

dr hab. Artur Pokropek, prof. IFiS PAN

Instytut Badań Edukacyjnych – Państwowy Instytut Badawczy

Polskie Towarzystwo Diagnostyki Edukacyjnej

Kiedy badany przestaje się starać Rozpoznawanie odpowiedzi przy niskim zaangażowaniu i ich wpływ na trafność wyników

Streszczenie

Rozdział dotyczy rozpoznawania odpowiedzi udzielanych bez należytego wysiłku w testach poznawczych i w kwestionariuszach samoopisowych. Punktem wyjścia jest założenie, na którym opiera się większość pomiarów edukacyjnych i psychologicznych, a mianowicie, że osoba badana odpowiada najlepiej, jak potrafi. W sytuacjach o niskiej stawce założenie to bywa zawadne, a odpowiedzi udzielone bez zaangażowania obniżają trafność wniosków formułowanych na podstawie wyników. W pierwszej części uporządkowano terminologię z dwóch niezależnych tradycji badawczych, psychometrii testów poznawczych oraz metodologii sondaży. Omówiono pojęcia motywacji testowej, wysiłku, szybkiego zgadywania, nieuważnego odpowiadania, satysfakcjonowania, wytrzymałości testowej oraz danych procesowych. W części drugiej przedstawiono trzy grupy metod detekcji. Pierwszą stanowią metody oparte wyłącznie na wektorze odpowiedzi, od klasycznych wskaźników nieuważnego odpowiadania po ich nową iterację wykorzystującą uczenie maszynowe. Drugą jest proste filtrowanie odpowiedzi za pomocą progów czasu reakcji. Trzecią jest modelowanie prawdopodobieństwa braku zaangażowania na podstawie czasu reakcji oraz kolejnych zmiennych paradanych, prowadzone zarówno w ramach modeli IRT, jak i metodami uczenia maszynowego. Działanie tych metod zilustrowano czterema badaniami empirycznymi, od analizy wzorców odpowiedzi w kwestionariuszach po klasyfikację ruchów kursora. Piąte omawiane badanie, wykorzystujące śledzenie ruchu gałek ocznych, ma odmienny charakter, pokazuje bowiem, jakiej części procesu odpowiadania obecne rejestry nie obejmują, i wyznacza kierunki dalszych prac. Rozdział zamyka omówienie otwartych problemów dziedziny, przede wszystkim braku zewnętrznego wzorca odniesienia, oraz programu badań obejmującego walidację zewnętrzną, standaryzację definicji danych procesowych między platformami i udostępnienie praktykom prostych narzędzi detekcji.

Słowa kluczowe: zaangażowanie testowe, nieuważne odpowiadanie, szybkie zgadywanie, dane procesowe

1. Wprowadzenie

Testy i kwestionariusze towarzyszą nam na każdym etapie życia. Egzaminy państwowe, badania psychologiczne, sondaże opinii, oceny kompetencji pracowników, międzynarodowe porównania systemów edukacyjnych: wszystkie te sytuacje opierają się na silnym założeniu. Zakłada się, że osoba badana odpowiada najlepiej, jak potrafi, a jej odpowiedzi odzwierciedlają faktyczne umiejętności,

postawy lub opinie.

Założenie to bywa zawodne. Osoba badana nie zawsze ma powód, aby wkładać wysiłek w odpowiadanie. Gdy wynik nie ma dla niej konsekwencji, test nie decyduje o ocenie ani zatrudnieniu, a kwestionariusz jest jednym z wielu w mieścu, motywacja słabnie. Skutki ponoszą badacze i użytkownicy wyników. Rankingi krajów, mapy postaw społecznych, oceny skuteczności programów edukacyjnych opierają się na odpowiedziach, z których część pochodzi od respondentów zaangażowanych, a część od tych, którzy odpowiadali byle jak.

Ten rozdział dotyczy sposobów rozpoznawania odpowiedzi udzielonych bez należytego wysiłku. Interesują nas zarówno testy poznawcze, w których istnieje prawidłowa odpowiedź, jak i kwestionariusze samoopisowe, w których osoba badana ocenia własne postawy. W obu przypadkach niskie zaangażowanie zostawia ślady w danych. Rozdział pokazuje, jak te ślady rozpoznać, jak je interpretować i co z nich wynika dla trafności wyników. Sekcja 2 porządkuje pojęcia, sekcja 3 opisuje metody detekcji, sekcja 4 omawia pięć badań empirycznych, a sekcja 5 podsumowuje wnioski.

2. Przegląd pojęć

Terminologia dotycząca rozpoznawania odpowiedzi udzielonych bez wysiłku wyrosła z dwóch tradycji badawczych. Pierwsza to psychometria testów poznawczych, druga to metodologia sondaży i kwestionariuszy. Obie rozwijały się w dużej mierze niezależnie i czasem używają tych samych słów w różnych znaczeniach. Zaczniemy więc od porządku terminologicznego. Literatura specjalistyczna posługuje się głównie terminologią angielską, dlatego przy każdym pojęciu podaję angielski odpowiednik w nawiasie.

Motywacja testowa (*test-taking motivation*) oznacza gotowość respondenta do włożenia wysiłku w rozwiązanie zadania. W klasycznym ujęciu teorii oczekiwań i wartości¹ motywacja zależy od tego, w jakim stopniu respondent oczekuje sukcesu oraz jaką wartość przypisuje zadaniu. Test o niskiej stawce z definicji nie dostarcza żadnej użytecznej wartości poza wewnętrzną. Motywacja musi wtedy opierać się na ciekawości, poczuciu obowiązku wobec ankietera albo dyspozycji do sumiennego wykonywania powierzonych zadań². Motywacja testowa jest podstawą wysiłku (*test-taking effort*), ten zaś jest jej obserwowalną alokacją. Wise i Kong³ wprowadzili miarę wysiłku wyprowadzoną z czasu odpowiedzi (*response time effort*), obliczaną z rozkładu czasów odpowiedzi na poszczególne zadania.

Szybkie zgadywanie (*rapid guessing*) polega na odpowiadaniu tak szybko, że w tym czasie nie jest możliwe rozwiązanie zadania. To behawioralny sygnał braku wysiłku na poziomie pojedynczej odpowiedzi. Rozkład czasów odpowiedzi

¹ A. Wigfield, J.S. Eccles, *Expectancy-value theory of achievement motivation*, „Contemporary Educational Psychology” 2000, nr 25(1), s. 68–81.

² S.L. Wise, C.E. DeMars, *Low examinee effort in low-stakes assessment: Problems and potential solutions*, „Educational Assessment” 2005, nr 10(1), s. 1–17.

³ S.L. Wise, X. Kong, *Response time effort: A new measure of examinee motivation in computer-based tests*, „Applied Measurement in Education” 2005, nr 18(2), s. 163–183.

w testach komputerowych często wykazuje bimodalność. Widać dwa tryby: szybki (zgadywanie) i wolny (rozwiązywanie zadań)⁴. Na tej obserwacji opiera się większość metod detekcji wykorzystujących czas reakcji. Obok szybkiego zgadywania funkcjonuje zgadywanie po namyśle (*solution-based guessing*): respondent poświęca zadaniu dużo czasu, ale ostatecznie i tak zgaduje. Zjawiska tego nie da się wychwycić prostym progiem czasowym.

Odpowiednikiem szybkiego zgadywania po stronie kwestionariuszy jest nieuważne odpowiadanie (*careless responding*, *insufficient effort responding*, w skrócie C/IER). Chodzi o odpowiadanie bez wystarczającej uwagi⁵. Zjawisko ma wiele form. Najprostszą jest wypełnianie kolumnowe (*straight-lining*), czyli zaznaczanie tej samej odpowiedzi w kolejnych pytaniach. Odpowiedź środkowa (*midpoint responding*) polega na konsekwentnym wybieraniu środka skali. Naprzemienne odpowiedzi skrajne (*extreme alternating*) to przeplatanie odpowiedzi na krańcach skali. Wzór przekątny (*diagonal bouncing*) polega na zmienianiu odpowiedzi o jeden punkt skali w każdym kolejnym zadaniu. Odpowiedź losowa (*random responding*) to odpowiedzi bez związku z treścią pytań. W literaturze polskiej używa się czasem terminu „niedostateczne zaangażowanie”, jednak w publikacjach specjalistycznych dominuje anglojęzyczny skrót C/IER.

Szerszą teoretyczną ramę dla obu problemów daje koncepcja satysfakcjonowania (*satisficing*)⁶. Respondent, dla którego zadanie przekracza iloczyn jego zdolności i motywacji, stosuje skróty poznawcze. Słabe satysfakcjonowanie polega na wybraniu pierwszej akceptowalnej odpowiedzi bez pełnego przemyślenia. Silne satysfakcjonowanie polega na pominięciu etapu zrozumienia pytania. Nieuważne odpowiadanie można traktować jako behawioralną konsekwencję silnego satysfakcjonowania.

Osobna rodzina pojęć dotyczy wytrwałości w sesji testowej. Wytrzymałość testowa (*test endurance*, *persistence*) oznacza zdolność utrzymywania wysiłku w miarę trwania testu. Empirycznym śladem ograniczonej wytrzymałości jest efekt pozycji (*position effect*), czyli spadek wykonania zadania zależny od jego pozycji w teście. Borgonovi i Biecek⁷ pokazali, że kraje różnią się istotnie pod względem tempa, w którym uczniowie tracą wysiłek w badaniach PISA. Balart i Oosterveen⁸ wykazali, że kobiety utrzymują wysiłek dłużej niż mężczyźni, a długość testu wpływa na estymowane różnice płciowe w matematyce. Wytrzymałość jest zatem po części cechą osoby, po części stanem zależnym od

⁴ D.L. Schnipke, D.J. Scrams, *Modeling item response times with a two-state mixture model: A new method of measuring speededness*, „Journal of Educational Measurement” 1997, nr 34(3), s. 213–232.

⁵ A.W. Meade, S.B. Craig, *Identifying careless responses in survey data*, „Psychological Methods” 2012, nr 17(3), s. 437–455; P.G. Curran, *Methods for the detection of carelessly invalid responses in survey data*, „Journal of Experimental Social Psychology” 2016, nr 66, s. 4–19; M.K. Ward, A.W. Meade, *Dealing with careless responding in survey data: Prevention, identification, and recommended best practices*, „Annual Review of Psychology” 2023, nr 74, s. 577–596.

⁶ J.A. Krosnick, *Response strategies for coping with the cognitive demands of attitude measures in surveys*, „Applied Cognitive Psychology” 1991, nr 5(3), s. 213–236.

⁷ F. Borgonovi, P. Biecek, *An international comparison of students' ability to endure fatigue and maintain motivation during a low-stakes test*, „Learning and Individual Differences” 2016, nr 49, s. 128–137.

⁸ P. Balart, M. Oosterveen, *Females show more sustained performance during test-taking than males*, „Nature Communications” 2019, nr 10, s. 3798.

architektury samego testu.

Ostatnie pojęcie porządkujące to dane procesowe (*process data* albo *paradata*). Termin ten obejmuje dane zbierane obok właściwych odpowiedzi. Klasycznym przykładem jest czas reakcji na zadanie. Nowsze dane procesowe obejmują liczbę akcji na ekranie (kliknięcia, przewinięcia, użycie kalkulatora), sekwencje ruchów myszy, ruchy gałek ocznych rejestrowane eye-trackerem oraz miary fizjologiczne, takie jak średnica źrenicy.

Powyższe pojęcia zebrane są w tabeli 1 jako podsumowanie słownika, do którego czytelnik może wracać w trakcie lektury.

Tabela 1. Kluczowe pojęcia i ich anglojęzyczne odpowiedniki

Termin polski	Termin angielski	Znaczenie
Motywacja testowa	<i>test-taking motivation</i>	Gotowość respondenta do włożenia wysiłku w rozwiązanie zadania
Wysiłek	<i>test-taking effort</i>	Realizacja motywacji, faktyczny wysiłek włożony w odpowiadanie
Wysiłek wyprowadzony z czasu odpowiedzi	<i>response time effort</i>	Miara wysiłku obliczana z rozkładu czasów odpowiedzi na poszczególne zadania
Szybkie zgadywanie	<i>rapid guessing</i>	Odpowiadanie tak szybko, że rozwiązanie zadania nie było w tym czasie możliwe
Zgadywanie po namyśle	<i>solution-based guessing</i>	Długi czas na zadaniu zakończony odpowiedzią, która nie wynika z rozwiązania
Nieuważne odpowiadanie	<i>careless responding, insufficient effort responding (C/IER)</i>	Odpowiadanie w kwestionariuszu bez wystarczającej uwagi
Wypełnianie kolumnowe	<i>straight-lining</i>	Zaznaczanie tej samej odpowiedzi w kolejnych pytaniach
Odpowiedź środkowa	<i>midpoint responding</i>	Konsekwentne wybieranie środka skali
Odpowiedź losowa	<i>random responding</i>	Odpowiedzi bez związku z treścią pytań
Wzór przekątny	<i>diagonal bouncing</i>	Odpowiedzi zmieniające się o jeden punkt skali w każdym zadaniu
Naprzemienne odpowiedzi skrajne	<i>extreme alternating</i>	Przeplatanie odpowiedzi na krańcach skali
Satysfakcjonowanie	<i>satisficing</i>	Stosowanie skrótów poznawczych, gdy zadanie przekracza możliwości lub motywację respondenta
Wytrzymałość testowa	<i>test endurance, persistence</i>	Zdolność utrzymywania wysiłku w miarę trwania testu
Efekt pozycji	<i>position effect</i>	Spadek wykonania zadania zależny od jego pozycji w teście
Dane procesowe	<i>process data, paradata</i>	Dane zbierane obok właściwych odpowiedzi (czas reakcji, kliknięcia, ruchy myszy, ruchy gałek ocznych)

Dwa rozróżnienia porządkują dalszą część rozdziału. Pierwsze rozdziela detekcję od modelowania. Detekcja to decyzja, czy dana odpowiedź została

udzielona z zaangażowaniem. Modelowanie to sposób, w jaki uwzględnia się tę informację w estymacji parametrów pomiaru. Można wykryć odpowiedź niezaangażowaną i ją odrzucić, można ją wykryć i zważyć, można też modelować proces generujący odpowiedź razem z procesem generującym zaangażowanie. Drugie rozróżnienie oddziela testy poznawcze od kwestionariuszy samoopisowych. W testach istnieje prawidłowa odpowiedź, więc niezaangażowanie zostawia ślad w postaci błędu. W kwestionariuszach takiego śladu nie ma, a detekcja musi opierać się na wzorcach odpowiedzi oraz na czasach ich udzielania.

3. Przegląd metod detekcji

Metody rozpoznawania odpowiedzi udzielonych bez wysiłku można uporządkować w trzy grupy według tego, z jakich danych korzystają. Pierwsza grupa opiera się wyłącznie na samych odpowiedziach, czyli na wektorze odpowiedzi respondenta, i poszukuje w nim charakterystycznych wzorców. Druga sięga po czas reakcji i stosuje go w najprostszy możliwy sposób, jako kryterium odrzucania odpowiedzi zbyt szybkich. Trzecia, najbardziej zaawansowana, modeluje prawdopodobieństwo braku zaangażowania, najpierw na podstawie samego czasu reakcji, a następnie z użyciem kolejnych zmiennych paradanych, zarówno w ramach modeli IRT, jak i za pomocą uczenia maszynowego. Kolejność ta odpowiada zarazem rosnącym wymaganiom, od metod możliwych do zastosowania na każdym zbiorze danych po metody wymagające bogatych rejestrów procesu i znacznych zasobów obliczeniowych.

Pierwsza grupa wyrosła z metodologii kwestionariuszy, w których nie ma poprawnej odpowiedzi, a jedynym materiałem jest to, co respondent zaznaczył. Klasyczne wskaźniki nieuważnego odpowiadania oblicza się na poziomie osoby. Najprostszy z nich, najdłuższy ciąg (*longstring*), rejestruje, ile identycznych odpowiedzi respondent zaznaczył z rzędu. Statystyki dopasowania osoby (*person-fit*) mierzą, jak bardzo wzorzec odpowiedzi odbiega od przewidywań modelu psychometrycznego, a dystans Mahalanobisa, jak bardzo odbiega on od reszty populacji. Metody synonimów i antonimów sprawdzają spójność odpowiedzi na pary pytań o zbliżonym lub przeciwnym znaczeniu⁹. Wskaźniki te są tanie, łatwe w interpretacji i dostępne w wielu pakietach statystycznych. Mają jednak wspólne ograniczenie, ponieważ każdy z nich wykrywa jeden wzorzec, najczęściej wypełnianie kolumnowe, natomiast wzorce bardziej złożone, takie jak wzór przekątny czy odpowiedzi losowe, pozostają dla nich niewidoczne.

Nową iterację tej samej grupy stanowi podejście oparte na uczeniu maszynowym. Materiał wejściowy pozostaje ten sam, wciąż jest nim wyłącznie wektor odpowiedzi, zmienia się natomiast narzędzie analizy. Zamiast pojedynczego wskaźnika nastawionego na jeden wzorzec stosuje się klasyfikatory, takie jak głębokie sieci neuronowe (*deep neural networks*, DNN) czy maszyny wektorów podpierających (*support vector machines*, SVM), zdolne wychwycić wiele wzorców jednocześnie, również nieliniowych. Główną przeszkodą jest brak danych treningowych z etykietami, nie wiadomo bowiem, którzy respondenci

⁹ A.W. Meade, S.B. Craig, dz. cyt.; P.G. Curran, dz. cyt.

naprawdę się nie starali. Pokropek¹⁰ (zob. sekcja 4.1) proponuje rozwiązanie tego problemu, w którym klasyfikatory trenuje się na danych syntetycznych, generowanych metodą Monte Carlo z zaplanowanymi wzorcami niezaangażowania, a wyuczoną regułę stosuje następnie w danych empirycznych. Dawna idea poszukiwania wzorców w samych odpowiedziach zyskuje w ten sposób znacznie skuteczniejsze narzędzie.

Druga grupa obejmuje proste filtrowanie odpowiedzi za pomocą progów czasu reakcji. Rozumowanie jest tu następujące. Odpowiedź udzielona szybciej, niż pozwala na to przeczytanie i rozwiązanie zadania, nie mogła powstać w wyniku rozwiązywania, należy ją zatem odrzucić, a cała trudność sprowadza się do pytania, gdzie postawić granicę. Wise i Kong¹¹ zaproponowali próg stały, na przykład trzy sekundy, jednakowy dla wszystkich zadań. Progi normatywne uwzględniają, że zadania różnią się typowym czasem rozwiązywania, i wyznaczają granicę jako ułamek tego czasu, na przykład dziesięć procent mediany. Progi zakotwiczone w dokładności (*accuracy-anchored thresholds*) poszukują punktu, w którym prawdopodobieństwo poprawnej odpowiedzi zaczyna przewyższać poziom czystego zgadywania. Porównanie tych trzech wariantów na danych PIAAC wypadło na korzyść progów zakotwiczonych w dokładności, które dały najstabilniejsze wyniki¹². Zarazem metaanaliza procedur progowych pokazała, że sam wybór progu istotnie zmienia szacowany odsetek niezaangażowania¹³. Filtrowanie jest tanie, szybkie i szeroko wdrożone, ma jednak dwie zasadnicze słabości. Punkt odcięcia pozostaje arbitralny, a decyzja zerojedynkowa, zaangażowany albo nie, pomija całą niepewność tego rozstrzygnięcia.

Trzecia grupa zastępuje proste filtrowanie modelowaniem statystycznym. Zamiast odrzucać odpowiedzi poniżej progu, szacuje się prawdopodobieństwo, że dana odpowiedź powstała bez zaangażowania, i uwzględnia tę niepewność w estymacji. Pierwszym predyktorem tego prawdopodobieństwa był czas reakcji, wykorzystywany w modelach mieszanek, które zakładają, że dane pochodzą z co najmniej dwóch procesów, rozwiązywania i zgadywania. Kolejne modele różnią się poziomem, na którym umieszczają to rozróżnienie. Model HYBRID¹⁴ dzieli osoby, przyjmując, że respondent albo rozwiązuje zadania zgodnie z modelem IRT, albo zgaduje losowo. Model IRT z korektą na wysiłek (*effort-moderated IRT*)¹⁵ schodzi na poziom pojedynczej odpowiedzi, ale nadal opiera się na progu czasowym, ponieważ odpowiedzi poniżej progu traktuje jako niepełne obserwacje i usuwa z estymacji. Pełne przejście

¹⁰ A. Pokropek, *Detecting careless response patterns in Likert scales using supervised machine learning and Monte Carlo simulations*, „Current Psychology” 2026, nr 45, s. 518.

¹¹ S.L. Wise, X. Kong, dz. cyt.

¹² F. Goldhammer, T. Martens, G. Christoph, O. Lüdtke, *Test-taking engagement in PIAAC*, OECD Education Working Papers No. 133, OECD Publishing, Paryż 2016.

¹³ J.A. Rios, J. Deng, *Does the choice of response time threshold procedure substantially affect inferences concerning the identification and exclusion of rapid guessing responses? A meta-analysis*, „Large-scale Assessments in Education” 2021, nr 9(1), s. 1–25.

¹⁴ K. Yamamoto, *HYBRID model of IRT and latent class models*, ETS Research Report No. RR-89-41, Educational Testing Service, Princeton 1989.

¹⁵ S.L. Wise, C.E. DeMars, *An application of item response time: The effort-moderated IRT model*, „Journal of Educational Measurement” 2006, nr 43(1), s. 19–38.

od progu do prawdopodobieństwa dokonuje się w modelu przynależności stopniowalnej z czasem reakcji (*Grade of Membership Response Time Model*, GoM-RT)¹⁶, omówionym szerzej w sekcji 4.2. W modelu tym każda odpowiedź może należeć częściowo do klasy rozwiązywania i częściowo do klasy zgadywania, a stopień przynależności przewidywany jest na podstawie czasu reakcji. Hierarchiczny model odpowiedzi latentnych (*hierarchical latent response model*)¹⁷ łączy w jednej strukturze zgadywanie i pomijanie zadań, czyli dwa zachowania, które w praktyce często idą w parze.

Naturalnym rozwinięciem tej logiki jest włączanie kolejnych zmiennych paradanych. Czas reakcji mówi, jak długo respondent był przy zadaniu, ale nie mówi, co w tym czasie robił, a pokazują to dopiero liczba akcji na ekranie, użycie kalkulatora lub notatnika, sekwencje ruchów myszy, wreszcie ruchy gałek ocznych. Rozbudowa może przebiegać drogą modelową, jak u Pokropka i Khorramdel¹⁸, którzy włączają liczbę akcji, typ zadania i cechy respondenta jako predyktory zgadywania do jednego modelu IRT (zob. sekcja 4.3). Może też przebiegać drogą uczenia maszynowego, jak u Pokropka, Żółtaka i Muszyńskiego¹⁹, którzy analizują sekwencje ruchów kursora, traktowane jako przybliżenie uwagi wzrokowej (zob. sekcja 4.4). Najbogatszym źródłem obserwacji procesu pozostaje śledzenie ruchu gałek ocznych, po które sięga Maddox ze współpracownikami²⁰ (zob. sekcja 4.5). W obu wariantach zasada jest ta sama. Im więcej zmiennych opisujących proces odpowiadania, tym precyzyjniej można przewidzieć prawdopodobieństwo braku zaangażowania, tym wyższe są jednak koszty zbierania i przetwarzania danych.

Wybór spośród tych grup zależy przede wszystkim od typu instrumentu. W kwestionariuszach samoopisowych nie ma poprawnej odpowiedzi, dlatego naturalnym materiałem jest wektor odpowiedzi i metody pierwszej grupy, w razie potrzeby w wersji wykorzystującej uczenie maszynowe. W testach poznawczych osią detekcji jest czas reakcji, przy czym filtrowanie progowe wystarcza jako szybka diagnoza, natomiast tam, gdzie istotna jest trafność estymacji, lepiej sprawdza się modelowanie prawdopodobieństwa. Drugim kryterium są dostępne dane, im bogatsze bowiem paradane rejestruje platforma badawcza, tym pełniej można wykorzystać metody trzeciej grupy.

¹⁶ A. Pokropek, *Grade of membership response time model for detecting guessing behaviors*, „Journal of Educational and Behavioral Statistics” 2016, nr 41(3), s. 300–325.

¹⁷ E. Ulitzsch, M. von Davier, S. Pohl, *A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response*, „British Journal of Mathematical and Statistical Psychology” 2020, nr 73(S1), s. 83–112.

¹⁸ A. Pokropek, L. Khorramdel, *Analyzing test-taking behaviors through process data using the IRT Explanatory Model for Guessing* [w:] L. Khorramdel i in. (red.), *Innovative digital-based international large-scale assessments*, Springer, Cham 2025, s. 559–584.

¹⁹ A. Pokropek, T. Żółtak, M. Muszyński, *Approximate Areas of Interest for enhanced understanding of student motivation and task interaction in IEA assessments* [raport techniczny], IEA, Amsterdam 2025.

²⁰ B. Maddox, A.P. Bayliss, P. Fleming, P.E. Engelhardt, S.G. Edwards, F. Borgonovi, *Observing response processes with eye tracking in international large-scale assessments: Evidence from the OECD PIAAC assessment*, „European Journal of Psychology of Education” 2018, nr 33(3), s. 543–558.

4. Pięć przykładów empirycznych

Sekcję tę tworzy pięć badań empirycznych, które pełnią jednak dwie różne role. Cztery pierwsze pokazują rozwój prac nad detekcją i praktyczne zastosowanie opisanych wcześniej grup metod, od wzorców w samych odpowiedziach (4.1) przez modelowanie prawdopodobieństwa zgadywania na podstawie czasu reakcji (4.2) i jego rozszerzenie o kolejne dane procesowe (4.3) po klasyfikację ruchów kursora metodami uczenia maszynowego (4.4). Badanie piąte (4.5) ma odmienny charakter. Nie proponuje kolejnej metody detekcji, lecz pokazuje, jakiej części procesu odpowiadania nie obejmują dzisiejsze rejestry, i wskazuje, co należy badać dalej.

4.1. Wykrywanie wzorców w kwestionariuszach: Pokropek (2026)

Problem, od którego wychodzi pierwsze badanie, pojawił się już przy omawianiu pierwszej grupy metod. Klasyczne wskaźniki nieuważnego odpowiadania wykrywają w zasadzie jeden wzorzec, wypełnianie kolumnowe, a respondent, który zamiast zaznaczać wciąż tę samą odpowiedź porusza się po skali, po przekątnej, wokół środka albo losowo, pozostaje niewykryty. Pokropek²¹ proponuje rozwiązanie w trzech krokach. Najpierw generuje się syntetyczne odpowiedzi na skalę Likerta z pięcioma zaplanowanymi wzorcami niezaangażowania, którymi są wypełnianie kolumnowe, wzór przekątny, odpowiedź środkowa, naprzemienne odpowiedzi skrajne i odpowiedź losowa. Następnie na tych danych trenuje się klasyfikatory z rodziny głębokich sieci neuronowych oraz maszyn wektorów podpierających. Na końcu wyuczony klasyfikator stosuje się do danych empirycznych, w których część respondentów proszono o wypełnienie kwestionariusza „byle jak”, a część odpowiadała uważnie.

W symulacjach klasyfikatory rozpoznawały zaplanowane wzorce ze skutecznością przekraczającą 85 procent. Istotniejsze okazały się jednak wyniki empiryczne. Grupę niezaangażowaną od uważnej istotnie różnicowały trzy wzorce, mianowicie wzór przekątny (12 wobec 1 procenta respondentów), odpowiedź środkowa (20 wobec 7 procent) i odpowiedź losowa (48 wobec 30 procent). Dwa pozostałe, czyli wypełnianie kolumnowe i naprzemienne odpowiedzi skrajne, nie różnicowały grup wcale.

Płyną stąd dwa wnioski. Po pierwsze, dane syntetyczne pozwalają ominąć główną barierę uczenia maszynowego w tej dziedzinie, czyli brak dużych zbiorów z etykietami, ponieważ klasyfikator wystarczy wytrenować raz, a potem stosować w kolejnych badaniach. Po drugie, i jest to wniosek mniej wygodny, najpopularniejszy w praktyce wskaźnik, analiza najdłuższego ciągu, mierzy akurat ten wzorzec, który najslabiej odróżnia respondentów niezaangażowanych od uważnych. Poprzestanie na nim oznacza przeoczenie znacznej części nieuważnego odpowiadania.

4.2. Model mieszanek z czasem reakcji: Pokropek (2016)

Drugie badanie dotyczy testów poznawczych i trzeciej grupy metod, czyli modelowania prawdopodobieństwa zgadywania. Przed jego publikacją praktyka знаła dwa sposoby postępowania ze zgadywaniem. Można było

²¹ A. Pokropek, *Detecting careless response patterns...*, dz. cyt.

filtrować odpowiedzi po czasie reakcji, co wymagało arbitralnej decyzji o progu, albo stosować modele mieszanek na poziomie osoby, co z kolei zakładało, że respondent albo rozwiązuje wszystkie zadania, albo wszystkie zgaduje. Tymczasem rzeczywiste zachowanie jest mieszane, ten sam respondent jedno zadanie rozwiązuje, a przy innych zgaduje. Brakowało modelu, który dopuszczałby mieszaną na poziomie pojedynczej odpowiedzi.

Taką konstrukcją jest model przynależności stopniowalnej z czasem reakcji (GoM-RT). Definiuje on dwie klasy latentne dla każdej odpowiedzi z osobna. Pierwsza to rozwiązywanie, opisywane standardowym modelem IRT, druga to zgadywanie, w którym prawdopodobieństwo poprawnej odpowiedzi jest losowe i niezależne od zdolności respondenta. O tym, w jakim stopniu dana odpowiedź należy do klasy zgadywania, nie decyduje arbitralny próg, lecz predykcja oparta na czasie reakcji.

Model sprawdzono dwutorowo. W symulacjach Monte Carlo, w których manipulowano poziomem zgadywania, jego korelacją ze zdolnością oraz separowalnością rozkładów czasów reakcji, GoM-RT dawał najniższy błąd estymacji parametrów zadań we wszystkich warunkach. Przewagę zachowywał zwłaszcza tam, gdzie szybkie odpowiedzi rozwiązywane trudno odróżnić od zgadywanych, a więc dokładnie tam, gdzie model Rascha ignorujący zgadywanie oraz procedury filtrowania traciły precyzję. W realnych danych PIAAC estymacja modelem GoM-RT korygowała poziomy sprawności nieznacznie w górę, co jest kierunkiem oczekiwanym, ponieważ ignorowanie zgadywania systematycznie zaniża estymowaną kompetencję.

GoM-RT był pierwszym modelem IRT, w którym zgadywanie jest specyficzne dla pojedynczej odpowiedzi i warunkowane czasem reakcji, a nie progiem. Stał się punktem wyjścia do dalszej linii prac²² oraz rozszerzeń proponowanych przez innych autorów²³. Wniosek praktyczny brzmi, że modelowanie zgadywania na poziomie odpowiedzi jest wykonalne i redukuje obciążenie estymatorów bez odrzucania jakichkolwiek obserwacji. Ceną pozostaje koszt obliczeniowy.

4.3. Wzbogacenie czasu reakcji o dane procesowe: Pokropek i Khorramdel (2025)

Trzecie badanie zaczyna się tam, gdzie kończy się poprzednie. GoM-RT opiera się w zasadzie na jednym wskaźniku procesu, czyli czasie reakcji, oraz na założeniu, że krótszy czas oznacza wyższe ryzyko zgadywania. Oba elementy okazują się zbyt ubogie. Zadania różnią się typowym czasem rozwiązywania, a obok szybkiego zgadywania istnieje zgadywanie po namyśle, w którym respondent spędza nad zadaniem dużo czasu i mimo to zgaduje. Potrzebny był model elastyczniejszy, uwzględniający różne typy zgadywania oraz cechy zadań, respondentów i ich interakcje.

²² A. Pokropek, L. Khorramdel, dz. cyt.

²³ G. Nagy, E. Ulitzsch, *A multilevel mixture IRT framework for modeling response times as predictors or indicators of response engagement in IRT models*, „Educational and Psychological Measurement” 2022, nr 82(5), s. 845–879; E. Ulitzsch, M. von Davier, S. Pohl, dz. cyt.

Odpowiedzią jest wyjaśniający model IRT dla zgadywania (*IRT Explanatory Model for Guessing*, IRT-EMG), zaproponowany przez Pokropek i Khorramdel²⁴ jako bezpośrednie rozszerzenie GoM-RT. Nowość polega na rozbudowanej strukturze predyktorów zgadywania, uporządkowanych na trzech poziomach. Obejmują one cechy zadania, na przykład format odpowiedzi, cechy pojedynczej odpowiedzi, w tym czas reakcji i liczbę akcji wykonanych na ekranie, oraz cechy respondenta, takie jak wiek, płeć, wykształcenie i szacowana zdolność. Model zastosowano do danych PIAAC z trzech krajów o różnych poziomach sprawności, Japonii, Wielkiej Brytanii i Włoch.

Na uwagę zasługują trzy wyniki. Po pierwsze, szacowany odsetek zgadywania wahał się od kilku do kilkunastu procent i był konsekwentnie wyższy w krajach o niższych wynikach. Po drugie, relacja między czasem reakcji a prawdopodobieństwem zgadywania okazała się nie monotoniczna, lecz U-kształtna, ryzyko zgadywania jest bowiem wysokie zarówno przy bardzo krótkich, jak i przy bardzo długich czasach odpowiedzi. Stanowi to empiryczne potwierdzenie, że zgadywanie po namyśle jest zjawiskiem odrębnym od szybkiego zgadywania. Po trzecie, liczba akcji na ekranie okazała się niezależnym predyktorem zgadywania, a spośród cech respondenta najsilniej chroniła przed nim liczba lat wykształcenia, i to spójnie we wszystkich trzech krajach.

Wniosek jest dwojaki. W wymiarze metodologicznym łączenie różnych typów danych procesowych w jednym modelu daje obraz zgadywania o znacznie większej rozdzielczości niż sam czas reakcji. W wymiarze praktycznym, skoro ryzyko zgadywania zależy od formatu zadania, i to w sposób różny w różnych dziedzinach pomiaru, dobór formatów przy konstruowaniu testu jest jednocześnie decyzją o skali przyszłego zgadywania.

4.4. Ruchy kursora i przybliżone obszary zainteresowania: Pokropek, Żółtak i Muszyński (2025)

Czwarte badanie zamyka część poświęconą detekcji i mierzy się z problemem kosztów. Najdokładniejszą obserwację uwagi wzrokowej daje śledzenie ruchu gałek ocznych, wymaga ono jednak specjalnego sprzętu i kalibracji, przez co nie skaluje się do dużych badań. Ruch kursora myszy częściowo koreluje z ruchem wzroku, a rejestrować go można w każdym badaniu komputerowym bez dodatkowych urządzeń. W ankietach metoda oparta na ruchach kursora osiągała już dokładność około 95 procent. Otwarte pozostawało pytanie, czy sprawdzi się także w testach poznawczych, w których różnice motywacyjne są subtelniejsze.

Pokropek, Żółtak i Muszyński²⁵ proponują metodę przybliżonych obszarów zainteresowania (*Approximate Areas of Interest*, AAOI). Ekran testowy dzieli się na siatkę stref o różnej gęstości, a dla każdego respondenta i każdego zadania rejestruje się sekwencję pozycji kursora na tych strefach. Sekwencje klasyfikuje dwukierunkowa sieć rekurencyjna typu LSTM, rozstrzygając, czy respondent odpowiadał uważnie czy bez zaangażowania. Badanie objęło ponad tysiąc

²⁴ A. Pokropek, L. Khorramdel, dz. cyt.

²⁵ A. Pokropek, T. Żółtak, M. Muszyński, dz. cyt.

uczniów klas siódmych i ósmych polskiej szkoły podstawowej, a motywacją manipulowano w schemacie wewnątrzosobowym, w którym ten sam uczeń rozwiązywał zadania raz w warunkach uważnym, raz z instrukcją demotywującą.

Najlepszy model osiągnął 67 procent trafnych klasyfikacji, a więc wyraźnie mniej niż 95 procent uzyskiwane analogiczną metodą w ankietach. Różnica ta jest pouczająca, ponieważ testowanie poznawcze okazuje się trudniejszym obiektem detekcji. Manipulacja motywacją nie zmienia w oczywisty sposób parametrów zadań, a ślad, jaki zostawia w zachowaniu myszy, jest dyskretny. Ujawniły się też drobne, lecz spójne różnice płciowe we wskaźnikach ruchu kursora. Mimo umiarkowanej skuteczności bilans wypada na korzyść metody, AAOI daje bowiem tanie i skalowalne przybliżenie śledzenia wzroku, możliwe do zastosowania w zwykłej klasie szkolnej bez specjalnego sprzętu.

4.5. Czego nie rejestrują obecne metody. Eye-tracking w PIAAC: Maddox i współpracownicy (2018)

Piąte badanie różni się od czterech poprzednich. Nie proponuje kolejnej metody detekcji, lecz sprawdza, jakiej części procesu odpowiadania nie widzą narzędzia opisane dotąd, i tym samym wyznacza granice całego omawianego podejścia. Nawet bogate zapisy komputerowe rejestrują wyłącznie akcje, kliknięcia, przewinięcia i czasy. Nie pokazują tego, co dzieje się między akcjami, a więc które fragmenty ekranu respondent czyta i w jakiej kolejności, kiedy odrywa wzrok od monitora, kiedy wewnętrznie wycofuje się z zadania. Standardowe metody walidacji, takie jak protokół głośnego myślenia (*think aloud*) czy wywiady poznawcze, tej luki nie wypełniają. Stąd pytanie badawcze, co śledzenie ruchu gałek ocznych wnosi do walidacji i interpretacji zadań PIAAC.

Maddox, Bayliss, Fleming, Engelhardt, Edwards i Borgonovi²⁶ przeprowadzili eksploracyjne badanie laboratoryjne z udziałem czternastu dorosłych, głównie pracowników i studentów uniwersytetu, na standardowym protokole PIAAC w wersji brytyjskiej. Użyto eye-trackera w formie okularów, mniej precyzyjnego niż stacjonarny, ale rejestrującego również spojrzenia poza ekran. Dodatkowo mierzono średnicę źrenicy i przewodnictwo skóry jako pośrednie miary obciążenia poznawczego i pobudzenia. Analizowano sekwencje spojrzeń na wyznaczone obszary zainteresowania, czasy fiksacji oraz spojrzenia poza monitor.

Wyniki ujawniły cztery zjawiska niewidoczne dla klasycznych zapisów. Po pierwsze, respondenci traktują fragmenty instrukcji jako punkty odniesienia i wracają do nich wzrokiem w trakcie odpowiadania. Po drugie, znaczna część rozwiązywania zadań liczbowych odbywa się poza ekranem, na kalkulatorze, na kartce lub w wymianie spojrzeń z ankieterem, a według szacunków autorów jest to nawet do 40 procent aktywności. Po trzecie, dwie identyczne poprawne odpowiedzi mogą kryć skrajnie różne procesy dochodzenia do nich, zapis widzi bowiem jedynie dwa poprawne kliknięcia o różnym czasie, podczas gdy eye-tracker pokazuje dwie odmienne strategie rozwiązywania. Po czwarte, sekwencje spojrzeń pozwalają rozróżniać strategie czytania tekstu.

²⁶ B. Maddox i in., dz. cyt.

Dla całego rozdziału znaczenie tego badania polega nie tyle na wynikach szczegółowych, ile na diagnozie. Skoro nawet do 40 procent aktywności pozostaje poza zasięgiem rejestrów, na których opierają się metody detekcji, każda z nich pracuje na obrazie niepełnym. Badanie ma przy tym oczywiste ograniczenia, małą próbę, warunki laboratoryjne i wysoko wykształconą grupę uczestników, a sami autorzy podkreślają, że eye-tracking jest daleki od wdrożenia w głównym nurcie badań. Wskazuje jednak, co należy badać dalej. Otwiera drogę do walidacji zadań, wykrywania ich różnicowego funkcjonowania (*differential item functioning*, DIF) i obserwacji strategii rozwiązywania, a przede wszystkim wyznacza kierunek kolejnej generacji danych procesowych, w której obserwacja procesu odpowiedzi wychodzi poza ekran komputera.

5. Wnioski

Obraz wyłaniający się z tego rozdziału to obraz dziedziny w ruchu. Proste filtrowanie po progu czasu reakcji, przez dwadzieścia lat dominujące w praktyce, ustępuje modelowaniu prawdopodobieństwa zgadywania, modele oparte na samym czasie reakcji ustępują modelom z wieloma zmiennymi paradanymi, a klasyczne wskaźniki wzorców odpowiedzi zyskują nową iterację w postaci uczenia maszynowego. Kilka ustaleń można już uznać za solidne. Progi zakończone w dokładności dają stabilniejsze wyniki niż progi stałe²⁷. Modele mieszanek redukują obciążenie estymatorów bez odrzucania obserwacji²⁸. Bogatsze dane procesowe ujawniają informacje, których czas reakcji nie niesie²⁹, a uczenie maszynowe trenowane na danych syntetycznych wykrywa wzorce niedostępne klasycznym wskaźnikom³⁰. Ten dorobek nie powinien jednak przesłaniać faktu, że najważniejsze problemy dziedziny pozostają otwarte, o czym wprost przypomina omówione na końcu badanie z użyciem eye-trackera.

Problem pierwszy jest fundamentalny, nie istnieje bowiem wzorzec odniesienia (*ground truth*) dla odpowiedzi niezaangażowanej w warunkach polowych. Wszystkie omawiane metody wykrywają ślady niezaangażowania, ale żadna nie może udowodnić, że wykrywa je trafnie. Eksperymentalne manipulacje instrukcją motywacyjną, jak w badaniach Pokropka, Żółtaka i Muszyńskiego³¹ czy Pokropka³², dają porównanie warunków, lecz nie pewność, że osoba w warunkach niezaangażowanym rzeczywiście odpowiadała bezmyślnie w każdym zadaniu. Skutki tej luki widać w rozrzucie skuteczności. Metody osiągające 95 procent trafnych klasyfikacji w kwestionariuszach dają w testach poznawczych zaledwie 67 procent, ponieważ różnice motywacyjne są tam subtelniejsze, a parametry zadań pozostają względnie stabilne mimo zmian zaangażowania. Dopóki nie powstaną powtarzalne protokoły eksperymentalne pozwalające porównywać metody na wspólnym materiale, wyniki detekcji trzeba traktować jako szacunki, nie zaś rozstrzygnięcia.

²⁷ F. Goldhammer, T. Martens, G. Christoph, O. Lüdtke, dz. cyt.

²⁸ A. Pokropek, *Grade of membership response time model...*, dz. cyt.; E. Ulitzsch, M. von Davier, S. Pohl, dz. cyt.

²⁹ A. Pokropek, L. Khorramdel, dz. cyt.; A. Pokropek, T. Żółtak, M. Muszyński, dz. cyt.; B. Maddox i in., dz. cyt.

³⁰ A. Pokropek, *Detecting careless response patterns...*, dz. cyt.

³¹ A. Pokropek, T. Żółtak, M. Muszyński, dz. cyt.

³² A. Pokropek, *Detecting careless response patterns...*, dz. cyt.

Problem drugi dotyczy praktyki, która nie nadąża za metodologią. Najczęściej stosowany wskaźnik, czyli wypełnianie kolumnowe, ma słabszą moc dyskryminacyjną, niż się powszechnie sądzi, ponieważ grupy uważne i niezaangażowane różnicowały dopiero wzorce złożone, przekątny, środkowy i losowy³³. Filtrowanie odpowiedzi po arbitralnym progu jest tanie, ale traci informację i wprowadza własne obciążenia³⁴, podczas gdy modelowanie prawdopodobieństwa zachowuje wszystkie obserwacje i pozwala korygować estymaty. Jeżeli już progi, to zakotwiczone w dokładności, nie stałe. Zalecenia te są znane, a mimo to rzadko wdrażane. Powód jest prosty. Nowocześniejsze metody wymagają kompetencji i narzędzi, którymi przeciętny badacz nie dysponuje, a dopóki modele mieszanek i klasyfikatory nie trafią do standardowych pakietów statystycznych w formie równie prostej jak wskaźnik najdłuższego ciągu, praktyka będzie wybierać to, co dostępne, a nie to, co trafne.

Problem trzeci ma charakter teoretyczny, wciąż bowiem nie wiadomo, czym właściwie jest niezaangażowanie, szumem do usunięcia czy konstruktem do zmierzenia. Coraz więcej przemawia za tym drugim. Zgadywanie rośnie wraz z trudnością zadania i maleje wraz ze zdolnością respondenta³⁵. Osoby lepiej wykształcone zgadują rzadziej, spójnie we wszystkich krajach i dziedzinach³⁶. Wytrzymałość testowa okazuje się po części stabilną cechą osoby³⁷, po części stanem wytwarzanym przez architekturę testu, taką jak jego długość czy ułożenie zadań, a wysiłek testowy koreluje ze wskaźnikami niekognitywnymi, takimi jak sumienność i wytrwałość. Odpowiedź niezaangażowana niesie zatem informację o osobie, a nie tylko jej brak. Pytaniem bez ustalonej odpowiedzi pozostaje, co z tą informacją robić, czy korygować wyniki, raportować je osobno, czy może traktować jako odrębny wynik pomiaru.

Z tych trzech problemów wynika program pracy na najbliższe lata. Potrzebna jest walidacja zewnętrzna metod detekcji, oparta na powtarzalnych i współdzielonych protokołach eksperymentalnych. Potrzebna jest standaryzacja definicji danych procesowych między platformami badawczymi, a więc ustalenie, jakie zdarzenia w interfejsie rejestruje się w jednakowy sposób i jak je etykietować, bez tego bowiem wyniki z różnych systemów pozostaną nieporównywalne. Potrzebne są dalsze badania nad tym, czego obecne rejestry nie obejmują, w duchu omówionej pracy z użyciem eye-trackera, ponieważ każda metoda detekcji jest tylko tak dobra, jak dane, na których pracuje. Coraz pilniejsza staje się refleksja etyczna, paradane przestają być bowiem technicznym zapisem, a stają się informacją wpływającą na ocenę wyniku osoby badanej, choć ich status, a więc to, kto ma do nich prawo, o czym trzeba respondenta poinformować i do czego wolno je wykorzystać, nie jest uregulowany. Wreszcie potrzebne są narzędzia

³³ Tamże.

³⁴ T. Deribo, U. Kroehne, F. Goldhammer, *Model-based treatment of rapid guessing*, „Journal of Educational Measurement” 2021, nr 58(2), s. 281–303.

³⁵ F. Goldhammer, T. Martens, O. Lüdtke, *Conditioning factors of test-taking engagement in PIAAC: An exploratory IRT modelling approach considering person and item characteristics*, „Large-scale Assessments in Education” 2017, nr 5, s. 18.

³⁶ A. Pokropek, L. Khorramdel, dz. cyt.

³⁷ F. Borgonovi, P. Biecek, dz. cyt.

dla praktyków, proste, udokumentowane i dostępne w powszechnie używanych pakietach. Detekcja niezaangażowania osiągnęła dojrzałość techniczną. Teraz musi osiągnąć dojrzałość praktyczną i instytucjonalną, a jest to zadanie w równym stopniu dla teoretyków, twórców oprogramowania i organizatorów badań.

Literatura

- Balart P., Oosterveen M., *Females show more sustained performance during test-taking than males*, „Nature Communications” 2019, nr 10, s. 3798, <https://doi.org/10.1038/s41467-019-11691-y>.
- Borgonovi F., Biecek P., *An international comparison of students' ability to endure fatigue and maintain motivation during a low-stakes test*, „Learning and Individual Differences” 2016, nr 49, s. 128–137, <https://doi.org/10.1016/j.lindif.2016.06.001>.
- Curran P.G., *Methods for the detection of carelessly invalid responses in survey data*, „Journal of Experimental Social Psychology” 2016, nr 66, s. 4–19, <https://doi.org/10.1016/j.jesp.2015.07.006>.
- Deribo T., Kroehne U., Goldhammer F., *Model-based treatment of rapid guessing*, „Journal of Educational Measurement” 2021, nr 58(2), s. 281–303, <https://doi.org/10.1111/jedm.12290>.
- Goldhammer F., Martens T., Christoph G., Lüdtke O., *Test-taking engagement in PIAAC*, OECD Education Working Papers No. 133, OECD Publishing, Paryż 2016, <https://doi.org/10.1787/5jlzfl6fhxs2-en>.
- Goldhammer F., Martens T., Lüdtke O., *Conditioning factors of test-taking engagement in PIAAC: An exploratory IRT modelling approach considering person and item characteristics*, „Large-scale Assessments in Education” 2017, nr 5, s. 18, <https://doi.org/10.1186/s40536-017-0051-9>.
- Krosnick J.A., *Response strategies for coping with the cognitive demands of attitude measures in surveys*, „Applied Cognitive Psychology” 1991, nr 5(3), s. 213–236, <https://doi.org/10.1002/acp.2350050305>.
- Maddox B., Bayliss A.P., Fleming P., Engelhardt P.E., Edwards S.G., Borgonovi F., *Observing response processes with eye tracking in international large-scale assessments: Evidence from the OECD PIAAC assessment*, „European Journal of Psychology of Education” 2018, nr 33(3), s. 543–558, <https://doi.org/10.1007/s10212-018-0380-2>.
- Meade A.W., Craig S.B., *Identifying careless responses in survey data*, „Psychological Methods” 2012, nr 17(3), s. 437–455, <https://doi.org/10.1037/a0028085>.
- Nagy G., Ulitzsch E., *A multilevel mixture IRT framework for modeling response times as predictors or indicators of response engagement in IRT models*, „Educational and Psychological Measurement” 2022, nr 82(5), s. 845–879, <https://doi.org/10.1177/00131644211045351>.
- Pokropek A., *Grade of membership response time model for detecting guessing behaviors*, „Journal of Educational and Behavioral Statistics” 2016, nr 41(3), s. 300–325, <https://doi.org/10.3102/1076998616636618>.
- Pokropek A., *Detecting careless response patterns in Likert scales using supervised machine learning and Monte Carlo simulations*, „Current Psychology” 2026, nr 45, s. 518, <https://doi.org/10.1007/s12144-026-09186-3>.
- Pokropek A., Khorramdel L., *Analyzing test-taking behaviors through process data using the IRT Explanatory Model for Guessing* [w:] L. Khorramdel i in. (red.), *Innovative digital-based international large-scale assessments*, Springer, Cham 2025, s. 559–584, https://doi.org/10.1007/978-3-031-90951-1_23.

- Pokropek A., Żółtak T., Muszyński M., *Approximate Areas of Interest for enhanced understanding of student motivation and task interaction in IEA assessments* [raport techniczny], IEA, Amsterdam 2025, <https://www.iea.nl/publications/rd-outcomes/approximate-areas-interest-enhanced-understanding-student-motivation-and> [dostęp: 31.07.2026].
- Rios J.A., Deng J., *Does the choice of response time threshold procedure substantially affect inferences concerning the identification and exclusion of rapid guessing responses? A meta-analysis*, „Large-scale Assessments in Education” 2021, nr 9(1), s. 1–25, <https://doi.org/10.1186/s40536-021-00110-8>.
- Schnipke D.L., Scrams D.J., *Modeling item response times with a two-state mixture model: A new method of measuring speededness*, „Journal of Educational Measurement” 1997, nr 34(3), s. 213–232, <https://doi.org/10.1111/j.1745-3984.1997.tb00516.x>.
- Ullrich E., Davier M. von, Pohl S., *A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response*, „British Journal of Mathematical and Statistical Psychology” 2020, nr 73(S1), s. 83–112, <https://doi.org/10.1111/bmsp.12188>.
- Ward M.K., Meade A.W., *Dealing with careless responding in survey data: Prevention, identification, and recommended best practices*, „Annual Review of Psychology” 2023, nr 74, s. 577–596, <https://doi.org/10.1146/annurev-psych-040422-045007>.
- Wigfield A., Eccles J.S., *Expectancy-value theory of achievement motivation*, „Contemporary Educational Psychology” 2000, nr 25(1), s. 68–81, <https://doi.org/10.1006/ceps.1999.1015>.
- Wise S.L., DeMars C.E., *Low examinee effort in low-stakes assessment: Problems and potential solutions*, „Educational Assessment” 2005, nr 10(1), s. 1–17, https://doi.org/10.1207/s15326977ea1001_1.
- Wise S.L., DeMars C.E., *An application of item response time: The effort-moderated IRT model*, „Journal of Educational Measurement” 2006, nr 43(1), s. 19–38, <https://doi.org/10.1111/j.1745-3984.2006.00002.x>.
- Wise S.L., Kong X., *Response time effort: A new measure of examinee motivation in computer-based tests*, „Applied Measurement in Education” 2005, nr 18(2), s. 163–183, https://doi.org/10.1207/s15324818ame1802_2.
- Yamamoto K., *HYBRID model of IRT and latent class models*, ETS Research Report No. RR-89-41, Educational Testing Service, Princeton 1989, <https://doi.org/10.1002/j.2330-8516.1989.tb00320.x>.