• dotyczy stopnia w jakim na podstawie wyników testowych można wnioskować o przypuszczalnej pozycji badanego względem innej zmiennej -tzw. Kryterium
• wskazuje na zakres, w jakim wyniki testowe są empirycznie powiązane z ważnym dla badacza/diagnosty kryterium.
Np. wyniki testu stanowiącego egzamin wstępny na wyższą uczelnię można potraktować, jako wskaźnik późniejszych osiągnięć w trakcie studiów.

Trafność kryterialna

Trafność kryterialna określa skuteczność testu w diagnozowaniu i/lub prognozowaniu funkcjonowania jednostki w określonej sferze, stąd ma dwa aspekty:
Trafność diagnostyczna (concurrent validity) określa, w jakim zakresie można wykorzystywać test do określania aktualnej pozycji osoby badanej względem kryterium, np. udzielić odpowiedzi: “Czy osoba badana posiada cechę X?"
Trafność prognostyczna (predictive validity) mówi o tym, w jakim stopniu można -na podstawie wyników testowych przewidywać przyszłą pozycję osoby badanej względem zmiennej kryterialnej,
np. stwierdzić: “Jakie jest prawdopodobieństwo tego, że osoba badana będzie posiadać cechę X?"

Trafność teoretyczna (construct validity)

Trafność teoretyczna jest oszacowaniem stopnia, w jakim wnioski wyprowadzone na podstawie wyników testowych odzwierciedlają pozycję osoby badanej na pewnym teoretycznym kontinuum mierzonej cechy.

Trafność teoretyczna jest funkcją:
• definiowania -tak jasno jak to możliwe
-mierzonej cechy (konstruktu). CO TEST MIERZY,
• skumulowanych wyników testowych z wielu badań z zachowaniami osób badanych w takich sytuacjach, w jakich mierzony konstrukt jest traktowany jako ważna zmienna. GDZIE TEST SIĘ SPRAWDZIŁ

Test jest trafny treściowo, gdy :
• wszystkie jego pozycje należą do zdefiniowanego uniwersum (opisują pełen zakres dziedziny, której test ma dotyczyć), oraz
• test proporcjonalnie reprezentuje zdefiniowane uniwersum.
Do oceny trafności treściowej najczęściej korzysta się z pomocy sędziów -ekspertów, którzy na podstawie swojej wiedzy o tym, co ma być przedmiotem pomiaru oceniają, każdą pozycję testową wg skali:
1) ma zasadnicze znaczenie dla testu;
2) jest użyteczna, jednak nie ma zasadniczego znaczenia;
3) nie powinna znaleźć się w obrębie testu.

SPOSOBY BADANIA TRAFNOŚCI KRYTERIALNEJ
Trafność kryterialna= stopień zgodności wyników testowych i wskaźników zewnętrznych, które pełnią funkcję kryterium, czyli standardu, względem którego ocenia się jakość wyników testowych.
Cechy kryterium:
• Wysoka rzetelność,
• Trafność z punktu widzenia celu pomiaru.
• Brak kontaminacji (skażenia) kryterium, tzn. znajomość wyników, jakie osoba badana uzyskała w teście, nie może wpływać na ocenę wyniku tej osoby względem analizowanej zmiennej kryterialnej.

Wskaźnikiem trafności kryterialnej jest współczynnik korelacji między wynikami testu a poziomami zmiennej kryterialnej, zebranymi dla tej samej grupy badanych osób.
Im wyższa wartość współczynnika korelacji, tym wyższa trafność kryterialna testu.
Informacje o trafności testu powinny być na tyle wyczerpujące, by użytkownik testu mógł określić, czy na ich podstawie może wykorzystywać test swoich celów, czy charakterystyki próby, na której przeprowadzono badania walidacyjne, odpowiadają charakterystykom tej grupy osób, dla której test ma być stosowany oraz czy podane współczynniki trafności są wystarczająco wysokie.

Trafność teoretyczna charakteryzuje związek pomiędzy cechą psychologiczną (konstruktem), wywodzącą się z określonej teorii psychologicznej, a narzędziem pomiarowym (testem), będącym operacyjną miarą tej cechy.

UWAGA: Pojęcia, takie jak “lęk", “satysfakcja z pracy", “inteligencja”,“twórczość”, „miłość”, itd. są nieobserwowalne. Test jest sposobem ich operacyjnego zdefiniowania.

SPOSOBY BADANIA TRAFNOŚCI TEORETYCZNEJ

Trafność teoretyczna odnosi się wprost do pytania o przedmiot pomiaru testowego.
Podstawowe metody badania trafności teoretycznej testu:

- analiza różnic międzygrupowych,
- analiza macierzy korelacji,
- analiza czynnikowa,
- analiza struktury wewnętrznej testu,
- analiza zmian nieprzypadkowych wyników testu,
- analiza procesu rozwiązywania testu.

Analiza różnic międzygrupowych. Metoda ta polega na weryfikowaniu hipotez dotyczących różnego zachowania się dwóch grup osób, które różnią się nasileniem cechy mierzonej przez test. Najczęściej są to tzw. grupy skrajne, tj. grupa o niskich wynikach oraz grupa o wysokich wynikach w teście.
Analiza macierzy korelacji. Korelacje wyników ocenianego testu z wynikami testów mierzących podobne cechy powinny być wysokie (aspekt zbieżności), zaś korelacje z wynikami testów mierzących inne cechy powinny być niskie (aspekt różnicowy).
Analiza czynnikowa pozwala ona sprawdzić, czy otrzymane dane empiryczne są zgodne z zakładaną strukturą teoretyczną testu.

Analiza struktury wewnętrznej testu, jej zgodności wewnętrznej (homogeniczność) , tj. stopnia, w jakim dany test można uznać za miarę jednego tylko konstruktu.
Jedną z metod szacowania stopnia zgodności wewnętrznej jest
analiza współczynników korelacji między wynikiem każdej pozycji testu a ogólnym wynikiem w tym teście.
Analiza zmian nieprzypadkowych wyników testu. W przerwie między badaniami tym samym testem wprowadza się oddziaływanie eksperymentalne, wyprowadzone z teorii mierzonej cechy.
Wynik porównywania powinien być zgodny z założonymi efektami manipulacji (wyniki post-testu powinny się obniżyć lub podwyższyć zgodnie z hipotezami).

STRONNICZOŚĆ TESTÓW

W psychometrii „stronniczość" to błąd systematyczny , który jest związany z wynikami testowymi osób należących do konkretnej podgrupy populacji (rasowej, klasowej, narodowościowej lub wiekowej).
Pojęciu „stronniczość”
-szczególnie z powodu potocznych skojarzeń ze słowem <stronniczość>
-przypisuje się często znaczenie inne niż to ustalone w psychmetrii
W statystyce termin „stronniczość" =„obciążenie" oznacza systematyczne niedoszacowywanie lub przeszacowywanie parametru populacyjnego na podstawie danych z próby. (np. M -jest nieobciążonym estymatorem a SD obciążonym)

Psychometryczna „stronniczość” testu nie ma nic wspólnego z uczciwością, równością, uprzedzeniami, czy preferencjami.

Stronniczość to cecha testów, w których błędy systematyczne obniżają ich trafność (przede wszystkim teoretyczną, ale także treściową, kryterialną i prognostyczną)
Stronniczość zupełnie nie wiąże się z błędami losowymi.

Źródła stronniczości testu
Niewłaściwa treść testu-osoby pochodzące z grup społecznych np. o niższym statusie mogą nigdy nie zetknąć się ze specyficznym materiałem, który złożył się na treść pozycji testowych. Może to dotyczyć zarówno języka, wiedzy, jak i wartości.

Pomiar różnych charakterystyk-ten sam test może mierzyć odmienne charakterystyki (zmienne psychologiczne), jeżeli stosowany jest w stosunku do osób pochodzących spoza kultury, która była “źródłem" pozycji testowych.

Język, w jakim test został sformułowany
-osoby poddane badaniu testowemu w innym -niż własny -języku uzyskują generalnie niższe wyniki, często z uwagi na trudności komunikacyjne.
Niewłaściwa próba standaryzacyjna
-jeżeli w próbie standaryzacyjnej nie są reprezentowane wszystkie grupy, które mogą być badane określonym testem, to test należy uznać za narzędzie stronnicze w stosunku do tych grup, które nie zostały w próbie standaryzacyjnej uwzględnione.

Stronniczość testu a trafność treściowa
Test może zostać uznany za stronniczy, jeżeli uniwersum pozycji testowych zostało trafnie określone tylko w stosunku do członków wybranej (wybranych) grupy (grup) (np. większości).

Stronniczość testu w aspekcie trafności teoretycznej ma miejsce, gdy test mierzy różne cechy (konstrukty psychologiczne) w wypadku różnych grup lub gdy mierzy tę samą cechę, lecz z różnym stopniem dokładności.
Do najczęściej stosowanych metod pozwalających szacować tego typu stronniczość należy analiza czynnikowa. Stwierdzenie identycznych czynników w grupach wyodrębnionych w ramach tej samej populacji traktuje się jak dowód, że test mierzy ten sam konstrukt we wszystkich grupach.

Techniki szacowania stronniczości testu
Empiryczne szacowanie stronniczości testu polega na ocenie funkcjonowania testu z punktu widzenia jego trafności kryterialnej. Stosowane testy powinny być -z założenia wysoko skorelowane z kryterium będącym podstawą podjęcia decyzji o charakterze kwalifikacyjnym / etykietującym (czy to diagnostycznych, czy prognostycznych).

Wskaźnikiem stronniczości testu jest zróżnicowanie wielkości korelacji między wynikami testu a wybranymi miarami kryterium w testowanych grupach. Tzn. jeśli grupa jest moderatorem siły lub/i kierunku związku pomiędzy wynikiem testowym a kryterium, to test można podejrzewać o stronniczość.
Test bezstronny powinien posiadać podobne korelacje z tymi samymi miarami kryterium dla

wszystkich analizowanych grup. Zgodnie z definicją stronniczości jako nierówności linii regresji w grupach, tylko sytuacja przedstawiona na rys. 1 ilustruje wyniki zastosowania testu bezstronnego

Podsumowując
Stronniczość testów oznacza błąd systematyczny popełniany przy prognozowaniu wartości zmiennej kryterialnej dla osób z różnych grup, będący rezultatem:
• oparcia diagnozy lub prognozy na wspólnej linii regresji wyznaczonej dla wszystkich osób bez względu na ich populacyjną przynależność, lub
• oparcia diagnozy lub prognozy wyników kryterialnych osób należących do jednej grupy na równaniu regresji wyznaczonym dla innej.

Test oceniany jest pod kątem trafności diagnozy lub prognozy w stosunku do członków określonych grup pochodzących z tej samej populacji. Badanie stronniczości polega na wyznaczeniu linii regresji dla każdej z grup, a następnie na ocenie ich zgodności.

Czynniki zniekształcające wyniki testowe
1) Zgadywanie.
Kontrolowanie zgadywania:
• poprzez instrukcję testową: wyrównywanie tendencji do zgadywania (zachęcanie do zgadywania) lub eliminowanie zgadywania (informacja o stosowaniu korekty wyników);
• zastosowanie statystycznej poprawki na zgadywanie:
2) Style odpowiadania
Styl odpowiadania -tendencja osoby testowanej do wybierania określonej opcji odpowiedzi niezależnie od treści pozycji kwestionariuszowej; tendencja do:
• zgadzania się albo zaprzeczania,
• udzielania odpowiedzi ekstremalnych albo centralnych (pośrednich),
• udzielania odpowiedzi nieuważnych lub niekonsekwentnych,
• udzielania odpowiedzi ekstremalnych albo centralnych (pośrednich),
• udzielania odpowiedzi nieuważnych lub niekonsekwentnych,
• losowych (przypadkowych),
• niezdecydowanych (opcje „?” lub opuszczenia odpowiedzi),
• pozornie oryginalnych lub konwencjonalnych,
• impulsywnych, itd..
Ważne jest odróżnienie tendencyjności ogólnej (podatność narzędzia lub procedury badania) oraz różnic indywidualnych w stylu reagowanie – skale kontrolne badają różnice indywidualne, wyjątkowo mogą być zastosowane do analizy sytuacji badania
Źródła stylów odpowiadania.
Brak motywacji osoby testowanej do wzięcia udziału w testowaniu lub poczucie zagrożenia badaniem
-styl odpowiadania związany jest z chęcią ukrycia faktycznego obrazu osobowości (inteligencji).
Własności pozycji oraz zastosowane opcje odpowiedzi niezrozumiałość oraz niejasność pozycji oraz nieadekwatność zastosowanego formatu odpowiedzi.
Zalecenia.
Właściwe sformułowanie językowe pozycji oraz właściwy format odpowiedzi (eliminowanie lub dodawanie odpowiedzi pośrednich, dostosowanie formatu do preferencji osób badanych).

Zrównoważenie skali pod względem klucza odpowiedzi (równoliczność pozycji, dla których diagnostyczne są odpowiedzi twierdzące z tymi, dla których diagnostyczne są zaprzeczenia; niezbędne do zbudowania skal kontrolnych).

3) Tendencyjne udzielanie odpowiedzi społecznie aprobowanych albo społecznie nieaprobowanych:
- tendencja do dysymulowania (aprobata społeczna),
- tendencja do symulowania.

Dysymulowanie- tendencja osoby testowanej do przedstawiania się w nieprawdziwie korzystnym świetle, która jest związana ze zmienną aprobaty społecznej. Tendencja osoby testowanej do kierowania się społecznym wartościowaniem zachowania przy odpowiadaniu, prowadząca do zaprzeczania posiadania cech społecznie niepożądanych oraz przypisywania sobie cech społecznie pożądanych.

RADZENIE SOBIE Dbałość by przynajmniej niektóre pozycje:
• były neutralne.
• były subtelne pod względem trafności fasadowej.
• Specjalne formułowane językowo– unikanie dużych kwantyfikatorów czasu, “miękkie” opisywanie zachowań (np. Można odnieść wrażenie, że nie jestem szczery).
• Zmiana procedury badania np. poprzez użycie komputera.

Sposoby wykrywania.
• Osobno mierzona zmienna aprobaty społecznej silnie koreluje z pozycjami lub/iwynikami testu.
• Laicy nie mają problemu z wypełnieniem testu tak, by otrzymać najwyższe oceny lub doskonały profil.

Sposoby zwalczania na etapie testowania.
• Specjalna instrukcja.
• Zmiana procedury badania na uniemożliwiającą powrót i poprawianie poprzednich odpowiedzi (np. komputeryzacja testu).
• Wprowadzenie kontrolnej skali aprobaty społecznej, na podstawie wyników której specjalnie traktuje się osoby o wysokich wynikach
• Diagnoza psychologiczna oparta o metody wykorzystujące badanie dokumentów oraz informacji od (licznych) osób trzecich.

Symulowanie –tendencja osoby testowanej do udzielania odpowiedzi, bezpodstawnie przedstawiających ją w niekorzystnym świetle, np. wskazujących na zaburzenia psychiczne lub niepożądane cechy osobowości.
Tendencja do symulowania może być sytuacyjną reakcją na konkretne badanie lub trwałym syndromem cech osobowości, związanym z agrawacją, obniżoną samooceną, ekscentrycznością i zaburzeniami psychicznymi.

Tendencyjne udzielanie odpowiedzi społecznie nieaprobowanych. Sposoby radzenia.
• Unikanie pytań o zachowania regulowane społecznymi normami.
• Unikanie patosu, przesady i afektacji w formułowaniu treści pozycji.

Sposoby wykrywania.

• Osobno mierzona tendencja do symulowania silnie koreluje z pozycjami lub/iwynikami testu.
• Laicy nie mają problemu z wypełnieniem testu tak, by otrzymać najniższe oceny lub najgorszy profil.
• Osobno mierzona tendencja do symulowania silnie koreluje z pozycjami lub/iwynikami testu.
• Laicy nie mają problemu z wypełnieniem testu tak, by otrzymać najniższe oceny lub najgorszy profil.

Konstruowanie testów –etapy opracowywania testu
Jakość testu zależy od jakości jego elementów składowych, czyli pozycji testowych oraz procedury testowania.

Konstruowanie testu to długotrwały i kosztowny proces wymagający w pierwszej kolejności:
– konceptualizacji tego, co ma być przedmiotem pomiaru (zdefinowania mierzonego zmiennej oraz jej operacjonalizacji);
– określenia formatu pozycji i formatu odpowiedzi;
– zgromadzenia dużej puli pozycji testowych;
– analizy językowej i treściowej powyżej wzmiankowanej puli;
W wyniku powyższych zabiegów powstaje pierwsza (eksperymentalna) wersja testu. A to dopiero początek procesu konstruowania testu.

Konstruowanie testu to długotrwały i kosztowny proces wymagający w dalszej kolejności:
•przeprowadzenia badań pilotażowych i oszacowania właściwości psychometrycznych poszczególnych pozycji;
•podjęcia decyzji, które z pozycji testowych nadają się do kolejnej wersji budowanego narzędzia
•przeprowadzenia badań nową wersją testu i ponownego sprawdzenia właściwości psychometrycznych poszczególnych pozycji;
1) Jeżeli okaże się, że pozycji testowych spełniających wymagania jest zbyt mało, procedura rozpoczyna się od początku (od gromadzenia puli pozycji testowych).
2) Jeżeli okaże się, że pozycje testowe spełniają wymagania przechodzi się do walidacji narzędzia.

Określanie celu pomiaru
Test, którego wyniki mają być podstawą selekcji, powinien dobrze różnicować osoby pożądanym poziomie mierzonej zmiennej od osób o wynikach niepożądanych (najczęściej osoby o wysokich wynikach skalowych), ale test przeznaczony do diagnozy indywidualnej powinien dobrze różnicować osoby na całym kontinuum mierzonej zmiennej (zarówno osoby o wysokich, jak i niskich wynikach skalowych).
Jasne określenie celu głównego obszaru zastosowania testu a więc rodzaju decyzji, jakie będą podejmowane na podstawie jego wyników testowych, w dużym stopniu zwiększa prawdopodobieństwo to, że ostateczna wersja testu spełni te oczekiwania.

Procedura operacjonalizacji
Teoretyczne pojęcia psychologiczne denotują z reguły szeroki zakres zjawisk, a więc trudno

je jednoznacznei i precyzyjnie zdefiniować.
Co to jest: „lęk", „konformizm", „inteligencja", „ekstrawersja", „agresja" czy „ugodowość".
Dla powodzenia procesu budowy testu niezbędne jest bardzo precyzyjne zdefiowanie docelowo mierzonej zmiennej, co jest możliwe w zasadzie tylko poprzez zastosowanie definicji operacyjnej.
Czyli przez wskazanie obserwowalnych wskaźników niskiego, średniego i wysokiego natężenia mierzonej zmiennej..
Procedura operacjonalizacji = procedura wiązania pojęć teoretycznych z obserwacyjnymi wskaźnikami.
Operacjonalizacja
1. określa, co należy zrobić i co należy obserwować, aby mierzone zachowania można było potraktować jako wskaźniki mierzonej zmiennej (cechy) psychologicznej;
2. pozwala użytkownikowi testu (także autorowi testu) na powiązanie mierzonej cechy (pojęcia teoretycznego) z konkretnym zbiorem obserwacji.

Zasady konstruowania pozycji testowych
Wyjściowy zbiór pozycji testowych (pula pozycji) powinien być duży (zalecane się skonstruowanie 2-3 x tylu pozycji, ile liczyć ich ma ostateczna wersja testu), co zwiększa szansę na uzyskanie odpowiedniej liczby pozycji spełniających wymagane kryteria.
Zanim pozycje testowe zaczną być generowane testowe, trzeba zdecydować jaki będzie najbardziej odpowiedni i korzystny format pozycji testowych. Za Pophamem (1981) można wyróżnić dwa główne formaty pozycji:
• otwarte (constructed-response format) –odpowiedź swobodna, spontaniczna w formie uzupełnienia (completion item), krótkiej wypowiedzi (ang. short answer) lub opowiadania (ang. essay)
• zamknięte (selected-response format) –odpowiedź wybierana z przygotowanej puli odpowiedzi.

Opcje odpowiedzi
! Problem opcji/wartości centralnej w formie “nie wiem” , (?), “tak i nie”, “trochę tak, trochę nie:.
Wybór opcji centralnej przez osobę testowaną:
1) oznacza, że jej zachowanie zależy od sytuacji;
2) jej odpowiedź wynika z trudności w ocenie zachowania;
3) jej odpowiedź wskazuje, że nie potrafiła ona jednoznacznie rozkodować treści pozycji (pozycja jest niejasna);
4) jej odpowiedź wskazuje, że rzeczywiście ma przeciętne nasilenie cechy.

Jednak brak opcji środkowej (parzysta liczba opcji) powoduje wyrwę w kontinuum szacowania nasilenia zachowania (postawy), dlatego też dopuszczalna (pożądana) jest ta wartość środkowa o ile wyraźnie wyraża ona umiarkowane nasilenie zachowania (postawy).

Opcje odpowiedzi Szerokość formatu:
odpowiedzi wielokategorialne dwukategorialne.
Wady formatów dwukategorialnych:

1) Zmuszają do udzielenia zdecydowanej odpowiedzi.
2) Nie gwarantują odpowiedniego zróżnicowania odpowiedzi.
Wady formatów wielokategorialnych:
1) Są trudne, zwłaszcza dla osób gorzej wykształconych. 2) Są podatne na style odpowiadania.

Opcje odpowiedzi-optymalna liczba
•osoby dobrze funkcjonujące poznawczo 4 -9 opcji
•osoby słabo funkcjonujące poznawczo 2 -4
Wniosek: 4 opcje są zawsze dobre (ale być może 5 są lepsze)

Opcje odpowiedzi -inne problemy.
•Zakotwiczenie skali: najlepiej jak najbardziej skrajne
•Kierunek zakotwiczenia: negatywny-pozytywny (np.: ZNz Nz Z ZZ) .
•Wartości liczbowe są wtórne (podstawowe określenia słowne), choć nie bez znaczenia.
•Opcje powinny być ujednolicone dla wszystkich pozycji (także w skali kontrolnej).

Analiza pozycji testowych
Ocena pozycji testowych powinna zostać przeprowadzona w trzech aspektach: językowym, treściowym i statystycznym.
Analiza językowa prowadzona jest pod kątem sprawdzenia poprawności gramatycznej, zrozumiałości, stosowanego słownictwa oraz zrozumiałości całej pozycji
Pozycje:
1. nie powinny być zbyt długie (preferowane są pozycje jednozdaniowe, chyba że mają stanowić opis pewnej sytuacji lub problemu)
2. powinny być raczej zbudowana ze zdań prostych niż złożonych, a jeżeli dana pozycja wymaga dłuższego sformułowania;
• powinna być sformułowana w trybie oznajmującym (przeczenia, zwłaszcza podwójne, mogą prowadzić do nieporozumień interpretacyjnych).

3. Język pozycji powinien być prosty (nie powinny one zawierać trudnych zwrotów lub wrażeń);
4. Pozycje powinny być napisane językiem dostosowanym do przeciętnych kompetencji językowych osób z populacji, dla której przygotowywany jest test.
5. Pozycje powinny być poprawnie sformułowana gramatycznie.
6. Należy unikać przysłówków np. „czasami", „rzadko", „niekiedy", „kilka" „wiele" „nigdy" czy „zawsze", ponieważ osoby testowane mogą różne je interpretować.

Analiza treściowa = ocena pozycji testowych z uwagi na zgodność treści poszczególnych pozycji testowych z założeniami teoretycznymi. Najczęściej zespół ekspertów proszony jest o ocenę zgodności treści poszczególnych pozycji z przyjętymi wcześniej założeniami.
Członkowie zespołu udzielają odpowiedzi na dwa pytania:
1. Czy wszystkie pozycje testowe można traktować jako operacjonalizację mierzonej cechy?
2. Czy pozycje testowe reprezentują uniwersum zachowań, ważnych z punktu widzenia tej cechy.
Wysoki współczynnik zgodności między sędziami jest podstawą oceny jakości pozycji testowej.

Badanie pilotażowe
Próbne pilotażowe musi zostać przeprowadzone na próbce osób z populacji, dla której test jest przeznaczony.
Próba pilotażowa powinna być liczna.
Najczęściej zaleca się przebadanie 10 osób na każdą pozycję testową, ale nie mniej niż 200 osób (lepiej 400).
Im większa próba, tym lepiej. Jeżeli test jest długi, to preferowaną procedurą jest podział testu na części i zbadanie każdą z części innej próbki osób.

Badanie pilotażowe powinno przebiegać w takich samych warunkach i wg takich samych procedur, w jakich stosowany będzie gotowy test.
• instrukcja, limity czasowe, charakter badania (indywidualny czy grupowy), oraz atmosfera w trakcie badania powinny być takie, jakie będą w trakcie właściwego badania tym testem.
Wyniki otrzymane w badaniu pilotażowym są analizowane a każda pozycja testowa jest opisywana za pomocą wybranych statystyk. Najczęściej są to wskaźnik trudności pozycji i współczynnik mocy dyskryminacyjnej. Do ostatecznej wersji testu włączane są tylko te pozycje, których właściwości statystyczne będą satysfakcjonujące.

Wskaźnik trudności pozycji testowych (item-difficulty index) oblicza się w testach zdolności, umiejętności oraz wiadomości (bo tylko w nich jest odpowiedź poprawna) i wybiera się pozycje testowych, które mają odpowiedni z punktu widzenia celu testowania -poziom trudności.

T = proporcja osób, które poprawnie odpowiedziały na daną pozycję testową (pi-), wyrażoną w procentach. Im wyższa wartość T, tym łatwiejsza jest dana pozycja testowa.
Dla ułatwienia (utrudnienia*) ze względu na taki właśnie sposób interpretacji wielkości współczynnika T czasami nazywa się go wskaźnikiem łatwości zadania.
* wybrać właściwe

Badanie pilotażowe. Wskaźnik trudności pozycji testowych
Jeżeli celem testowania jest różnicowanie badanych osób na całym kontinuum zmienności, to dobrą pozycją testową jest taka pozycja, która gwarantuje to zróżnicowanie, tj. gdy wskaźnik trudności jest równy lub niewiele mniejszy lub większy niż 50%.

! Ale pod warunkiem, że pozycje testowe nie korelują ze sobą. Bo jeśli silnie korelują, to w efekcie otrzymujemy podział tylko na dwie kategorie: tych, którzy mają maksymalną wiedzę na ten temat, i tych, którzy nic nie wiedzą.

Jeśli celem jest precyzyjne zróżnicowanie badanej grupy, wówczas konieczne, aby pozycje testowe posiadały zróżnicowaną trudność (od najłatwiejszych do najtrudniejszych patrz *wykład 3*
zróżnicowanie to powinno być tym większe, im większa jest korelacja między pozycjami.
Pozycje testowe dobiera się tak, by średnia trudność pozycji całego testu wynosiła 50%.

W przypadku testów przeznaczonych do celów selekcyjnych pozycje testowe powinny być o takiej trudności, jaka jest najbliższa pożądanemu współczynnikowi selekcji.
Np. jeżeli celem jest wybranie najlepszych 30% kandydatów, to optymalne są pozycje,

których wskaźnik trudności zbliżone są do 30%. Im bliżej punktu selekcji znajdują się wskaźniki trudności pozycji, tym lepiej z uwagi realizację celu pomiaru.

Badanie pilotażowe. Współczynnik mocy dyskryminacyjnej.
Współczynnik mocy dyskryminacyjnej to stopień, w jakim dana pozycja testowa różnicuje badaną populację w zakresie zachowania, które dany test ma mierzyć. Wartość WMD jest interpretowana następująco:
• pozycja testowa o dodatniej mocy dyskryminacyjnej jest częściej rozwiązywana przez osoby badane o wysokich ogólnych wynikach w teście, a więc różnicuje testowanych w zgodzie z innymi pozycjami testu; Do ostatecznej wersji testu powinny wejść tylko o dodatniej i wysokiej mocy dyskryminacyjnej.

Wskaźnik dyskryminacji wymaga ustalenia punktu podziału (w punkcie mediany) osób badanych na dwie grupy: tzw. dolną grupę (tj. grupę o niskich wynikach w teście) i grupę górną (tj. grupę osób o wysokich wynikach w teście).

Operacyjnym wskaźnikiem pozycji danej osoby na kontinuum mierzonego konstruktu jest ogólny wynik w teście.
Im wyższą dodatnią wartość współczynnika korelacji między pozycją a wynikiem ogólnym, tym lepsze właściwości różnicujące, w pożądanym kierunku posiada pozycja testowa. Najczęściej stosowane są trzy współczynniki korelacji, odpowiadające trzem modelowym sytuacjom:

• współczynnik korelacji dwuseryjnej,
(a) rozkład wyników cechy mierzonej przez pozycję testową jest w rzeczywistości zmienną ciągłą o rozkładzie normalnym, a jedynie niedoskonałość narzędzia pomiarowego (pozycji testowej) sprawia, że jest to zmienna dyskretna, oraz
(b) rozkład ogólnych wyników w teście też jest normalny.

• współczynnik korelacji punktowo-dwuseryjnej oraz
(a) rozkład wyników cechy mierzonej przez pozycję testową jest zmienną dychotomiczną oraz
(b) rozkład ogólnych wyników w teście jest rozkładem normalnym.

• współczynnik korelacji punktowo-czteropolowej (phi).
jest obliczany wtedy, kiedy zarówno wynik pozycji testowej, jak i ogólny wynik w teście są traktowane jako zmienne dychotomiczne (taki wynik może dawać dychotomiczne kryterium, np. sukces vs porażka).

Badanie pilotażowe. Współczynnik mocy dyskryminacyjnej.
•Przy obliczaniu współczynnika korelacji między wynikami danej pozycji a ogólnym wynikiem w teście zaleca się, by wynik analizowanej pozycji został wyłączony z ogólnego wyniku testowego;

Badanie pilotażowe. Trafność pozycji. Współczynnik trafności pozycji.
Współczynnik trafności pozycji zależy od wielkości korelacji między wynikami danej pozycji a wynikami interesującego nas kryterium oraz od odchylenia standardowego tej pozycji.

Im wyższe zatem odchylenie standardowe (s) wyników dla danej pozycji, tym większy wkład tej pozycji do wariancji a więc i trafności testu. Współczynnik trafności jest przydatny, gdy celem jest konstrukcja testu o maksymalnej trafności kryterialnej.

Badanie pilotażowe. Rzetelność pozycji. Współczynnik rzetelności pozycji zależy od wielkości korelacji między wynikami danej pozycji z wynikiem ogólnym w teście oraz od odchylenia standardowego tej pozycji.
Wybieranie pozycji o wysokich współczynnikach rzetelności daje narzędzie bardzo homogeniczne (spójne wewnętrznie).

Stronniczości pozycji testowych Stronniczość pozycji testowej = stały błąd pomiaru.
•Stronniczość pozycji testowych polega na tym, że poszczególne pozycje testowe są mniej lub bardziej trudne dla osób należących do różnych podgrup, wyodrębnianych z tej samej populacji a źródłem tej tego są czynniki nie związane z konstruktem mierzonym przez test.
•Ocena stronniczości musi być oparta o dokładna analizę celu tworzonego narzędzia.
•Stwierdzenie, że pozycja testowa w zróżnicowany sposób funkcjonuje w dwóch grupach badanych, nie jest podstawą do stwierdzania ich stronniczości.
•Ale stronniczości należy unikać.

Ostateczna rewizja testu
•Crocker i Algina (1986) proponują zrealizować obie fazy tworzenia testu (ocenę pozycji i walidację krzyżową) w jednym badaniu. Wtedy strategia postępowania jest następująca:
a) wszystkie pozycje testowe wchodzące w skład puli pozycji testowych daje się do rozwiązania bardzo dużej grupie osób badanych;
b)losowo przydziela się część wypełnionych arkuszy testowych do analizy zadań, a część do walidacji krzyżowej. Jeżeli efektem analizy zadań będzie zaakceptowanie 20 pozycji testowych, to wyniki dla tych 20 pozycji z drugiej grupy osób badanych zostaną wykorzystane do oceny trafności testu.

Teoria odpowiadania na pozycje testu
Teoriaodpowiedzi na pozycje testowe (Item response theory; IRT) pozwala na określenie związku pomiędzy odpowiedziami udzielanymi przez osobę badaną a nieobserwowalną cechą leżącą u podstaw zachowań testowych.
Modele formułowane w ramach IRT mają postać funkcji matematycznych, wiążących prawdopodobieństwo udzielenia odpowiedzi prawidłowej (lub zgodnej z kluczem) na daną pozycję testową z ogólnym poziomem mierzonej cechy u osoby badanej.

Ograniczenia modelu klasycznego
W KTT przyjmuje się, że:
•Związek pomiędzy wynikiem prawdziwym a wynikiem otrzymanym w teście jest związkiem prostoliniowym
•Przedziały ufności są takie same dla wszystkich wyników, a wartość błędu pomiaru zależy od badanej próbki.
•Wartość parametrów charakteryzujących pozycje testowe również zależy od próbki.

Ograniczenia modelu klasycznego W IRT:
•Związek pomiędzy wynikiem prawdziwym a wynikiem otrzymanym nie musi być związkiem liniowym;
•Szerokość przedziałów ufności jest inna w środku, a inna na krańcach rozkładu (błąd większy dla wyników skrajnych);
•Błąd standardowy pomiaru nie jest związany z badana próbką osób;
•Parametry opisujące pozycje testowe nie są zależne od badanej próbki osób.
•W ramach IRT oszacowania poziomu badanej cechy dokonuje się oddzielnie dla każdej odpowiedzi testowej, uwzględniając zarazem parametry danej pozycji testu.

Założenia IRT W teorii odpowiadania na pozycje testu przyjmuje się trzy podstawowe założenia:
•o wymiarach przestrzeni latentnej,
•o lokalnej niezależności pozycji testowych i
•o krzywej charakterystycznej pozycji testowej.

Teoria odpowiadania na pozycje testu Założenia IRT Założenie o wymiarach przestrzeni latentnej.
•Test, który mierzy jedną cechę latentna, jest testem jednowymiarowym. Testami jednowymiarowymi są np. testy zdolności (np. matematycznych, językowych czy myślenia technicznego). Wszystkie zależności statystyczne stwierdzane pomiędzy pozycjami testowymi są wyjaśniane przez odwołanie się do jednej cechy latentnej.
•Cechę latentną oznacza się jako theta (Θ) i przyjmuje, że jest ona ciągła, a ponieważ skala jest najczęściej wyrażana w postaci konwencjonalnych wartości z, to w praktyce wszystkie wyniki mieszczą się w przedziale od -4z do +4z.

Założenia IRT Założenie o lokalnej niezależności pozycji testowych
W IRT przyjmuje się, że odpowiedzi każdej osoby badanej na jedną pozycje testową nie zależą od jej odpowiedzi na jakąkolwiek inną pozycję tego testu.
Oznacza to zatem, że rozkład wyników poszczególnych pozycji testowych zależy jedynie od parametru Θ; wyniki pozycji testowych są statystycznie niezależne.
Jeżeli test jest rzeczywiście jednowymiarowy (założenie 1), to założenie o lokalnej niezależności pozycji testowych jest również spełnione.
Wówczas możemy przyjąć, że cecha latentna jest mierzona w sposób niezależny k razy, gdzie k oznacza liczbę pozycji testowych.

Założenia IRT Założenie o krzywej charakterystycznej pozycji testowej
Krzywa charakterystyczna pozycji testowej(item characteristic curve - ICC) to graficzny obraz funkcji matematycznej, wiążącej prawdopodobieństwo udzielenia odpowiedzi prawidłowej na daną pozycję testową z poziomem cechy, operacyjnie wyznaczonym przez ogólny wynik w teście.

•Mierzona cecha ICC jest zmienną ciągłą, a prawdopodobieństwo sukcesu (prawdopodobieństwo udzielenia prawidłowej odpowiedzi na daną pozycję testową) jest funkcją ogólnego poziomu zdolności.
•Ogólny poziom zdolności jest szacowany na podstawie wyniku, jaki osoby badane otrzymały w całym teście.
•Krzywa ICC nie reprezentuje liniowego związku między prawdopodobieństwem sukcesu a

ogólnymi zdolnościami osób badanych.

Parametry pozycji testowej i skala cechy latentnej (ICC)
Każdą ICC można opisać za pomocą trzech parametrów:
–parametru a -współczynnika mocy dyskryminacyjnej (różnicującej),
–parametru b -współczynnika trudności, oraz
–parametru c -współczynnika zgadywania.
Wartości tych parametrów są ustalane empirycznie.

Parametry pozycji testowej i skala cechy latentnej (ICC)
Parametr a. W IRT współczynnikowi mocy dyskryminacyjnej pozycji testowej, czyli parametrowi a, odpowiada na wykresie kąt nachylenia (stopień stromości) krzywej ICC w punkcie przegięcia.
–W klasycznej teorii testów współczynnik mocy dyskryminacyjnej jest miarą tego, jak dobrze dana pozycja testowa różnicuje badaną populację.

Parametr b. Trudność pozycji testowej (bi) jest wartością cechy latentnej θ, dla której prawdopodobieństwo prawidłowej odpowiedzi na daną pozycję [P(θ)] jest równe 0,5. Pozycja jest tym trudniejsza, im wyższa wartość bi.
Trudność, czyli mediana poziomu cechy latentnej dla pozycji, informuje, w którym miejscu na skali θ dana pozycja najlepiej różnicuje badanych. łatwe pozycje dobrze różnicują osoby o niskich umiejętnościach, trudne – o wysokich. Teoretyczny zakres zmienności tego parametru obejmuje przedział (-; +), jednak wartości typowe ograniczają się do przedziału (-3 ; +3).

Parametr c reprezentuje prawdopodobieństwo, z jakim osoba badana o niskich wartości cechy latentnej może odpowiedzieć poprawnie na daną pozycję testową.
•Parametr ten zazwyczaj nazywa się współczynnikiem zgadywania, jako że przyjmuje się, iż osoba badana udzieliła odpowiedzi prawidłowej, stosując strategię nie wynikającą z posiadanej wartości.
•Graficznie współczynnik zgadywania jest reprezentowany za pomocą dolnej asymptoty krzywej ICC.

•W typowej sytuacji testowania prawdopodobieństwo zgadywania oblicza się jako 1/m, gdzie m oznacza liczbę możliwych kategorii. Jednakże w wypadku krzywych ICC wartość ta rzadko będzie równa 1/m. W IRT bowiem przyjmuje się, iż badany, zgadując prawidłową odpowiedź, nie czyni tego w sposób losowy.
•Ponieważ współczynnik zgadywania jest tożsamy z prawdopodobieństwem udzielenia odpowiedzi prawidłowej, dlatego przybiera on wartości od 0,00 do 1,00. W praktyce współczynnik ten najczęściej mieści się w przedziale od 0,00 do 0,40. Im mniejsza wartość c, tym oczywiście lepiej dla testu.

Parametry pozycji testowej i skala cechy latentnej (ICC)
Przewaga krzywych ICC nad klasycznymi wskaźnikami dobroci pozycji testowych polega na tym, że na ich podstawie można określić zależność między prawdopodobieństwem poprawnej odpowiedzi na konkretną pozycję testową a różnymi wartościami cechy latentnej

Teoria odpowiadania na pozycje testu Podsumowanie

•Pomiar psychologiczny jest pomiarem pośrednim. Pozycję danej osoby na kontinuum cechy, która nie jest bezpośrednio obserwowalna (kontinuum latentne), możemy określić tylko na podstawie jej zachowania w ściśle określonych zadaniach. Aby to można było zrobić, musimy dysponować modelem wiążącym konstrukt psychologiczny (cechę latentną) z poziomem zachowań.
•W IRT buduje się modele wiążące poziom nieobserwowalnej cechy psychologicznej z odpowiedzią na każdą kolejną pozycję testową. Zaletą tych modeli jest to, że poziom mierzonej cechy może zostać oszacowany na podstawie każdej pozycji testowej pod warunkiem, że znane są jej parametry, a statystyczne właściwości tych pozycji są bezpośrednio wiązane z zachowaniami testowymi.
•W klasycznej teorii testów model jest prosty: przyjmuje się, że wynik, jaki otrzymała dana osoba w teście, jest sumą dwóch składowych -wyniku prawdziwego tej osoby i błędu pomiaru.

Jeśli walidacja testu potwierdzi jego dobroć psychologiczna następnymi etapami są:
•Normalizacja wyników testu.
•Publikacja testu.
•Aktualizacja norm –najpóźniej po 10 latach.
•Rewizja testu (po około 25 latach lub wcześniej po stwierdzeniu ewidentnych wad diagnostycznych).

Normy testowe są niezbędne dla interpretacji wyników testu.
Wynik surowy osoby testowanej jest nieinterpretowalny bez informacji o wynikach, otrzymanych przez osoby z odpowiedniej grupy odniesienia.
Wyróżniamy dwa rodzaje norm:
•wyniki progowe (pomiędzy grupami kontrastowymi) oraz
•normy bazujące na rozkładzie wyników testu (w grupie odniesienia –normalizacyjnej)

Pojęcie normy psychometrycznej
Interpretacja normatywna lub zorientowana na normy (norm-referenced) = nadawanie znaczenia wynikowi surowemu przez odniesienie go do innych wyników otrzymanych w tym samym teście.
Normą -w sensie psychometrycznym -jest „standard ilościowy, wyznaczony przez średnią, medianę lub inną miarę tendencji centralnej obliczoną dla grupy przedstawicieli danego typu (gatunku)" (Ricks, 1993).
Normą psychometryczną jest zatem typowy wynik w teście otrzymany przez określoną grupę osób.

Istotą normatywnej interpretacji wyników testowych jest odwołanie się do sposobu wykonania danego testu przez określoną grupę osób.
Grupa ta stanowi tzw. grupę odniesienia, inaczej nazywaną też grupą normalizacyjną. Wybór właściwej grupy normalizacyjnej jest krytycznym czynnikiem decydującym o jakości interpretacji wyników testowych.
Korzystanie z norm jest niezbędne w diagnozie. W selekcji (mamy przyjąć określoną liczbę kandydatów) rozsądniej jest odwołać się do wyników surowych i przyjąć tych, którzy uzyskali najwyższe wyniki w teście (pod warunkiem, że wykorzystywany w tym celu test jest trafny).

Znaczenie grupy odniesienia. Zgodnie ze Standardami... (1985a, s. 28): „normy przedstawiane w podręczniku testowym powinny zostać opracowane dla wyraźnie zdefiniowanych populacji. Populacje te muszą odpowiadać tym grupom osób, z którymi badający testem będzie zazwyczaj porównywał osoby badane".
Grupa normalizacyjna musi być reprezentatywna próbką populacji, dla której przeznaczony jest test. Uwaga H.
• **Normy ogólnokrajowe** odwołują się do wyników -reprezentujących z założenia -populację ogólną.
• **Normy lokalne** odwołują się do rozkładów częstości wyników testowych w grupach o mniejszym zakresie i są wykorzystywane dla realizacji wąsko zdefiniowanych celów.

Wyniki progowe (punkty odsiewowe) są stosowane w podejściu zorientowanym na trafność kryterialną.
Celem takich norm jest diagnoza jakościowa –klasyfikacji osób badanych do jednej z n grup jakościowo odmiennych pod względem danego kryterium.
Punkty odsiewowe służą jako norma dla wyników testu – wskazują co oznacza wysoki i niski wynik testu.

Wynik progowy jest zazwyczaj wyrażany w postaci wartości liczbowych wyników surowych stanowiących granicę pomiędzy jakościowo różnymi grupami, np.: “10/11” (najwyższy wynik w jednej grupie/najniższy wynik w drugiej grupie).
Wynik progowy jest estymowany w regresji logistycznej pomiędzy wynikami testu a prawdopodobieństwem przynależności do grup kryterialnych, jako punkt (wynik surowy), w którym prawdopodobieństwo przynależności do którejkolwiek z grup jest większe-równe 50%; poniżej tego wyniku osoba testowana ma mniej niż 50% szans, że należy do wybranej grupy, zaś powyżej, że ma powyżej 50% szans na przynależność.

Diagnoza oparta o progi odsiewowe powinna być poddana walidacji.
Na podstawie odsetków trafnych i błędnych diagnoz można wyliczyć wskaźniki trafności diagnostycznej procedury:
•wrażliwość,
•specyficzność,
•pozytywna wartość predykcyjna,
•negatywna wartość predykcyjna,
•ogólny wskaźnik błędnych klasyfikacji.

Wskaźniki te wylicza się na podstawie liczby diagnoz:
•“prawdziwie negatywnych” (trafienia negatywne; np. osoby zdrowe zdiagnozowane jako zdrowe),
•“prawdziwie pozytywnych” (trafienia pozytywne; np. osoby chore zdiagnozowane jako chore),
•“fałszywie pozytywnych” (fałszywe alarmy; np. osoby zdrowe zdiagnozowane jako chore) i
•“fałszywie negatywnych” (ominięcia; np. osoby chore zdiagnozowane jako zdrowe).

Wskaźnik specyficzności = proporcja osób z “dolnej” grupy kontrastowej poprawnie zakwalifikowanych na postawie wyników testu do ogółu osób z tej grupy, np. liczba osób zdrowych trafnie ocenionych jako zdrowe w stosunku do ogólnej liczby zdrowych = A/(A+C).
Wskaźnik wrażliwości = proporcja osób z “górnej” grupy kontrastowej poprawnie

zakwalifikowanych na postawie wyników testu do ogółu osób z tej grupy, np. liczba osób chorych trafnie ocenionych jako chore w stosunku do ogólnej liczby chorych = D/(B+D).

Wskaźnik pozytywnej wartości predykcyjnej = proporcja osób z "górnej" grupy kontrastowej poprawnie zakwalifikowanych na postawie wyników testu do ogółu osób zakwalifikowanych do tej grupy na podstawie testu, np. liczba osób chorych trafnie ocenionych jako chore w stosunku do ogólnej liczby osób ocenionych jako chore na podstawie inwentarza = D/(C+D).
Wskaźnik negatywnej wartości predykcyjnej = proporcja osób z "dolnej" grupy kontrastowej poprawnie zakwalifikowanych na postawie wyników testu do ogółu osób zakwalifikowanych do tej grupy na podstawie testu, np. liczba osób zdrowych trafnie ocenionych jako zdrowe w stosunku do ogólnej liczby osób ocenionych jako zdrowe na podstawie inwentarza = A/(A+B).

Ogólny wskaźnik błędnych klasyfikacji = liczba diagnoz fałszywych w stosunku do ogólnej liczby osób badanych = (B+C)/(A+B+C+D).

Wskaźniki wskazują w jakim typie diagnozy test wykazuje obniżoną trafność, np. może dobrze diagnozować osoby zdrowe o niskich wynikach, ale "mylić" się w obszarze wyników wysokich, typowych dla osób chorych (ale uzyskiwanych często także przez osoby zdrowe). Efekty te zależą od charakterystyki rozkładu wyników w obu grupach kryterialnych i można nimi manipulować zmieniając próg odcięcia. Jednak zmiana progu zawsze powoduje polepszenie jednego ze wskaźników kosztem pogorszenia wskaźnika komplementarnego.

Normy bazujące na rozkładzie wyników testu
Celem norm opartych na rozkładzie wyników w grupie normalizacyjnej jest uzyskanie diagnozy ilościowej
–oceny intensywności mierzonej cechy (jako własności różnicowej w grupie odniesienia).
Istnieją trzy rodzaje norm:
•normy rangowe (porządkowa skala pomiarowa)
•normy typu równoważnikowego (tzw. równoważniki wieku i równoważniki klasy) oraz
•skale standaryzowane (interwałowa skala pomiarowa)

Normy rangowe. Skala centylowa.
Skala centylowa jest preferowana, gdy rozkład wyników surowych testu znacznie odbiega od rozkładu normalnego (rozkład jest asymetryczny, ma nieprawidłową gęstość i nie ma procedury pozwalającej przetransformowany go w rozkład normalny). W skali centylowej punktem odniesienia (standardem wykonania testu) jest mediana a centyle wskazują na częstość uzyskania danego wyniku w grupie normalizacyjnej.
Centyl to punkt na skali, poniżej którego leży określony odsetek rozkładu.

Zalety i wady skali centylowej:
• Skala centylowa pozwala na ocenę wyniku danej osoby w stosunku do wyników innych osób należących do określonej populacji. Są to wyniki czytelne, i dlatego chętnie stosowane.
• Skala centylowa nie odzwierciedla kształtu rozkładu wyników surowych.
• Rozkład otrzymywany w rezultacie przeliczenia wyników surowych na centyle jest prostokątny -niezależnie od kształtu wyjściowego rozkładu wyników. (Rozkład prostokątny to inaczej rozkład równoprawdopodobny, czyli rozkład, w którym wszystkie wartości zmiennej

pojawiają się z tym samym prawdopodobieństwem.)
•Jeżeli rozkład wyników surowych jest rozkładem normalnym, to skala centylowa prowadzi do przeceniania wielkości różnic pośrodku rozkładu, a niedocenianiu ich na krańcach tego rozkładu (w rozkładzie normalnym najwięcej wyników lokuje się w środku rozkładu).
•Normy centylowe są normami typu rangowego (porządkowego). Oznacza to, że normy tego typu dobrze odzwierciedlają uporządkowanie osób badanych w grupie normalizacyjnej, nie odzwierciedlają natomiast względnych różnic między tymi osobami.

Normy równoważnikowe (rozwojowe)
Normy typu równoważnikowego (rozwojowe) pozwalają określić, „jak daleko na drodze normalnego rozwoju znalazła się jednostka" (Anastasi, Urbina, 1999, s. 84).
Normy te mają charakter opisowy, a wyniki wyrażone w nich „są psychometrycznie surowe i nie nadają się do precyzyjnej obróbki statystycznej".
Rodzaje norm równoważnikowych:
– równoważniki wieku i
– równoważniki klasy.

Normy równoważnikowe (rozwojowe)
Równoważniki wieku to liczby wskazujące na kolejny rok i miesiąc życia badanych osób, odpowiadające średniej arytmetycznej lub medianie wykonania testu na danym etapie rozwoju.
Jednym z rodzajów norm typu równoważników wieku jest tzw. **wiek umysłowy.**

Normy równoważnikowe -Równoważniki klasy.
Równoważniki klasyczne to liczby wskazujące na rok i miesiąc nauczania w roku szkolnym, odpowiadające średniej arytmetycznej lub medianie wykonania testu na danym etapie rozwoju.
Ponieważ rok szkolny liczy zazwyczaj dziesięć miesięcy, dlatego normy tego typu można łatwo wyrazić w systemie dziesiętnym.
Normy tego typu oblicza się w ten sposób, że określa się średni wynik w teście dla dzieci będących aktualnie w określonej klasie.
Wyniki liczbowe, odpowiadające kolejnym miesiącom nauczania, otrzymuje się zasadniczo przez interpolację, choć oczywiście można również badać dzieci w każdym miesiącu nauki szkolnej (Anastasi, Urbina, 1999, s. 86).

Normy standardowe
• Normy standardowe powstają przez przekształcenie wyników surowych otrzymanych w teście na wyniki standardowe z według wzoru.

Normy standardowe –transformacje skali z Ze skalą wyników z nie spotykamy się jednak w podręcznikach testowych.
Interpretowanie wyników testowych w skali z może być kłopotliwe (Punkt 0 nie oznacza początku skali, a wartość średnią.
Nadto co innego oznaczają wyniki ujemne, a co innego wyniki dodatnie). Dlatego zaproponowano, aby dokonując kolejnej transformacji liniowej, przekształcić wyniki zw taki sposób, by początek skali znajdował się po lewej stronie, a kolejne punkty skali miały wyłącznie wartości dodatnie.

Normy standardowe –transformacje skali z

• Transformacja liniowa skali z typu polega na wybraniu dla nowej skali umownej wartości średniej i odchylenia standardowego.

Kulturowa adaptacja testu= dostosowanie oryginalnej wersji testu do warunków użycia w innej kulturze (w innym języku).

Adaptacja testu polega na zrealizowaniu specjalnych procedur decentrujących = uniwersalizujących lub centrujących, w kulturze dla której jest realizowana.
Stosowanie narzędzia, które nie zostało zaadaptowane kulturowo prowadzi do stronniczości.
Aby test nie był stronniczy niezbędne jest kulturowe zrównoważenie testu.

Aspekty równoważności kulturowej testu

1) równoważność teorii psychologicznych,
2) równoważność wymiarów psychologicznych
3) równoważność pojęć psychologicznych,
4) równoważność wskaźników dla cech -zachowań,
5) równoważność procedury badania.

Z psychometrycznego punktu widzenia ocena równoważności kulturowej testu polega na upewnieniu się, że testy (oryginalny i adaptowany) są:

1) równoważne fasadowo (mają identyczną / bardzo podobną formę),
2) równoważne psychometrycznie (mają identyczne / bardzo podobne wskaźniki trafności)
3) równoważne funkcjonalnie (psychologicznie),
4) są wiernymi tłumaczeniami lub adaptacja wiernie rekonstruuje oryginalne narzędzie.

Metody uzyskiwania zaadaptowanego kulturowo narzędzia:

• transkrypcja (wierne tłumaczenie),
• translacja,
• trawestacja,
• parafraza,
• rekonstrukcja.

Transkrypcja= możliwie najwierniejsze tłumaczenie oryginalnych pozycji.
Celem adaptacji poprzez transkrypcję jest dochowanie wierności tłumaczenia oraz równoważności fasadowej testu.
Adaptacja transkrypcyjna wymaga założenia, że mierzone konstrukty oraz ich wskaźniki są uniwersalne (stałe kulturowo)
Z reguły adaptacje transkrypcyjne są ułomne językowo i niedobre psychometrycznie.

Translacja= wierne tłumaczenie oryginalnych pozycji, jednak niezbędnymi modyfikacjami językowymi zapewniającymi kumunikatywność pozycji.
Translacja wymaga przyjęcia założenia, że że konstrukty i ich behawioralne wskaźniki są uniwersalne kulturowo, choć słowa (pojęcia językowe) używane w danej kulturze do opisu zachowania mogą się istotnie różnić od pojęć oryginalnych. (np. słowo respekt, respect).

Trawestacja = kongenialne tłumaczenie oryginalnych pozycji, dopuszczalne jest wprowadzenie licznych modyfikacji, podyktowanych względami językowymi i psychologicznymi (treściowymi lub psychometrycznymi).
Trawestacja wymaga założenia, że mierzone konstrukty psychologiczne są uniwersalne, ale nie ich sposoby artykulacji ich wskaźników (pojęcia) nie są.

Parafraza = opracowanie zupełnie nowego narzędzia, dla którego test oryginalny jest jedynie inspiracją.
Oryginalne pozycje są wykorzystywane jako baza do generowana pozycji testu adaptowanego. Parafraza wymaga założenia, że mierzone konstrukty psychologiczne są uniwersalne kulturowo, ale ich ekspresja w zachowaniach nie jest uniwersalna (szanujesz, to stoisz "na baczność", czy w pozie uniżonej).
Parafraza jest czasochłonna, ale prowadzi do uzyskania wysoce równoważnych psychometrycznie adaptacji testu.

Rekonstrukcja = opracowanie całkowicie nowego narzędzia, dla którego oryginalny model teoretyczny i strategia konstrukcji jest jedynie inspiracją.
Rekonstrukcja wymaga założenia, że zarówno zachowania, jak i konstrukty psychologiczne nie są uniwersalne kulturowo.
Rekonstrukcja często prowadzi do zupełnie innej niż oryginalna wersji testu i jest równie czaso-i zasobochłonna jak konstrukcja oryginalnego testu.
Rekonstrukcja prowadzi do uzyskania całkowicie dostosowanych kulturowo wersji testu.

Adaptacja demograficzna = przystosowanie testu oryginalnie przeznaczonego do badania określonej grupy demograficznej (płeć, wiek, niepełnosprawność, rasa, status społecznoekonomiczny, środowisko subkultura) do stosowania w innej grupie.
Adaptacja demograficzna testu wymaga zastosowania podobnych procedur przystosowujących jak adaptacja kulturowa.

Interpretacja wyników testowania
Interpretacja, to formułowanie wniosków psychologicznych na podstawie wyników testu; nadawanie wynikom testu znaczenia psychologicznego.
Proces interpretacji wyników testu jest zakotwiczony w trafności pomiaru i polega na bezpośrednim odwołaniu się do trafności teoretycznej i/lub kryterialnej.

Interpretacja zwykle przybiera postać:
a) opisu zachowań osoby poddanej testowaniu (cech osobowości lub zdolności i inteligencji);
b) przewidywań zachowań osoby poddanej testowaniu w warunkach naturalnych, jej funkcjonowania w życiu codziennym.
Interpretacje opierają się na danych empirycznych zebranych w trakcie walidacji testu i ma oczywiście charakter wnioskowania indukcyjnego (a więc nie gwarantującego pewności).

Podstawą (warunkiem koniecznym, choć niewystarczajacym) dla trafności interpretacji jest rzetelne oszacowanie natężenia mierzonej cechy oraz wyznaczenie przedziału ufności dla tego oszacowania.
Rodzaje interpretacji wyników testowych
1) Interpretacja kliniczna

2) Interpretacja statystyczna

Interpretacja kliniczna
• Najczęściej przybiera postać wniosków jakościowych, a jeśli są to wnioski ilościowe, to oszacowania są bardzo zgrubne.

• Interpretowany jest profil wyników w poszczególnych skalach i jest skoncentowana na stworzeniu idiograficznego “portretu” psychologicznego osoby poddanej testowaniu. Podstawa interpretacji jest agregacja treści psychologicznych poszczególnych skal oraz “odgadnięcie” znaczenia uzyskanej (profilowej) konfiguracji cech. Interpretacja kliniczna wiąże się z idiograficznym podejściem do diagnostyki psychologicznej i w jej efekcie formułowane są wnioski raczej jakościowe niż ilościowe, a jeśli ilościowe -to są one mało precyzyjne

Interpretacja statystyczna
• Charakterystyczna dla nomotetycznego podejściem do diagnostyki psychologicznej, dostarcza przede wszystkim wniosków ilościowych, ale wymaga sformalizowania procesu wnioskowania na podstawie uzyskanych danych.

• Interpretowany jest agregat wyników testowych, ale skale łączone są wg kryteriów statystycznej mocy predykcyjnej lub dyskryminacyjnej (np. dla stanu zdrowia psychicznego, sukcesu w nauce, powodzenia w wykonywaniu określonej pracy, itd.).

• Dla interpretacji tych tworzy się liczbowe wskaźniki pewności (dokładności).

Interpretacja kliniczna vs statystyczna
•Przewidywania zachowań są zdecydowanie bardziej trafne przy wnioskowaniu statystycznym.
•Podejście statystyczne jest mechaniczne i nie wymaga w zasadzie wiedzy psychologicznej oraz jest mało twórcze.

Struktura podręcznika do testu
•Prezentacja teoretycznych podstaw testu.
•Opis procedury konstrukcji i walidacji testu.
•Opis danych ilustrujących rzetelność i trafność pomiaru (wraz z charakterystyką osób).
•Opis procedury praktycznego stosowania testu i obliczania wyników.
•Opis procedury interpretacji wyników testu.
•Tabelaryczne zestawienia normalizacyjne.