

Klucz Bremera: opracowanie i walidacja wrażliwego na potrzeby dzieci testu kompetencji językowych dla nauczycieli

Tamas Rotschild¹

Universität Bremen

Fachbereich 12 – Erziehungs- und Bildungswissenschaften

<https://orcid.org/0000-0001-9244-2815>

Streszczenie

Cel: Język jest potężnym narzędziem w rękach nauczycieli, ale nie jest neutralny – niesie za sobą niuanse, które wykraczają poza dosłowność. Te niuanse naładowane emocjami i ocenami, wywierają skumulowany wpływ na wewnętrzny świat dzieci. Dlatego w zawodzie nauczyciela, oprócz wrażliwości na potrzeby uczniów, niezbędne są wysokie kompetencje językowe.

Metoda: W celu podniesienia kompetencji językowych przyszłych pedagogów, w ramach badań doktoranckich na Uniwersytecie w Bremie w Niemczech, opracowano i zweryfikowano *Test oceny sytuacyjnej*. Pozytcje testowe oraz opcje odpowiedzi przygotował zespół trzech czynnych nauczycieli, kierowany przez doktoranta i wspierany przez dwóch profesorów uniwersyteckich z doświadczeniem w tej dziedzinie, zgodnie z ustalonymi praktykami rozwoju testów.

Wyniki: W testach pilotażowych wzięło udział 47 nauczycieli z różnych szkół publicznych oraz 55 studentów z trzech uczelni. Po iteracyjnych usprawnieniach test sfinalizowano w formie 13 pozycji, wykazując silną spójność wewnętrzną, co potwierdza obliczona wartość $\omega = 0,81$. Test i jego użyteczność zostały pozytywnie ocenione przez uczestników.

Konkluzja: W badaniu stwierdzono, że narzędzie to może być wykorzystywane w różnych dziedzinach edukacji oraz w programach szkolenia nauczycieli w celu poprawy praktyk i wyników edukacyjnych.

Słowa kluczowe: kompetencje językowe, perspektywa dziecka, *Test oceny sytuacyjnej*, ocena edukacyjna, szkolenie nauczycieli

¹ Adresy do korespondencji: tam_rot@uni-bremen.de; rotschild.tommy@gmail.com.

Klucz Bremera to narzędzie oceny zaprojektowane w celu sprawdzenia kompetencji nauczycieli w zakresie stosowania określonych wyborów semantycznych i pragmatycznych, które kształtują językowe oraz psychologiczne właściwości komunikacji uwzględniającej potrzeby dzieci (Rotschild, 2024). Według Rotschilda komunikaty konstruowane i przekazywane w języku wrażliwym na potrzeby dziecka mają pozytywny wpływ na intrapsychiczny świat dzieci, wspierają ich dobrostan emocjonalny, poczucie własnej wartości oraz rozwój poznawczy. Regularne komentarze i informacje zwrotne nauczycieli na temat cech wyników i zachowań dzieci w trakcie typowego dnia szkolnego mają istotny wpływ na ich rozwój (Johnston, 2004). Komunikaty nauczycieli stanowią główne źródło doświadczeń psychologicznych dzieci, pomagając im w zrozumieniu siebie oraz kształtowaniu własnego obrazu (Burns, 1982). Znaczenie posiadania pozytywnego obrazu samego siebie jest niezwykle ważne. Badania wykazują, że samoakceptacja ma kluczowe znaczenie dla rozwoju i zachowania zdrowia psychicznego, wpływając na różnorodne psychologiczne, behawioralne i edukacyjne rezultaty (Craven i Marsh, 2008; Ybrandt, 2007). Język, choć jest potężnym narzędziem w kształtowaniu postrzegania siebie i zachowania bez użycia siły, nie jest neutralny. Każde wypowiedziane słowo i fraza przekazują znaczenia wykraczające poza dosłowność, niosąc za sobą emocje i oceny. Nie wywierają one natychmiastowego, magicznego efektu, lecz raczej działają poprzez skumulowany wpływ, zarówno indywidualnie, jak i w sposobie konstruowania zdań (Bolinger, 1980). Poprzez komunikaty nauczyciele nieświadomie przekazują dzieciom, jakimi osobami są w ich oczach, zachęcając je do bycia takimi ludźmi (Johnston, 2004). Dzisiejsze komunikaty stają się jutrzejszą samoakceptacją dziecka. Dlatego niezwykle istotne jest, aby nauczyciele świadomie wybierali język, który wzmacnia samoocenę dzieci i pozwala im definiować siebie w pozytywny sposób.

Niniejszy artykuł omawia rozwój koncepcyjny oraz walidację *Klucza Bremera – Testu oceny sytuacyjnej (SJT)*, opracowanego w celu oceny wrażliwych na potrzeby dzieci kompetencji językowych nauczycieli oraz przewidywania ich zachowań komunikacyjnych podczas kontaktów z dziećmi w sytuacjach trudnych i wymagających wrażliwości. Stosowanie *Klucza Bremera* może dostarczyć danych, które pomogą w opracowywaniu materiałów instruktażowych i kursów szkoleniowych, mających na celu poprawę umiejętności przyszłych nauczycieli w tworzeniu pozytywnego, wspierającego środowiska nauki i rozwoju dzieci poprzez skuteczne używanie języka. Oprócz potencjalnej użyteczności w szkoleniu nauczycieli *Klucz Bremera* ma szerokie zastosowanie w szkolnictwie. Może być wykorzystywany w ramach ciągłego rozwoju zawodowego nauczycieli, przyczyniając się do doskonalenia porozumiewania się w klasie. Co więcej, jego użycie w programach nauczania dla przyszłych nauczycieli może pomóc w przygotowaniu ich do efektywnego nawiązywania relacji z uczniami. W kontekście rekrutacji i selekcji nauczycieli instytucje edukacyjne mogłyby włączyć ten instrument do procesu naboru. Dzięki systematycznej ocenie kompetencji językowych kandydatów oraz ich skłonności do określonych zachowań komunikacyjnych można by wybrać osoby najlepiej przygotowane do skutecznej pracy edukacyjnej. Instrument ten byłby także wsparciem przy tworzeniu strategii edukacyjnej. Decydenci mogliby wykorzystać jego wyniki do formułowania zasad mających na celu poprawę ogólnej jakości

nauczania, szczególnie w zakresie wspierania efektywnego używania języka i komunikacji w klasie. Podsumowując, instrument ten może być kluczowym czynnikiem w ciągłych wysiłkach na rzecz poprawy praktyk i wyników edukacyjnych.

Metoda

Opracowanie Testu oceny sytuacji w do oszacowania kompetencje językowe nauczycieli

Korzystając z ustalonych działań rozwojowych (Delgado-Rico i in., 2012; Patterson i in., 2015; Pollard i Cooper-Thomas, 2015; Whetzel i in., 2020) oraz opierając się na wynikach najnowszych badań (Reiser i in., 2022; Smith i in., 2020, 2022), budowa *Klucza Bremera* przebiegała w ośmiu krokach, co ilustruje rysunek 1. Wśród uczestników znaleźli się eksperci, nauczyciele oraz studenci niemieckich uniwersytetów z wydziałów edukacji lub pokrewnych dziedzin (np. psychologii i pracy socjalnej w szkole). Na wszystkich etapach uczestnicy byli informowani o celu badania, a ich udział był dobrowolny. Testy były anonimowe; od uczestników nie wymagano podawania żadnych danych osobowych, co zapewniło im anonimowość. Adresy IP nie były rejestrowane ani przechowywane.

Rysunek 1

Proces konstruowania Klucza Bremera



Krok 1: Definicja pojęcia

Pojęcie „języka wrażliwego na potrzeby dziecka” (CSL – *Child-Sensitive Language*), które ma być oceniane, skonceptualizowano bazując na wynikach badań przedstawionych przez Rotschilda (2023) – zob. tabela 1 (s. 24).

Tabela 1*Konceptualizacja pojęcia „język wrażliwy na dziecko” i jego kluczowych komponentów***Język wrażliwy na potrzeby dziecka (CSL)**

Język opracowany w celu uwzględnienia zdolności poznawczych, indywidualności, wrażliwości emocjonalnej dzieci oraz kształtowania postrzegania siebie. Uwzględnia unikalne doświadczenia i perspektywy dzieci, mając na celu pielęgnowanie ich poczucia własnej wartości, pewności i pozytywnego obrazu siebie.

CSL₁ Język świadczący o zrozumieniu i wrażliwości na emocje dziecka i jego potrzeby oraz doświadczenia.

CLS₂ Język, który powstrzymuje się od wartościowania lub krytyki.

CLS₃ Język, w którym unika się używania słów, zwrotów lub tonów, które mogą wywoływać strach lub niepokój u dzieci.

CLS₄ Język wyrażający szczere uznanie i zachętę dla dzieci.

CLS₅ Język, który oferuje dzieciom słowa wsparcia, otuchy i wskazówki.

Adnotacja. Na podstawie Delgado-Rico i in. (2012).

CSL jest niejawnym, wielowymiarowym konstruktym wyższego rzędu, zdefiniowanym przez Lawa i in. (1998). Integruje on kilka powiązanych, ale odrębnych konstruktów niższego rzędu, z których każdy ma własny zestaw obserwowalnych wskaźników (Johnson i in., 2012). Każdy z tych konstruktów w unikalny sposób przyczynia się do nadrzędnej koncepcji CSL, czyniąc ją bogatszą i bardziej znaczącą w porównaniu z poszczególnymi wymiarami rozpatrywanymi osobno. Interakcje między różnymi wymiarami mogą tworzyć synergie, które zwiększają ogólną funkcjonalność i wpływ konstruktów (Carver, 1989).

Krok 2: Specyfikacja testu

Na tym etapie opracowano szczegółowe ramy rozwoju systemu oceniania. Specyfikacje obejmowały określenie grupy docelowej, celu jej testowania oraz wybranej metody. Z uwagi na kluczową rolę nauczycieli, psychologów oraz szkolnych pracowników socjalnych w kształtowaniu samooceny dzieci poprzez ich komentarze dotyczące wyników w nauce i zachowania to właśnie oni zostali uznani za grupę docelową. Mając na uwadze, jak istotną rolę pełni język w ich zawodzie, za cel testu przyjęto ocenę kompetencji językowych praktyków wrażliwych na potrzeby dzieci. Aby skutecznie zrealizować ten cel, wybrano *Test oceny sytuacji* (SJT) z pytaniami wielokrotnego wyboru ze względu na jego sprawdzoną skuteczność w zapewnianiu trafnej, rzetelnej i opartej na dowodach oceny pozanaukowych umiejętności, takich jak komunikacja, które są kluczowe dla efektywnej pracy edukacyjnej (Al Hashmi i Klassen, 2020; Durksen i Klassen, 2018). W teście SJT badanym prezentuje się scenariusze związane z ich pracą, a także listę możliwych działań do wyboru. Uczestnicy muszą wybrać najlepsze działanie lub takie, które najprawdopodobniej podjęliby w danej sytuacji (McDaniel i Nguyen, 2001; McDaniel i in., 2007). Test SJT ma wiele zalet, do których należą łatwość badania dużych grup przy jednoczesnym pośrednictwie Internetu,

silne właściwości psychometryczne oraz pozytywny odbiór przez uczestników (Lievens i in., 2008).

Krok 3: Scenariusze i opcje odpowiedzi

Na tym etapie opracowano 36 rzeczywistych scenariuszy, z których każdy zawierał trzy opcje odpowiedzi. Scenariusze stworzono we współpracy z trzema nauczycielami, a ich poprawność i praktyczne znaczenie zweryfikowało dwóch ekspertów. Dystraktory, czyli nieprawidłowe odpowiedzi, zaczerpnięto z życiowych sytuacji, aby odzwierciedlać powszechne błędne przekonania lub nieporozumienia obecne w środowisku zawodowym. Takie podejście zapewnia autentyczność testu i jego adekwatność do rzeczywistych wyzwań, z którymi mierzą się praktycy. Klucze, czyli prawidłowe odpowiedzi, oparto na wytycznych Rot-schilda (2024) dotyczących komunikacji wrażliwej na potrzeby dziecka, bazując na dowodach empirycznych. Dzięki temu scenariusze i opcje odpowiedzi bezpośrednio oceniają kompetencje w zakresie stosowania strategii komunikacji wrażliwej na potrzeby dziecka, co czyni test solidnym narzędziem do oceny profesjonalistów w tej dziedzinie. Tworzenie pytań testowych było zgodne z sugestiami Haladyna i in. (2002). Decyzję o ograniczeniu liczby opcji do trzech na pytanie podjęto na podstawie wyników badań empirycznych (Rodríguez, 2005). Tabela 2 przedstawia typowy przykład pytania testowego.

Tabela 2

Przykładowy element testu

W klasie uczniowie otrzymali zadanie pracy nad zestawem ćwiczeń matematycznych. Leon jednak siedzi przy biurku i nie rozpoczął pracy. Wydaje się niezaangażowany, rozgląda się po klasie, podczas gdy inni uczniowie pracują.

- a Leonie, co mogę zrobić, aby pomóc ci rozpocząć ćwiczenia?
 - b Leonie, dlaczego nie wykonujesz ćwiczeń?
 - c Leonie, czy chcesz zacząć wykonywać ćwiczenia?
-

Krok 4: Konstrukcja testu

Pozycje testowe oceniono pod kątem reprezentatywności i podobieństwa treści, co doprowadziło do wykluczenia 12 z nich. Kolejne 24 pozycje podzielono na dwa zestawy: wersję A i wersję B. Każda pozycja w wersji A miała odpowiadającą jej pozycję w wersji B, tworząc łącznie 24 unikatowe pytania rozłożone na oba zestawy. Klucze odpowiedzi zostały równomiernie rozdzielone pomiędzy różne opcje, zgodnie z zaleceniami Haladyna i Downinga (1985), co zapewniło równą liczbę poprawnych odpowiedzi dla każdej opcji. Rozkład poprawnych odpowiedzi był zrównoważony i losowy, aby uniknąć przypadkowego grupowania odpowiedzi.

Krok 5: Pilotaż 1.

Każdą wersję testu przeprowadzono w formie bezpośredniej w kwietniu 2023 r., w formacie papier-olówek, w grupie 20 nauczycieli (nazywanych „egzaminowanymi ekspertami”), którzy uczestniczyli w programie rozwoju zawodowego na uniwersytecie. Wśród uczestników było 11 kobiet i 9 mężczyzn. Ich doświadczenie zawodowe wahało się od 4 do 18 lat ($M = 7,8$).

Krok 6: Ocena i weryfikacja

Faza oceny

Celem tej fazy było: (1) wybranie pozycji testowych z obu wersji do włączenia w ostateczną wersję testu oraz (2) ich poprawa i modyfikacja w celu uzyskania optymalnej skuteczności. Proces oceny rozpoczął się od analizy poziomu trudności (p) każdej pozycji testowej. Choć statystyczna moc analizy jest ograniczona przy próbie liczącej 10 osób, obliczenie poziomu trudności dostarczyło cennych wstępnych informacji o wydajności każdej pozycji, co pomogło zidentyfikować elementy wymagające modyfikacji. Wyniki przedstawiono w tabeli 3.

Tabela 3*Dane analizy pozycji*

Pozycja (I)	Wersja A p	Wersja B P
1	0,30	0,30
2	0,20	0,70
3	0,70	0,01
4	0,30	0,20
5	0,60	0,20
6	0,10	0,40
7	0,30	0,30
8	0,00	0,50
9	0,20	0,60
10	0,10	0,40
11	0,60	0,10
12	0,00	0,40

Dane analityczne zostały uzupełnione o komentarze profesora uniwersyteckiego, eksperta w tej dziedzinie, oraz 13 z 20 uczestników, którzy wzięli udział w dobrowolnej otwartej dyskusji grupowej, przeprowadzonej tydzień po teście. W wyniku tej analizy wybrano osiem pozycji (1, 2, 3, 4, 5, 7, 8 i 12) z wersji A oraz pięć pozycji (1, 2, 4, 5 i 12) z wersji B jako potencjalne do włączenia do drugiego projektu testu. Kryteria wyboru obejmowały jasność, niski poziom złożoności

poznawczej oraz typowość scenariuszy opisywanych przez te pozycje. Na podstawie komentarzy uczestników opracowano także 14 pozycję, którą włączono do poprawionej wersji testu. Liczbę pozycji ograniczono do 14, aby test był bardziej komfortowy dla uczestników, nie powodował zmęczenia oraz był łatwy w przeprowadzeniu, punktowaniu i interpretacji wyników.

Faza weryfikacji

Pozycje testowe poprawiono na podstawie komentarzy i wniosków z przeglądu literatury (Boland i in., 2010; Holland i Stevens, 2021; Mayer i Moreno, 2003; Zhang i in., 2022), aby wyrównać poziom trudności, zapewnić odpowiedni rozkład cech dyskryminacyjnych oraz prawidłową ocenę kompetencji. Imiona zneutralizowano pod względem płci tam, gdzie pozwalał na to kontekst językowy, aby uniknąć uprzedzeń związanych z płcią. Stereotypy złagodzono w celu zapewnienia etyczności testu i dokładności wyników. Język uproszczono, tak aby zminimalizować dwuznaczność, a długość pytań skrócono, eliminując zbędne informacje zarówno z treści pytań, jak i z opcji odpowiedzi. Dodatkowo zmieniono lub zastąpiono dystraktory, aby poprawić ich funkcjonalność.

Poprawiony test opracowano za pomocą LimeSurvey i umieszczono na serwerze Uniwersytetu w Bremie. Do 14 specyficznych dla danej dziedziny pytań dodano 5 pytań zwrotnych w celu uzyskania opinii od uczestników:

1. Czy uważasz, że scenariusze przedstawione w teście są realistyczne i reprezentatywne dla rzeczywistych sytuacji komunikacyjnych w klasie?
2. Czy opcje odpowiedzi były adekwatne i odpowiednie do danego scenariusza?
3. Czy uważasz, że opcje odpowiedzi były wiarygodne w kontekście podanych scenariuszy?
4. Czy elementy testu były jasno sformułowane?
5. Czy uważasz, że test umożliwia ocenę umiejętności rozróżniania pozytywnie i negatywnie sformułowanych komunikatów?

Każde pytanie zawierało 5-punktową skalę odpowiedzi: (a) *Tak/ Bardzo dobrze* [5], (b) *Przeważnie/ Dobrze* [4], (c) *Umiarkowanie* [3], (d) *Raczej nie/ Słabo* [2], (e) *Wcale* [1]. Na końcu testu dodano pytanie otwarte: „Jak oceniasz treść i strukturę tego testu? Czy skłonił Cię on do refleksji nad własną komunikacją? Czego nauczyłeś(-aś) się z tego testu?” w celu zebrania dodatkowych opinii od uczestników.

Krok 7: Pilotaż 2.

Drugi pilotaż odbył się online w listopadzie i grudniu 2023 r. Test skonfigurowano tak, aby uniemożliwić uczestnikom pomijanie pytań, wylogowanie się i ponowne logowanie, a także dwukrotne przystąpienie do testu. Nie było limitu czasu na jego ukończenie. W badaniu wzięło udział 82 uczestników: 27 nauczycieli akademickich oraz 55 studentów uniwersytetów studiujących edukację lub pokrewne dziedziny, takie jak psychologia, w trzech niemieckich instytucjach. Nauczyciele zostali zrekrutowani poprzez bezpośrednie wiadomości e-mail, natomiast studenci otrzymali link do testu pośrednio od swoich profesorów.

Krok 8: Ocena i weryfikacja 2.

Faza oceny

Ocena na tym etapie miała na celu zweryfikowanie wprowadzonych zmian, aby potwierdzić, że skutecznie poprawiają one wiarygodność i przejrzystość pytań. Dodatkowo analiza miała zidentyfikować ewentualne pozostałe lub nowe problemy, zapewniając, że pytania są wolne od błędów i tworzą rzetelne oraz trafne narzędzie oceny.

Wyniki

Analiza pozycji

Poszczególne pozycje przeanalizowano w celu określenia ich poziomu trudności (p) i mocy dyskryminacyjnej (D). Wyniki przedstawiono w tabeli 4.

Tabela 4

Dane z analizy wydajności 2. pozycji

Pozycja	Oryginalny przedmiot	p	D
1	A1	0,30	0,50
2	B5	0,50	0,20
3	A5	0,70	0,30
4	A2	0,00	-0,10
5	B2	0,70	0,10
6	A4	0,70	0,70
7	A8	0,80	0,50
8	B1	0,20	0,40
9	B4	0,70	0,50
10	A3	0,50	0,20
11	–	0,50	0,40
12	A12	0,50	0,60
13	B12	0,60	0,50
14	A7	0,60	0,00

Adnotacja. A i B odnoszą się do wersji testowej w pilotażu 1.

Większość pozycji, z wyjątkiem pozycji 4 i 7, mieści się w średnim zakresie trudności ($0,30 < p < 0,70$), co oznacza, że mają potencjał do dostarczania zróżnicowanych informacji. Większość pozycji przekracza akceptowaną wartość (0,30),

co wskazuje na dobrą moc dyskryminacyjną, a różnorodne wskaźniki dyskryminacji pozwalają na dokładną ocenę szerokiego zakresu umiejętności (Ebel i Frisbie, 1991). Średnie poziomy trudności przyczyniają się do dobrze zbalansowanej konstrukcji testu. Elementy o umiarkowanej trudności i niższym współczynniku dyskryminacji mogą uzupełniać te trudniejsze z wyższym, co tworzy kompleksowy instrument oceny.

Warto jednak zauważyć, że pozycja 14 wykazała wskaźnik dyskryminacji równy 0,001, a ujemny wskaźnik dyskryminacji w pozycji 4 prawdopodobnie wynikał z uprzedzeń związanych z oczekiwaniami społecznymi. Oznacza to, że uczestnicy częściej wybierali odpowiedzi zgodne z normami społecznymi, przedstawiające ich w korzystnym świetle, niż poprawne, choć mniej społecznie pożądane odpowiedzi. Wysoka liczba błędnych odpowiedzi oraz wynikający z nich poziom trudności dodatkowo przyczyniały się do ujemnego wskaźnika dyskryminacji tej pozycji. Niskie wskaźniki dyskryminacji w pozycjach 2, 5 i 10 mogą wynikać z faktu, że koncentrują się one na ocenie podstawowej wiedzy. Chociaż te pozycje mogą słabo różnicować uczestników o niskich i wysokich wynikach, to odgrywają kluczową rolę w ocenie, czy uczestnicy mają podstawową wiedzę z zakresu testowanego przedmiotu. Dodatkowo niska moc dyskryminacyjna pozycji 2 i 10 mogła być spowodowana niefunkcjonalnymi dystraktorami.

Analiza dystraktorów

Na podstawie zaleceń Berka (1984) oraz Raymonda i in. (2019) siedem dystraktorów uznano za niefunkcjonalne, ponieważ wybrało je mniej niż 5% uczestników.

Rzetelność

Współczynnik *omega* (ω) obliczono w celu oszacowania rzetelności testu. W przeciwieństwie do powszechnie stosowanego współczynnika *alfa* (α), który zakłada równoważność *tau* (τ) (czyli równe ładunki czynnikowe), współczynnik *omega* uwzględnia różnice w relacjach między pozycjami a czynnikami oraz różne wartości ładunku czynnikowego oraz błędy związane z charakterem danej pozycji, co zapewnia dokładniejsze oszacowanie rzetelności w wielowymiarowych konstrukcjach (Gignac, 2013; Hayes i Coutts, 2020; Watkins, 2017). Obliczona wartość $\omega = 0,81$ przekracza standardowy poziom $\alpha = 0,7$, co wskazuje na dobrą spójność wewnętrzną (Kline, 2000).

Proces analizy obejmował trzy etapy, wykorzystując funkcjonalności SPSS i Microsoft Excel:

1. Przeprowadzono analizę czynnikową metodą maksymalnego prawdopodobieństwa, wyodrębniając 1 czynnik w celu oceny korelacji każdej pozycji z badanym konstruktą. Pozycja 14 została wykluczona ze względu na niski ładunek czynnikowy (0,025).
2. Przeprowadzono analizę czynnikową metodą maksymalnego prawdopodobieństwa, wyodrębniając 5 czynników odpowiadających różnym aspektom konstruktów oraz zidentyfikowano główne obciążenie każdego elementu.

3. Współczynnik (ω) obliczono zgodnie z ustalonym wzorem:

$$\omega = \frac{(\sum\lambda)^2}{(\sum\lambda)^2 + (\sum\varepsilon)}$$

gdzie suma $(\sum\lambda)^2$ kwadratów ładunków czynników pierwotnych i $(\sum\varepsilon)$ jest sumą wariancji błędu.

Obliczając rzetelność dla nadrzędnego konstruktów przy użyciu najwyższego obciążenia każdej pozycji, ocena niezawodności koncentrowała się na głównym konstrukcie zainteresowania. Uwzględnienie tylko najwyższych ładunków maksymalizuje wkład każdej pozycji w rzetelność i zwiększa trafność konstruktów, ponieważ pozycje te mają silniejszy związek z konstruktem bazowym. Pozycje o wysokim ładunku zapewniają także, że konstrukct jest dobrze reprezentowany w różnych jego wymiarach, co wspiera trafność treściową. Niemniej jednak takie podejście może prowadzić do utraty wkładu wieloczynnikowego. Pozycje, które obciążają wiele czynników, mogą przyczyniać się do wyniku złożonego jedynie na podstawie ich najwyższego obciążenia, co potencjalnie zaciemnia ich znaczenie dla czynników drugorzędnych. Wreszcie ta metoda może przeoczyć szerszy wkład pozycji o umiarkowanych, ale znaczących ładunkach.

Istotność

Istotność instrumentu potwierdzono na kilka sposobów. Przede wszystkim jego projekt był zgodny z ustalonymi procedurami. Oceniany konstrukct jasno zdefiniowano i oparto na dowodach empirycznych. Scenariusze, które odzwierciedlają rzeczywiste sytuacje w klasie zaczerpnięto z doświadczeń praktyków (ekspertów), nadając instrumentowi praktyczne zastosowanie. Opcje odpowiedzi zostały starannie opracowane na podstawie operacjonalizacji konstruktów, jak przedstawiono to w pracy Rotschilda (2023). Dodatkowo pozycje testowe poddano przeglądowi, ocenie i zatwierdzeniu przez dwóch ekspertów. Proces walidacji także wzbogacono o cenne spostrzeżenia zebrane podczas pierwszej fazy pilotażu, kiedy informacje zwrotne od uczestników dostarczyły jakościowej wiedzy na temat cech testu. Druga faza pilotażu zapewniła kluczowe dane statystyczne, które potwierdziły ogólną trafność testu.

Uczestnicy testu ($N = 82$) ocenili pozycje jako jasno napisane, ze średnim wynikiem 4,86 ($SD = 0,56$), co świadczy o wysokim poziomie jasności i wspiera ważność merytoryczną testu. Scenariusze testowe uzyskały średnią ocenę 3,97 ($SD = 0,82$), co potwierdza ich autentyczność i zwiększa trafność treści testu. Opcje odpowiedzi osiągnęły średnią 4,13 ($SD = 0,83$), co mówi o tym, że test mierzy realistyczne sytuacje decyzyjne, co również wzmacnia trafność treści. Ogólną trafność konstruktów i treści potwierdzono średnim wynikiem 3,76 ($SD = 0,79$). Ponadto test oceniono jako skuteczny w ocenie kompetencji komunikacyjnych wrażliwych na potrzeby dziecka, ze średnim wynikiem 3,94 ($SD = 0,73$), co dodatkowo przyczynia się do trafności konstruktów.

Użyteczność

Spośród różnych tematów dostępnych do dyskusji 35 uczestników (43%) zgłosiło w odpowiedzi na pytanie otwarte, że test skłonił ich do refleksji nad

własnym użyciem języka i znaczeniem subtelności językowych. To potwierdza praktyczność testu oraz jego potencjał jako narzędzia do rozwijania kompetencji komunikacyjnych.

Faza weryfikacji

Na podstawie danych psychometrycznych, poziomu złożoności oraz podobieństwa treści do innych pozycji pozycja 14 została wykluczona. Zastąpiono siedem niefunkcyjnych dystraktorów. Pozycję 4, wraz z pozycjami 2, 5 i 10, zachowano ze względu na ich diagnostyczny potencjał i pokrycie podstawowych umiejętności w testowanej dziedzinie. Ostateczna wersja instrumentu, składająca się z 13 pozycji, jest gotowa do wykorzystania w ocenach na dużą skalę.

Dyskusja

Opracowanie i walidacja *Klucza Bremera* stanowi znaczący postęp w ulepszaniu programów szkolenia nauczycieli. *Test oceny sytuacyjnej* (SJT) wykazuje dużą trafność merytoryczną i rzetelność w ocenie kompetencji językowych wrażliwych na potrzeby dzieci, zarówno wśród nauczycieli stażystów, jak i nauczycieli praktyków. Równowaga w zakresie trudności elementów i mocy dyskryminacyjnej zapewnia, że test jest nie tylko wymagający, lecz także sprawiedliwy, oferując szczegółową ocenę zdolności egzaminowanych do skutecznego poruszania się w interakcjach wrażliwych na potrzeby dzieci. Informacje zwrotne od uczestników potwierdziły autentyczność scenariuszy testowych oraz wiarygodność opcji odpowiedzi. Ponadto wzmocniono użyteczność testu w podnoszeniu świadomości znaczenia niuansów językowych oraz zachęcaniu do refleksji nad własnym użyciem języka. Dzięki solidnym właściwościom psychometrycznym i możliwości zastosowania, narzędzie to stwarza znaczącą szansę na poprawę nauczania i w rezultacie wyników edukacyjnych. *Klucz Bremera* jest przeznaczony do stosowania online przez nauczycieli stażystów na różnych uniwersytetach, koncentrując się na pięciu kluczowych domenach używania języka wrażliwego na potrzeby dzieci. Wyniki uzyskane przez stażystów będą oceniane na podstawie wcześniej ustalonego punktu odcięcia, który odzwierciedla poziom osiągniętych umiejętności, jakiego można słusznie oczekiwać od skutecznych praktyków. Po zebraniu danych zastosowane zostaną statystyki opisowe w celu podsumowania ogólnych wyników, w tym średnich wyników, odchylenia standardowego oraz odsetki stażystów, którzy spełniają wymagany poziom kompetencji. Wyniki następnie zebrano w szczegółowe raporty i udostępniono kierownikom wydziałów na każdym z uniwersytetów. Raporty te mają na celu zapewnienie wglądu w obszary, w których stażyści wyróżniają się, oraz w obszary, w których potrzebny jest ich dalszy rozwój. Wykładowcy będą zachęceni do korzystania z tych danych, aby zdobyć informacje o rozwoju edukacji komunikacyjnej specyficznej dla zawodu, co umożliwi tworzenie programów poprawiających profesjonalne

umiejętności stażystów. Jednak ważnym ograniczeniem *Testu Bremera* jest brak możliwości oceny niewerbalnych aspektów komunikacji takich jak język ciała, mimika i gesty, co w rzeczywistych warunkach nauczania, może powodować obniżenie jego trafności predykcyjnej. W związku ze skoncentrowaniem się na odpowiedziach na poziomie zdań w pojedynczych scenariuszach *Klucz Bremera* może nie w pełni ujawniać umiejętności komunikacyjne nauczycieli na poziomie dyskursu, w tym ich zdolności do używania języka wrażliwego na potrzeby dzieci w bieżących interakcjach.

Podsumowując, chociaż *Klucz Bremera* ocenia pewne aspekty, nie jest w stanie kompleksowo odnieść się do złożonej natury szerszych umiejętności komunikacyjnych wymaganych w kontekście edukacyjnym. Może on jednak stanowić skuteczne narzędzie diagnostyczne dostarczające wstępnych wyników i znacząco przyczynić się do wzbogacenia programów szkolenia nauczycieli. Aby zwiększyć trafność predykcyjną *Klucza Bremera*, przyszłe adaptacje lub dodatkowe oceny powinny uwzględniać metody oceniające zarówno werbalne, jak i niewerbalne aspekty komunikacji. Ponadto przyszłe badania powinny skupić się na ocenie możliwości wdrożenia i skuteczności *Klucza Bremera* w szerszych środowiskach edukacyjnych, aby jeszcze bardziej zwiększyć jego użyteczność i oddziaływanie.

Bibliografia

- Al Hashmi, W. A., Klassen, R. M. (2020). Developing a situational judgement test for admission into initial teacher education in Oman: An exploratory study. *International Journal of School & Educational Psychology*, 8(Supplement 1), 187–198. <https://doi.org/10.1080/21683603.2019.1630042>
- Berk, R. A. (1984). Conducting the Item Analysis. W: R. A. Berk (red.), *A Guide to criterion-referenced test construction* (s. 97–143). The John Hopkins University Press.
- Boland, R. J., Lester, N. A., Williams, E. (2010). Writing multiple-choice questions. *Academic Psychiatry*, 34, 310–316. <https://doi.org/10.1176/appi.ap.34.4.310>
- Bolinger, D. (1980). *Language – The loaded weapon*. Longman.
- Burns, R. B. (1982). *Self-concept development and education*. Holt.
- Carver, C. S. (1989). How should multifaceted personality constructs be tested? Issues illustrated by self-monitoring, attributional style, and hardiness. *Journal of Personality and Social Psychology*, 56(4), 577–585. <https://doi.org/10.1037/0022-3514.56.4.577>
- Craven, R. G., Marsh, H. W. (2008). The centrality of the self-concept construct for psychological well-being and unlocking human potential: Implications for child and educational psychologists. *Educational & Child Psychology*, 25(2), 104–118. <http://dx.doi.org/10.53841/bpsecp.2008.25.2.104>
- Delgado-Rico, E., Carrtero-Dios, H., Ruch, W. (2012). Content validity evidences in test development: An applied perspective. *International Journal of Clinical and Health Psychology*, 12(3), 449–460. <https://doaj.org/article/e3c64d10e477402ca1f9add91cb-467d5>

- Durksen, T. L., Klassen, R. M. (2018). The development of a situational judgment test of personal attributes for quality teaching in rural and remote Australia. *Australian Educational Researcher*, 45(2), 255–276. <https://doi.org/10.1007/s13384-017-0248-5>
- Ebel, R. L., Frisbie, D. A. (1991). *Essentials of educational measurement* (wyd. 5). Allyn & Bacon.
- Gignac, G. E. (2013). On the inappropriateness of using items to calculate total scale score reliability via coefficient alpha for multidimensional scales. *European Journal of Psychological Assessment*, 30(2), 130–139. <https://doi.org/10.1027/1015-5759/a000-181>
- Haladyna, T. M., Downing, S. M. (5–9 April, 1988). *Functional distractors: Implications for test-item writing and test design* [paper presentation]. Annual Meeting of the American Educational Research Association, New Orleans, USA. <http://files.eric.ed.gov/fulltext/ED293851.pdf>
- Haladyna, T. M., Downing, S. M., Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309–334. https://doi.org/10.1207/S15324818AME1503_5
- Haladyna, T. M., Downing, S. M., Steven, M. (31 March–4 April, 1985). *A quantitative review of research on multiple-choice item writing* [paper presentation]. Annual Meeting of the American Educational Research Association, Chicago, USA. <https://eric.ed.gov/?id=ED255580>
- Hayes, A. F., Coutts, J. J. (2020). Use Omega Rather than Cronbach’s Alpha for Estimating Reliability. But.... *Communication Methods and Measures*, 14(1), 1–24. <https://doi.org/10.1080/19312458.2020.1718629>
- Holland, J., Stevens, N. (2021). *Guidelines for the development of multiple-choice items & assessments*. RCSI University of Medicine and Health Sciences.
- Johnston, P. (2004). *Choice words*. Stenhouse.
- Johnson, R. E., Rosen, C. C., Chang, Ch.-H., Djurdjevic, E., Taing, M. U. (2012). Recommendations for improving the construct clarity of higher-order multidimensional constructs. *Human Resource Management Review*, 22(2), 62–72. <https://doi.org/10.1016/j.hrmr.2011.11.006>
- Kline, P. (2000). *The handbook of psychological testing* (wyd. 2). Routledge.
- Law, K. S., Wong, C., Mobley, W. M. (1998). Toward a taxonomy of multidimensional constructs. *The Academy of Management Review*, 23(4), 741–755. <https://doi.org/10.2307/259060>
- Lievens, F., Peeters, H., Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review*, 37(4), 426–441. <https://doi.org/10.1108/00483480-810877598>
- Mayer, R. E., Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, 38(1), 43–52. https://doi.org/10.1207/S15326985EP-3801_6
- McDaniel, M. A., Hartman, N. S., Whetzel, D. L., Grubb, W. L. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology*, 60(1), 63–91. <https://doi.org/10.1111/j.1744-6570.2007.00065.x>

- McDaniel, M. A., Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment*, 9(1–2), 103–113. <https://doi.org/10.1111/1468-2389.00167>
- Patterson, F., Zibarras, L., Ashworth, V. (2015). Situational judgement tests in medical education and training: Research, theory and practice: AMEE Guide No. 100. *Medical Teacher*, 38(1), 3–17. <https://doi.org/10.3109/0142159x.2015.1072619>
- Pollard, S., Cooper-Thomas, H. D. (2015). Best practice recommendations for Situational judgment tests. *Australasian Journal of Organisational Psychology*, 8, e7. <https://doi.org/10.1017/orp.2015.6>
- Raymond, M. R., Stevens, C., Bucak, S. D. (2019). The optimal number of options for multiple-choice questions on high-stakes tests: Application of a revised index for detecting nonfunctional distractors. *Advances in Health Sciences Education*, 24(1), 141–150. <https://doi.org/10.1007/s10459-018-9855-9>
- Reiser, S., Schacht, L., Thomm, E., Figalist, C., Janssen, L., Schick, K., Dörfler, E., Berberat, P. O., Gartmeier, M., Bauer, J. (2022). A video-based situational judgement test of medical students' communication competence in patient encounters: Development and first evaluation. *Patient Education and Counseling*, 105(5), 1283–1289. <https://doi.org/10.1016/j.pec.2021.08.020>
- Rodriguez, M. C. (2005). Three options are optimal for Multiple-Choice items: a Meta-Analysis of 80 years of research. *Educational Measurement*, 24(2), 3–13. <https://doi.org/10.1111/j.1745-3992.2005.00006.x>
- Rotschild, T. (2023). Why and how to foster learning-disabled children's emotional intelligence? *Insights into Learning Disabilities*, 20(2), 153–175. <https://files.eric.ed.gov/fulltext/EJ1401946.pdf>
- Rotschild, T. (2024). The impact of communication: A practical guide for teachers in fostering positive self-concept in children with learning disabilities. *Journal of Research in Special Educational Needs*, 00, 1–13. <https://doi.org/10.1111/1471-3802.12709>
- Smith, K. J., Flaxman, C., Farland, M. Z., Thomas, A., Buring, S. M., Whalen, K., Patterson, F. (2020). Development and validation of a situational judgement test to assess professionalism. *American Journal of Pharmaceutical Education*, 84(7), artykuł 7771, 985–992. <https://doi.org/10.5688/ajpe7771>
- Smith, K. J., Neely, S., Dennis, V. C., Miller, M. M., Medina, M. S. (2022). Use of situational judgment tests to teach empathy, assertiveness, communication, and ethics. *American Journal of Pharmaceutical Education*, 86(6), artykuł 8761. <https://doi.org/10.5688/ajpe8761>
- Watkins, M. W. (2017). The reliability of multidimensional neuropsychological measures: from alpha to omega. *The Clinical Neuropsychologist*, 31(6–7), 1113–1126. <https://doi.org/10.1080/13854046.2017.1317364>
- Whetzel, D. L., Sullivan, T. S., McCloy, R. A. (2020). Situational judgment tests: An overview of development practices and psychometric characteristics. *Personnel Assessment and Decisions*, 6(1), artykuł 1, 1–16. <https://doi.org/10.25035/pad.2020.01.001>
- Ybrandt, H. (2007). The relation between self-concept and social functioning in adolescence. *Journal of Adolescence*, 31(1), 1–16. <https://doi.org/10.1016/j.adolescence.2007.03.004>

Zhang, N., He, G., Shi, D., Zhao, Z., Li, J. (2022). Does a gender-neutral name associate with the research impact of a scientist? *Journal of Informetrics*, 16(1), artykuł 101251. <https://doi.org/10.1016/j.joi.2022.101251>