

**dr Aleksandra Jasińska-Maciążek, Agnieszka Glica,
Marzena Binkiewicz, dr Kinga Karteczka-Świątek,
dr hab. Paweł Grygiel, prof. UJ; Bartosz Kondratek,
dr hab. Artur Pokropek, prof. IFiS PAN**

Instytut Badań Edukacyjnych – Państwowy Instytut Badawczy

Jak oceniać skuteczność technologii edukacyjnych? Wyniki badania RCT narzędzia zeszyt.online

Streszczenie

Rozwój technologii opartych na sztucznej inteligencji wyprzedza tempo powstawania wiarygodnych dowodów dotyczących ich wpływu na uczenie się. Referat przedstawia sposób oceny technologii edukacyjnej na przykładzie ogólnopolskiego randomizowanego badania kontrolowanego platformy zeszyt.online, wspierającej adaptacyjne uczenie się matematyki. Do badania przystąpiło 20 352 uczniów z 548 szkół, losowo przypisanych na poziomie szkoły do grupy eksperymentalnej lub kontrolnej. Trzy pomiary przeprowadzono od marca do listopada 2025 r. Dostęp do platformy zwiększył końcowe wyniki matematyczne o 0,09 odchylenia standardowego, co odpowiada w przybliżeniu jednemu miesiącowi dodatkowego postępu. Nie stwierdzono zmian w lęku przed matematyką, lubieniu przedmiotu ani samoocenie matematycznej. Jednocześnie wykorzystanie narzędzia było silnie zróżnicowane: 18,7% uczniów z grupy eksperymentalnej nie użyło go ani razu, a mediana czasu pracy użytkowników wyniosła 2,5 godziny. Wyniki pokazują, że ocena technologii powinna łączyć identyfikację efektu przyczynowego z analizą faktycznego wdrożenia. Sam dostęp do rozwiązania nie jest równoznaczny z jego systematycznym użyciem, a skala efektów zależy także od sposobu powiązania technologii z pracą nauczyciela.

Słowa kluczowe: technologie edukacyjne, zeszyt.online, adaptacyjne uczenie się, osiągnięcia matematyczne, randomizowane badanie kontrolowane (RCT), wdrażanie innowacji edukacyjnych

Dlaczego technologie edukacyjne wymagają rygorystycznej oceny?

Wdrażaniu technologii edukacyjnych towarzyszą często wysokie oczekiwania: mają personalizować nauczanie, odciążać nauczyciela, zwiększać zaangażowanie uczniów i poprawiać osiągnięcia. Szczególnie duże nadzieje wiąże się obecnie z rozwiązaniami wykorzystującymi sztuczną inteligencję. Systemy adaptacyjne mogą na bieżąco analizować odpowiedzi ucznia, dobierać poziom trudności, proponować zadania pomocnicze oraz planować powtórki. Potencjalnie tworzą więc warunki do bardziej zindywidualizowanego uczenia się niż jednolity zestaw ćwiczeń kierowany równocześnie do całej klasy.

Sama atrakcyjność rozwiązania, pozytywne opinie użytkowników ani obserwowany wzrost wyników nie wystarczają jednak do stwierdzenia, że narzędzie jest skuteczne. Uczniowie rozwijają umiejętności również dzięki standardowemu nauczaniu, dojrzewaniu, pracy własnej i wielu innym doświadczeniom. Dlatego zasadnicze pytanie ewaluacyjne nie brzmi: „czy wyniki się poprawiły?”, lecz: „czy poprawa była większa, niż nastąpiłaby bez danej technologii?”. Odpowiedź wymaga wiarygodnego oszacowania sytuacji kontrfaktycznej, czyli wyniku, jaki osiągnęliby ci sami uczniowie, gdyby nie otrzymali dostępu do interwencji.

Randomizowane badanie kontrolowane (RCT) jest jednym z najmocniejszych schematów służących temu celowi. Losowy przydział do grupy eksperymentalnej i kontrolnej ogranicza ryzyko, że grupy różnią się systematycznie jeszcze przed rozpoczęciem programu. Dzięki temu różnicę zaobserwowaną po interwencji można – przy spełnieniu założeń projektu – interpretować przyczynowo. W edukacji randomizacja często odbywa się na poziomie szkół lub klas, aby ograniczyć przenikanie interwencji między uczniami.

Potrzeba takiej oceny jest szczególnie istotna wobec niejednorodnych wyników badań nad AI w nauczaniu matematyki. Metaanalizy wskazują przeciętnie dodatnie efekty, ale ich wielkość wyraźnie zależy od rodzaju systemu, wieku uczniów, czasu trwania interwencji i sposobu wdrożenia. Yi i współpracownicy (2025) oszacowali średni efekt na około 0,34 odchylenia standardowego, przy czym wyższe wyniki uzyskiwały inteligentne systemy tutoringowe i systemy adaptacyjne. Rozrzut rezultatów między badaniami jest jednak duży. Oznacza to, że pytanie: „czy AI działa w edukacji?” jest zbyt ogólne. Należy raczej pytać, które narzędzie działa, dla kogo, w jakich warunkach i jako element jakiego modelu nauczania.

Zeszyt.online jako adaptacyjne wsparcie nauki matematyki

Zeszyt.online jest internetową platformą, stworzoną przez zespół z Wydziału Matematyki i Nauk Informatycznych Politechniki Warszawskiej, przeznaczoną do wspierania nauczania matematyki w klasach IV–VIII szkoły podstawowej. Zawiera interaktywne zadania zgodne z podstawą programową i wykorzystuje algorytmy do dostosowywania dalszej ścieżki pracy do odpowiedzi, tempa oraz wzorców błędów ucznia. Po udzieleniu odpowiedzi uczeń otrzymuje natychmiastową informację zwrotną. W przypadku gdy uczeń ma trudności, system może zaproponować zadania pomocnicze, które rozkładają problem na prostsze elementy, a następnie zadanie sprawdzające opanowanie danej umiejętności.

Platforma wykorzystuje także powtórki rozłożone w czasie i aktywne przypominanie. Jej celem nie jest maksymalizacja liczby rozwiązanych zadań, lecz doprowadzenie do opanowania określonego zagadnienia. Nauczyciel decyduje o tematach, rodzaju aktywności i organizacji pracy: może wykorzystywać zadania podczas lekcji, jako pracę domową, powtórkę albo przygotowanie do sprawdzianu. Jest to zatem narzędzie projektowane jako uzupełnienie procesu dydaktycznego, nie zaś autonomiczny zamiennik nauczyciela.

Zakładany mechanizm wpływu obejmuje przede wszystkim lepsze dopasowanie poziomu zadań i informacji zwrotnej do potrzeb ucznia. Można oczekiwać, że praca w strefie optymalnej trudności zwiększy liczbę produktywnych prób rozwiązania problemu i ograniczy zarówno nudę, jak i przeciążenie.

Drugim możliwym mechanizmem jest oddziaływanie motywacyjno-emocjonalne: częstsze doświadczenia sukcesu mogłyby wzmacniać samoocenę matematyczną, zwiększać lubienie przedmiotu i redukować lęk. Badanie objęło zatem zarówno osiągnięcia, jak i te trzy postawy wobec matematyki.

Projekt ogólnopolskiego badania RCT

Badanie przeprowadzono od marca do listopada 2025 r. w naturalnych warunkach szkolnych. Szkoły zostały przypisane do warunku eksperymentalnego, w którym otrzymały dostęp do zeszyt.online, albo do warunku kontrolnego, w którym kontynuowano standardowe nauczanie matematyki bez platformy. Randomizacja odbywała się na poziomie szkoły. Do badania przystąpiło 20 352 uczniów z 1006 klas, 548 szkół i 113 gmin: 10 743 uczniów w grupie eksperymentalnej oraz 9609 w kontrolnej.

Pomiar obejmował trzy fale. Pierwszą zrealizowano na przełomie marca i kwietnia, przed udostępnieniem narzędzia. Druga odbyła się po około dwóch miesiącach, na przełomie maja i czerwca. Pomiar końcowy przeprowadzono w listopadzie, już po przejściu uczniów do klasy V. Nauczyciele grupy eksperymentalnej uczestniczyli w szkoleniu, lecz w toku badania samodzielnie decydowali o częstotliwości oraz sposobie wykorzystania platformy. Taka konstrukcja zbliża badanie do próby efektywności wdrożenia (*effectiveness trial*): ocenia skutki rozwiązania w realnej praktyce, a nie w ściśle kontrolowanym, laboratoryjnym scenariuszu.

Umiejętności matematyczne mierzono testami skonstruowanymi w odniesieniu do ram badania TIMSS dla klasy IV, obejmującymi liczby, geometrię i pomiary oraz przedstawianie danych. Zastosowano wiele wersji testu i niekompletny plan blokowy, aby objąć szeroki zakres treści, a jednocześnie ograniczyć obciążenie pojedynczego ucznia. Rzetelność wersji testowych była wysoka. Postawy wobec matematyki oceniano za pomocą skal lubienia matematyki, samooceny umiejętności oraz lęku matematycznego. Analizy psychometryczne potwierdziły możliwość porównywania wyników między grupami.

Główne analizy wykonano dla 11 439 uczniów, którzy uczestniczyli we wszystkich trzech pomiarach testu matematycznego. Efekt programu oszacowano zgodnie z zasadą *intention-to-treat*: porównano wszystkich uczniów przypisanych do grupy eksperymentalnej ze wszystkimi uczniami z grupy kontrolnej, niezależnie od tego, czy i jak intensywnie platforma była faktycznie używana. Takie podejście zachowuje korzyść z randomizacji i odpowiada na pytanie o efekt zaoferowania oraz wdrożenia programu w warunkach rzeczywistych. Modele uwzględniały początkowy poziom wyniku, zagnieżdżenie uczniów w klasach oraz – w analizach rozszerzonych – dodatkowe cechy ucznia i klasy. Stabilność wniosków sprawdzono też na pełnej próbie za pomocą wielokrotnej imputacji braków danych.

Wyniki: niewielki, lecz wiarygodny wzrost umiejętności

Przed rozpoczęciem interwencji grupy eksperymentalna i kontrolna nie różniły się istotnie pod względem umiejętności matematycznych, lęku, lubienia matematyki ani samooceny. Wynik ten jest zgodny z oczekiwaniem wynikającym z losowego przydziału i wzmacnia podstawę do późniejszego porównania grup.

W pomiarze końcowym uczniowie ze szkół, którym udostępniono zeszyt online, uzyskali przeciętnie wynik wyższy o 0,09 odchylenia standardowego niż uczniowie z grupy kontrolnej, po uwzględnieniu początkowego poziomu umiejętności matematycznych i czasu wypełnienia ankiety przed przystąpieniem do testu podczas trzeciego pomiaru. Efekt był istotny statystycznie i praktycznie nie zmieniał się po wprowadzeniu dodatkowych zmiennych kontrolnych. Podobny rezultat uzyskano w analizie pełnej próby z imputowanymi brakami danych. Można go przeliczyć orientacyjnie na około jeden miesiąc dodatkowego postępu w nauce.

Wartość 0,09 SD może na pierwszy rzut oka wydawać się mała, lecz jej interpretacja wymaga właściwego punktu odniesienia. W dużych i rygorystycznych ewaluacjach edukacyjnych mediana efektów wynosi około 0,10 SD (Evans i Yuan, 2022; Kraft, 2023). Rezultat zeszyt.online jest więc zbliżony do typowej wielkości efektu interwencji edukacyjnej. Jednocześnie pozostaje niższy od średnich efektów raportowanych dla systemów AI w części metaanaliz. Ta różnica nie musi oznaczać słabszej jakości narzędzia. Metaanalizy łączą badania o różnym rygorze, wieku uczestników i warunkach wdrożenia, a próby skuteczności prowadzone przy ścisłej kontroli realizacji zwykle dają wyższe oszacowania niż szerokie wdrożenia w szkołach (Durlak i DuPre, 2008).

Analiza pomiaru pośredniego wykazała, że dodatni efekt był widoczny już przed wakacjami i wynosił 0,07 SD. Po całym okresie wzrósł do 0,09 SD. Oznacza to, że przewaga grupy eksperymentalnej pojawiła się stosunkowo wcześniej. Jednocześnie po wakacjach korzystanie z platformy osłabło, co mogło ograniczyć dalszy przyrost efektu.

Nie stwierdzono natomiast istotnego wpływu interwencji na lęk przed matematyką, lubienie przedmiotu ani samoocenę matematyczną. Wynik ten pokazuje, że poprawa umiejętności nie musi automatycznie prowadzić do zmiany bardziej ogólnych postaw. Platforma była ukierunkowana przede wszystkim na ćwiczenie i adaptację procesu rozwiązywania zadań, a przeciętny czas ekspozycji mógł być zbyt krótki, aby zmienić utrwalone przekonania i emocje związane z matematyką.

Dostęp do technologii a jej rzeczywiste wykorzystanie

Jednym z ważnych rezultatów badania jest silne zróżnicowanie wdrożenia. W całym okresie 18,7% uczniów z grupy eksperymentalnej nie skorzystało z platformy ani razu. Wśród użytkowników mediana łącznego czasu pracy wyniosła tylko 2,5 godziny, choć pojedynczy uczniowie korzystali nawet około 60 godzin. Przed wakacjami nieaktywnych było 24% uczniów, natomiast po wakacjach – 43,7%. Dane te unaocniają podstawową różnicę między udostępnieniem technologii a jej użyciem.

W analizach eksploracyjnych uczniowie, którzy rzeczywiście korzystali z platformy, osiągnęli wyniki wyższe o około 0,13 SD od grupy kontrolnej. Uczniowie mający dostęp, lecz niekorzystający z narzędzia, nie różnili się od grupy kontrolnej. Przy bardziej szczegółowym podziale dodatnie różnice pojawiały się od około godziny łącznego użytkowania i były największe wśród uczniów pracujących najdłużej. Wynik ten jest zgodny z intuicją, że narzędzie może oddziaływać tylko wtedy, gdy jest używane.

Nie należy jednak traktować zależności między czasem korzystania a wynikiem jako dowodu przyczynowego. Intensywność nie była losowana. Uczniowie, którzy zdecydowali się pracować z systemem, różnili się od pozostałych: częściej były to dziewczynki, osoby bardziej lubiące matematykę i uzyskujące wyższe oceny. Z drugiej strony uczniowie z niższym początkowym wynikiem testowym, o ile rozpoczęli pracę, korzystali z platformy dłużej. W liczniejszych klasach prawdopodobieństwo podjęcia pracy było niższe, a zmiana nauczyciela po wakacjach wiązała się z krótszym czasem użytkowania.

Te wyniki mają znaczenie zarówno metodologiczne, jak i praktyczne. Metodologicznie pokazują, dlaczego podstawą wnioskowania o przyczynowym efekcie programu powinna pozostać analiza zgodna z losowym przydziałem, a analizy użytkowników należy traktować jako uzupełniające. Praktycznie wskazują, że technologia nie wdraża się sama. Jej użycie zależy od organizacji pracy klasy, ciągłości prowadzenia zajęć, motywacji uczniów i działań nauczyciela. Bez celowego włączenia platformy w tok lekcji jej potencjał może pozostać niewykorzystany, a korzyści mogą częściej trafiać do uczniów, którzy już na starcie chętniej angażują się w matematykę.

Jak oceniać skuteczność technologii edukacyjnych – wnioski

Przypadek zeszyt.online prowadzi do kilku szerszych wniosków dotyczących ewaluacji technologii edukacyjnych. Po pierwsze, ocena powinna wykraczać poza satysfakcję użytkowników i deklarowaną użyteczność. Pozytywna opinia jest ważna dla akceptacji rozwiązania, ale nie zastępuje pomiaru efektów uczenia się. W badaniu większość użytkowników dobrze oceniała łatwość obsługi i chciała dalej korzystać z platformy, jednak dopiero porównanie z grupą kontrolną pozwoliło określić, czy narzędzie rzeczywiście poprawia wyniki.

Po drugie, konieczne jest odróżnienie efektu dostępu od efektu regularnego używania. Analiza *intention-to-treat* odpowiada na pytanie istotne dla decydentów, jaki rezultat przynosi wdrożenie programu w zwykłych warunkach, wraz z naturalnym zróżnicowaniem wykorzystania. Równolegle trzeba mierzyć implementację – czas pracy, regularność, rodzaj zadań, aktywność nauczyciela i bariery organizacyjne. Bez tych danych trudno wyjaśnić, dlaczego efekt jest większy w jednych klasach, a mniejszy w innych.

Po trzecie, wielkość efektu należy interpretować w odniesieniu do realistycznych benchmarków. Oczekiwanie bardzo dużych efektów może prowadzić do odrzucania użytecznych rozwiązań, natomiast zachwyt nad każdą istotną statystycznie różnicą – do przeceniania ich znaczenia. Efekt 0,09 SD jest niewielki, ale porównywalny z typowymi rezultatami interwencji edukacyjnych

i uzyskany w dużym badaniu w warunkach naturalnych. Jego znaczenie zależy również od kosztów wdrożenia, skalowalności, trwałości i tego, czy narzędzie uzupełnia, czy wypiera inne skuteczne działania.

Po czwarte, technologię trzeba oceniać jako element systemu dydaktycznego, nie jako izolowany produkt. Literatura wskazuje, że większe efekty pojawiają się, gdy narzędzia adaptacyjne są integrowane z pracą nauczyciela, a dane z platformy służą do planowania dalszego nauczania (Holman i in., 2025). Wyniki badania zeszyt.online prowadzą do podobnego wniosku: dalsze wdrażanie powinno obejmować nie tylko dostęp techniczny i jednorazowe szkolenie, lecz także wsparcie nauczycieli w planowaniu regularnej pracy, monitorowanie zaangażowania uczniów i rozwiązania ograniczające spadek użycia w czasie.

Po piąte, potrzebne są kolejne badania pozwalające zidentyfikować warunki maksymalizujące efekty. Warto eksperymentalnie porównywać różne poziomy intensywności, modele integracji z lekcją oraz formy wsparcia nauczyciela. Istotne jest również badanie innych grup wiekowych, trwałości rezultatów i wpływu na uczniów o zróżnicowanym poziomie osiągnięć. W odniesieniu do postaw wobec matematyki bardziej adekwatne mogą być krótkoterminowe miary doświadczeń emocjonalnych podczas pracy, a także interwencje bezpośrednio ukierunkowane na poczucie kontroli, wartość zadania i lęk.

Podsumowanie

Ogólnopolskie badanie RCT dostarczyło dowodów, że udostępnienie platformy zeszyt.online szkołom prowadzi do niewielkiej, lecz wiarygodnej poprawy umiejętności matematycznych uczniów klas IV–V. Efekt 0,09 SD odpowiada w przybliżeniu jednemu miesiącowi dodatkowego postępu i jest zbliżony do typowych efektów interwencji edukacyjnych. Nie zaobserwowano równoległych zmian w lęku przed matematyką, lubieniu przedmiotu ani samoocenie matematycznej.

Najważniejszym warunkiem wykorzystania potencjału platformy jest rzeczywiste i systematyczne użycie. Znaczna część uczniów nie korzystała z narzędzia lub pracowała z nim bardzo krótko, a intensywność spadała po wakacjach. Dlatego skuteczna polityka wdrożeniowa powinna łączyć decyzję o udostępnieniu technologii z działaniami wspierającymi nauczycieli i uczniów w jej regularnym stosowaniu. Badanie pokazuje zarazem, że rzetelna ocena technologii edukacyjnej wymaga dwóch perspektyw: eksperymentalnego oszacowania, czy narzędzie powoduje zmianę wyników, oraz analizy implementacji, która wyjaśnia, w jakich warunkach jego potencjał faktycznie się ujawnia.

Literatura

- Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, 41(3–4), 327–350. <https://doi.org/10.1007/s10464-008-9165-0>
- Evans, D. K., & Yuan, F. (2022). How big are effect sizes in international education studies? *Educational Evaluation and Policy Analysis*, 44(3), 532–540. <https://doi.org/10.3102/01623737221079646>
- Holman, K., Marino, M., Vasquez, T., Taub, M., Hunt, J., & Tazi, Y. (2025). Artificial intelligence interventions in mathematics education: A systematic literature review. *Insights into Learning Disabilities*, 22(1), 65–91.
- Kraft, M. A. (2023). The effect-size benchmark that matters most: Education interventions often fail. *Educational Researcher*, 52(3), 183–187. <https://doi.org/10.3102/0013189X231155154>
- Yi, L., Liu, D., Jiang, T., & Xian, Y. (2025). The effectiveness of AI on K–12 students' mathematics learning: A systematic review and meta-analysis. *International Journal of Science and Mathematics Education*, 23(4), 1105–1126. <https://doi.org/10.1007/s10763-024-10499-7>

Informacja o finansowaniu:

Badanie realizowano w ramach zadania „Ewaluacja programów i przedsięwzięć istotnych dla rozwoju systemu kształcenia i uczenia się przez całe życie”, finansowanego ze środków Ministerstwa Edukacji Narodowej na podstawie umów nr MEN/2025/DIR/89 oraz MEN/2026/DIR/38.