Dr hab. Roman Dolata, prof. uczelni

Wydział Pedagogiczny Uniwersytetu Warszawskiego

rdolata@uw.edu.pl

Recenzja przygotowana na potrzeby postępowania habilitacyjnego dr Jacka Stańdo

Jacek Stańdo jest magistrem matematyki i informatyki oraz doktorem matematyki. Jego *Alma Mater* to Uniwersytet Łódzki. Studiował też na Uniwersytecie w Utrechcie. Obroniony w 2004 r. doktorat dotyczył zagadnień *stricte* matematycznych. W 2005 roku uzyskał stopień nauczyciela dyplomowanego.

Do oceny w ramach procedury habilitacyjnej dr Stańdo przedstawił wydaną w 2023 roku monografię *Trajektorie i walidacja efektów uczenia się na przykładzie przedmiotu matematyka*. Wydawnictwo Politechniki Łódzkiej (poziom I wg punktacji ministerialne). Główna część monografii liczy 294 strony, bibliografia to 18 stron, załączniki to kolejne 66 stron. Spis literatury jest bardzo obszerny.

Monografia ta będzie głównym przedmiotem oceny w mojej recenzji. Jednak najpierw ocenię pozostały dorobek Habilitanta pod kątem jego wkładu w rozwój dyscypliny.

1. Osiągnięcia i aktywność naukowa poza przedstawioną do oceny monografią

Po doktoracie Habilitant był autorem lub współautorem:

1) 20 rozdziałów w pracach zbiorowych

Większość tekstów mieściła się w obszarze pedagogiki. Żaden z tych rozdziałów nie był opublikowany w wydawnictwie z II poziomu, najnowszy jest datowany w 2013 roku. Teksty najczęściej dotyczyły technologii IT w ocenie osiągnięć uczniów. Częstym tematem była implementacja w polskim systemie egzaminacyjnym e-oceniania. Badania Habilitanta w tym zakresie są cennym wkładem w rozwój technologii pomiaru edukacyjnego w Polsce, brak im jednak szerszego kontekstu teoretycznego i umieszczenia testowanych rozwiązań w kontekście stanu wiedzy w tym zakresie na świecie. Prace te były rzadko cytowane przez innych badaczy.

2) 34 artykułów w czasopismach naukowych lub specjalistycznych

Dwie trzecie z tej puli publikacji dotyczyło bezpośrednio lub przynajmniej pośrednio pedagogiki. Trzeba jednak zauważyć, że spośród 19 artykułów opublikowanych w latach 2019-2023, aż 13 nie mieści się w polu pedagogiki. Dwa najwyżej punktowane artykuły (w momencie publikacji 140 pkt, aktualnie 20 pkt) ukazały się w czasopiśmie wydawanym przez MDPI.

Dorobek publikacyjny dr Stańdo oceniam jako ponadprzeciętny, choć jest on tematycznie bardzo zróżnicowany i trudno w nim dostrzec linię rozwojową prowadzącą do rozprawy habilitacyjnej. Widać raczej kilka obszarów zainteresowań. Jedyną dominantą tematyczną jest zastosowanie IT w szeroko rozumianej edukacji.

Przed doktoratem Habilitant opublikował trzy artykuły spoza pedagogiki.

Dr Stańdo był autorem 10 recenzji artykułów zgłoszonych do czasopism naukowych, głównie wydawanych przez MDPI lub podmioty krajowe.

Wpływ prac dr Stańdo na innych badaczy wg miar bibliometrycznych (w dniu 23.03.2024 r.) wygląda następująco:

1) Google Scholar: 217 cytowań, znaczna część tych cytowań dotyczy prac współautorskich spoza pedagogiki

2) Scopus: 56 cytowań, h-index 5; znaczna część cytowań dotyczy prac współautorskich spoza pedagogiki

Habilitant brał udział w komitetach naukowych lub organizacyjnych 123 konferencji o zasięgu międzynarodowych, głównie o tematyce IT. To imponujący dorobek w sferze organizacji nauki.

Dr Stańdo kierował lub był wykonawcą w 27 projektach, głównie wdrożeniowych z funduszu POWR i PO KL. Habilitant nie był kierownikiem ani wykonawcą w żadnym projekcie finansowym przez NCN ani konkursowo przyznawanych grantach UE.

Habilitant jest autorem 123 recenzji podręczników, programów i środków dydaktycznych na potrzeby MEN.

Różnorodna aktywność naukowa dr Stańdo jest niezwykle bogata. To z pewnością mocny atut Habilitanta.

Ocena przedstawionej do oceny monografii naukowej

Badania podstawowe mają za cel rozwój teorii, badania stosowane wykorzystują teorie do rozwiązywania problemów praktycznych. W obszarze tych drugich badań znajdą się też innowacje rozumiane jako nowatorska odpowiedź na potrzeby praktyki społecznej. To co łączy te typy badań, to metoda naukowa. Oczywiście podział na te dwa typy badań naukowych nie jest ostry i mogą się one przenikać.

Czytając uzasadnienie wyboru tematu jak i sformułowanie głównego problemu badawczego widzimy, że przedstawione do oceny osiągnięcie naukowe to przede wszystkim badanie stosowane: zasoby teoretyczne i metoda naukowa mają być narzędziami do zaproponowania innowacji. Ma ona być naukowo uzasadnionym systemem weryfikacji efektów uczenia się powiązanych z poziomami Polskiej Ramy Kwalifikacji w zakresie nauczania matematyki. Jednak pytania badawcze od 6 do 9 to problemy typowe dla badań podstawowych, na przykład PB7: „Czy istnieją różnice między liczbą powrotów do rozwiązania zadania przez uczniów lepszych i słabszych?” Mamy zatem do czynienia z pracą niejako dwuczęściową.

Jakie kryteria oceny należy zatem przyjąć do oceny przedstawionej pracy naukowej? Oczywiści głównym, niejako nadrzędnym kryterium, jest wkład prac badawczych Habilitanta do rozwoju dyscypliny. W wypadku badań stosowanych wkład ten będzie polegał na rozwiązaniu ważnego problemu praktycznego. Problem powinien być precyzyjnie zdefiniowany, a jego rozwiązanie powinno korzystać z adekwatnie dobranych zasobów teoretycznych nauk społecznych ze szczególnym uwzględnieniem pedagogiki. Badania mające zweryfikować skuteczność proponowanego rozwiązania, powinny w pełni respektować metodologię naukową.

Natomiast w wypadku drugiej części pracy adekwatne będą kryteria swoiste dla badań podstawowych. Punktem wyjścia musi być przedstawienie stanu badań danego zagadnienia. Następnie krytyczny ich przegląd musi doprowadzić do wartej testowania hipotezy. Może to być hipoteza w pełni oryginalna lub hipoteza już testowana, ale dla której wyniki dotychczasowych badań są rozbieżne lub budzą uzasadnione wątpliwości metodologiczne.

Ostatecznym kryterium jest rozwój teorii danego zjawiska lub zaproponowanie nowej, lepszej teorii.

Opisane kryteria wyznaczają strukturę mojej recenzji. Ponieważ praca jest wielowątkowa i trudno sformułować całościową ocenę, będę dla wyróżnionych kryteriów kolejno omawiał fragmenty pracy.

I. Badanie stosowane

Poniżej prześledzimy fragmenty pracy poświęcone głównemu badaniu stosowanemu. Zajmę się kolejno precyzją sformułowania problemu badawczego, adekwatnością wyboru zasobów teoretycznych i poprawnością metodologiczną badań sprawdzających skuteczność zaproponowanego rozwiązania

Precyzja sformułowania problemu

s. 132-133. Długi wywód dotyczący tego, czy walidacja jest sprawdzaniem czy potwierdzaniem jest wykwitem biurokracji. Choć w słowniku nauk technicznych faktycznie walidacja oznacza dostarczenie dowodu na coś, to zgodnie z SJP i w języku nauk społecznych to proces szacowania trafności czegoś. Czyli walidacja to proces sprawdzania, który może się skończyć pozytywnie lub negatywnie. W przytoczonych przez Autora dokumentach walidacja jest zresztą definiowana jako składająca się z kilku etapów. Zatem gdy walidacja jest w toku, to nieznany jest jej rezultat. Z pewnością spory terminologiczne to sprawa trzeciorzędna, byle było widomo, o co chodzi. Generalnie rozwlekłe rozważania terminologiczne (poza walidacją: kompetencja, weryfikacja) niepotrzebnie zajmuje czas i energię. A i tak kluczowa sprawa pozostaje mocno zagmatwana. Na stronie 133. czytamy, że *walidacja to potwierdzenie osiągnięcia przez osobę ubiegającą się o nadanie kreślonej kwalifikacji wyodrębnionej części lub całości efektów uczenia się wymaganych dla danej kwalifikacji*. Czyli walidacja dotyczy osiągnięcia przez jednostkę określonych efektów uczenia się. Natomiast – poza modelem dyskretnym walidacji – pozostałe są odnoszone do grupy/populacji.

S. 142. Kluczowe i oryginalne pojęcie testu walidacyjnego zostało wprowadzone bez głębszego uzasadnienia. Z pewnością trzeba było przywołać koncepcja oceniania kryterialnego (*criterion-referenced vs norm-referenced,* Robert Glaser) i uczenia do mistrzostwa (mastery-based learning, Benjamin Bloom).

Adekwatny wybór zasobów teoretycznych i uzasadnione w świetle stanu wiedzy sformułowanie problemu badawczego

S.36. Wątpliwości budzi dobór przykładów do omawianych kategorii taksonomicznych Blooma. Np. dziedzina poznawcza: *Opisz, jak skonstruować dwusieczną kąta*. Dziedzina psychomotoryczna: *Narysuj dowolny kąt, a następnie skonstruuj jego dwusieczną*. Czasami przykłady wydają się wręcz satyryczne: *Wyrażenie przez ruch ciałem obiektów, symboli matematycznych -> Reakcja ruchu głowy prawo/lewo na wyrażenie zaprzeczenia.* Zabrakło natomiast odniesień do adekwatnej dla tego obszaru koncepcji *embodied cognition (poznanie ucieleśnione)* czy klasycznej teorii reprezentacji enaktywnej (Bruner). Co najważniejsze – brak odniesień do problemu badawczego. Ta ostatnia uwaga dotyczy zresztą wielu zasobów teoretycznych przywołanych w pierwszej części pracy. Na stronie s. 87. i dalszych znajdujemy omówienie pojęcia *myślenie krytyczne*. W tym wypadku też brak odniesienia do problemu badawczego. Zabrakło również wykorzystania klasycznych prac z tego zakresu.

S. 139. Teoria odpowiedzi na zadanie testowe jest ważnym zasobem teoretycznym w tej pracy. Czytamy: *Modele IRT (Item Response Theory) są stosowane coraz częściej w badaniach edukacyjnych. Jest to narzędzie, dzięki któremu możemy analizować uczącego się w odniesieniu do wybranego zadania w teście, a nie całego testu. Zadanie może pełnić funkcję walidacji (lub weryfikacji) efektu uczenia się. Niestety jedną z głównych wad zastosowania probabilistycznej teorii testów jest konieczność dużej próby (powyżej 1000 jednostek).* Teoria odpowiedzi na zadanie testowe nie zawdzięcza swej nazwy temu, że pozwala, jak pisze Autor: *analizować uczącego się w odniesieniu do wybranego zadania w teście, a nie całego testu.* Wręcz odwrotnie. Modele IRT pozwalają z jednej strony określać parametry zadań, z drugiej strony parametr opisujący prawdopodobny poziom badanej cechy u rozwiązującego. Jedno i drugie wymaga kontekstu testu, czyli zbioru zadań uznanych za dobre wskaźniki badanej umiejętności i spełniającego wymagania związane z przyjętym modelem cechy ukrytej. Charakterystyka wykonania danego zadania przez rozwiązującego x nie jest domeną ani KTT, ani IRT. To raczej domena SOLO. Również mało precyzyjne jest określenie, że IRT wymaga *dużej próby (powyżej 1000 jednostek).* Trzeba odróżnić fazę konstrukcji testu (banku zadań) i fazę stosowania testu. Przyjmuje się, że wiarygodne określenie parametrów zadań

wymaga min. n=200 na jeden parametr (czyli modele Rascha, model 1PL). W fazie stosowania testu oczywiście żadnych wymagań dotyczących *n* nie ma.

S. 139-141. Opis podejścia IRT jest niezwykle zdawkowy, chaotyczny i niestety mało kompetentny. Brak odniesień do bogatej literatury, również w języku polskim. Nie omówiono kluczowych pojęć dla IRT. Podrozdział ten sprawia wrażenie, jakby wiedza o IRT była czerpana jedynie z programu komputerowego IRT Illustrator. Tytuł podrozdziału sugeruje, że model Rascha będzie w centrum uwagi. Niestety poza wzorem brak jego charakterystyki, przede wszystkim wyłożenia aksjomatyki tego podejścia i omówienia specyfiki modelu Rascha w stosunku do modeli 1PL, 2PL, 3PL i 4 PL.

S. 157. i dalsze. Opis różnych strategii oceniania zadań testowych oparty jest na bardzo skromnej bazie literatury przedmiotu. Ilustrowany jest przykładami tylko z polskiego „poletka". Reprodukcje kart odpowiedzi do egzaminów realizowanych przez CKE bez odniesienia do tematu pracy jest niepotrzebnym zwiększaniem objętości pracy.

s. 163. Opisane modele walidacji efektów uczenia tylko luźno wiążą się z istniejącymi teoriami pomiaru. Nazewnictwo jest mylące. Na przykład „model dyskretny" okazuje się analizą poszczególnych zadań traktowanych jako wskaźniki osiągnięcia efektu uczenia się. „Dyskretność", jak można rozumieć, ma polegać na ocenie dychotomicznej dla danego ucznia: wykonał w pełni poprawnie *vs* wykonał nie w pełni poprawnie zadane-walidator. Z kolei nazwa „klasyczna" walidacja sugeruje wykorzystanie KTT. Nie bardzo wiadomo w jaki sposób. Rdzeniem KTT jest teoria rzetelności i wynikająca z niej metoda szacowania przedziałowego wyniku. Nie widać odwołania do pojęcia rzetelności pomiaru, jest za to wykorzystanie jednej ze skal standardowych do przeliczania wyników surowych – skali staninowej. Zaproponowana definicja klasycznej walidacji brzmi: *Powiemy, że dokonano klasycznej walidacji efektu uczenia się W90×90, jeśli w zakresie od niskiego do najwyższego stanina zadanie mierzące efekt uczenia się jest bardzo łatwe, to znaczy rozwiązywalność wynosi powyżej 90% (s. 165).* Nie wiadomo, czego dotyczy określenie: *w zakresie od niskiego do najwyższego stanina*. Można domyślać się, że chodzi o wynik ogólny w jakimś teście (jakim, co on mierzy? jaką ma rzetelność?) przeliczony na skalę staninową. Czyli osiągnięcie efektu kształcenia mierzonego zadaniem X jest zwalidowane (potwierdzone), gdy uczniowie z poziomów skali staninowej 3-9 mają min. 90% szansy poprawnego wykonania go. Dlaczego wyklucza się kategorie: stanin 1 i 2? Dlaczego przyjęto poziom rozwiązywalności 90%? Na

koniec omawiania modelu „klasycznego” najważniejsza uwaga: w przeciwieństwie do modelu „dyskretnego”, ten odnosi się do grupy/populacji a nie do jednostki, a przecież zgodnie z definicją walidujemy efektu kształcenia na poziome jednostki. Czy chodzi o określenie prawdopodobieństwa, że jednostka z danej grupy/populacji uzyska walidację danego efektu kształcenia? Po co zatem ten dziwny „model klasyczny”? Najlepszym oszacowaniem tego prawdopodobieństwa byłaby łatwość zadania uznanego za walidator danego efektu uczenia się. Problem ten powtarza się też w kolejnych modelach, więc do tej kwestii jeszcze powrócę.

s. 167. Podrozdział *Walidacja probabilistyczna* zaczyna się od akapitu: *Naukowcy coraz częściej stosują walidację opartą nie na klasycznej teorii testu, ale wykorzystującą model Rascha (Gong 2014). Wyniki testu przedstawiane są za pomocą tzw. „mapy Wrighta”, pokazującej szczegóły konstrukcji (Boone i in. 2014).* Niestety nie dowiadujemy się w dalszej części, jak walidacja oparta na KTT równocześnie wykorzystuje model Rascha. Oczywiście w praktyce tworzenia testów w podejściu IRT wykorzystuje się narzędzia z KTT, ale jako pomocnicze. Jak by miała wyglądać fuzja KTT i IRT, Autor nie zdradza. Przywołany w przypisie artykuł (Gong 2014) jest typowym raportem z badania własności psychometrycznych testu i nie ma nic wspólnego z problematyką pracy (w przypisie zabrakło drugiego autora, w bibliografii jest). Mapa Wrighta pojawia się jak *filip z konopi* i nie bardzo wiadomo, co ma wyjaśniać (skala theta jakaś dziwna). Druga przewołana pozycja też nic nie wyjaśnia (to rozdział *Wright Maps: First Steps* z podręcznika do IRT)

Zdefiniowanie kryterium walidacji za pomocą parametru trudności zadania (parametr b) jest problematyczne. Dlaczego przyjąć b=-0,9 (jak rozumiem na skali theta), jako kryterium? Jednostka o poziomie cechy ukrytej -0,9 ma 50% szansy na rozwiązanie zadania o takiej właśnie trudności. Dlaczego kryterium dotyczy trudności zadania a nie poziomu umiejętności rozwiązującego? Dalej czytamy *W przypadku walidacji efektu uczenia się, w sytuacji, gdy nie uczestniczy cała populacja, współczynnik trudności powinien spełniać bardziej rygorystyczny warunek: $b < -1{,}6$. Ale dlaczego?* Co właściwie znaczy, że kryterium walidacji odnosi się do cechy zadania, a nie rozwiązującego? Czy należy przez to rozumieć, że walidacja jest pozytywna, gdy uczeń rozwiąże takie zadanie? Czy mówimy jednak o przewidywaniu i prawdopodobieństwie poradzenia sobie z tym zadaniem? Jeżeli chodzi o 100% pewności, to model tego nie przewiduje (prawdopodobieństwo 1 to górna asymptota krzywej

logistycznej). Jak nie 100%, to jakie prawdopodobieństwo? Poza tym parametr b zadania X nie jest autonomiczną cechą zadania, ale zadania X w kontekście testu lub banku zadań (theta to cecha ukryta wyekstrahowana z macierzy kowariancji konkretnego zestawu zadań). Dodatkowo w podejściu IRT ważne są parametry dopasowania zadania. Modele IRT dla każdego zadania podlegają empirycznemu testowi dopasowania do danych. Do testu/banku zadań wchodzą zadania o zadawalających wartościach współczynników dopasowania (w pierwszym etapie tworzenia testu/banku jest to procedura iteracyjna, następnie zakwalifikowane zadania stają się kotwicami w procedurze uzupełniania testu/banku).

s. 170. Podrozdział zatytułowany *Walidacja probabilistyczna z parametrem b* zaczyna się definicji kolejnego modelu walidacji: *Powiemy, że dokonano probabilistycznej walidacji efektu uczenia się z parametrem b (Wp(b)), jeśli w modelu Rascha zadanie mierzące efekt uczenia się ma współczynnik trudności* $b > -0{,}9$. Czym ta definicja różni się od poprzedniej? Przypomnijmy: *Powiemy, że dokonano probabilistycznej walidacji efektu uczenia się Wp, jeśli w modelu Rascha zadanie mierzące efekt uczenia się ma współczynnik trudności* $b < -0{,}9$. Na czym polega różnica? Trudno pojąć.

S. 173. Podsumowując efekty walidacji na wybranym przykładzie Autor pisze: *Zatem w klasie szóstej (drugi poziomu Polskiej Ramy Kwalifikacji)* ***nie osiągnięto*** *walidacji probabilistycznej efektu uczenia się: wykonuje obliczenia procentowe dla 5%, 10%, 20%, 25%, 50%, 75%. Powinien być spełniony warunek:* $b < -0{,}9$. *Wynik badań uświadamia trudny do zaakceptowania fakt, że dziś uczeń polskiej szkoły nie opanował umiejętności działania na procentach.* Faktycznie parametr b dla wybranego zadania wyniósł -0,78, czyli było trochę trudniejsze niż kryterium. Ale czy z tego wynika, że *uczeń polskiej szkoły (klasa VI) nie opanował umiejętności działania na procentach?* Kto to jest „uczeń polskiej" szkoły? Czy to tylko figura retoryczna bez empirycznego odpowiednika? Jeżeli tak, to twierdzenie to jest nieweryfikowalne. A może chodzi o większość? To by można zweryfikować, ale Autor nie umieścił informacji o rozkładzie wyników na skali theta. Gdyby nawet rozkład był znany, to i tak walidacja jest mocno niepewna, bo sądy w tej sprawie na podstawie IRT można wyrażać tylko jako prawdopodobieństwa. Na dodatek związek między wybranym efektem uczenia się a zadaniem nie jest bezdyskusyjny (sam Autor pisze o tym na s. 174: *…ten sam efekt uczenia się może być mierzony przy pomocy zadań w zmodyfikowanej wersji, a ich rozwiązywalności*

mogą się istotnie różnić.). Dobór wartości liczbowych z pewnością modyfikowałby trudność. Również kontekst opisywany przez treść zadania mógłby zmieniać wartość parametru b.

S. 175. W tym miejscu tekstu pojawia się *ad hoc* pojęcie *poziom walidacji*. Czytamy: *Przyjmijmy, że poziom A stanowi pozytywną walidację efektu uczenia się dla grupy, populacji, natomiast poziomy B i C stanowią negatywną walidację. Poziomy B i C są jedynie pomocne w dążeniu do osiągnięcia poziomu A.* Jak już pisałem, o i ile pierwszy model walidacji odnosi się do jednostki, to kolejne są tak sformułowane, że przedmiotem walidacji staje się grupa/populacja. Tu znów pojawia się ta kwestia. Przecież walidacja polega na potwierdzeniu, że dana osoba starająca się o uzyskania danej kwalifikacji osiągnęła wymagane efekty uczenia. Czy chodzi – jak już pisałem – o określenie prawdopodobieństwa wykonania danego zadania-walidatora przez losowo wybraną jednostkę z danej populacji? Ale do tego nie potrzeba IRT. Również samo zdefiniowanie poziomów nie jest w żaden sposób uzasadniane przez Autora.

S. 176. – s. 188. W rozdziale *Analiza krytyczna matury z matematyki w kontekście walidacji efektów uczenia* się czytamy: *Można zatem przyjąć, że uczeń, który zdał maturę na poziomie podstawowym z matematyki, na egzaminie zewnętrznym potwierdził wszystkie efekty uczenia się zapisane w podstawie programowej kształcenia ogólnego*. Wcześniej jednak sam Autor przywołuje definicję testu Daniela Koretza, która brzmi: *test to mała próbka zadań, której używamy do oszacowania opanowania przez uczniów (uczących się) szerokiego wachlarza wiadomości i umiejętności* (s. 141). Zatem założenie jest niezgodne z samą naturą testu. Rozumiem, że można mieć zastrzeżenia do używania logiki testu w egzaminowaniu, ale trudno testom maturalnym robić zarzut z tego, że są testami.

W mojej ocenie cały ten rozdział jest nic nie wnoszącym ćwiczeniem arytmetycznym. W podsumowaniu czytamy: *Świadectwo szkoły podstawowej, liceum lub technikum, świadectwo maturalne, dyplom inżyniera, dyplom magistra w zakresie matematyki jest jedynie potwierdzeniem ukończenia procesu uczenia się, a* ***nie walidacji efektów uczenia się****. Przeprowadzone w tym podrozdziale rozważania uzasadniają tę tezę.* Z jednej strony wniosek ten jest banalny, z drugiej nietrafny. Na przekład świadectwo maturalne nie jest potwierdzeniem ukończenia procesu uczenia się, a zaliczenia egzaminu przy ustalonym kryterium ilościowym. Czy test egzaminacyjny użyty do egzaminowania jest wysokiej jakości (rzetelność, trafność) i czy kryterium 30% jest sensowne, to inna sprawa. Nie sposób

wyobrazić sobie testowania egzaminacyjnego, które walidowało by wszystkie efekty uczenia się zapisane w PP. Jedynie ocenianie „w procesie", wewnątrzszkolne, miałoby szanse na zbliżenie się do takiej kompletności. Zresztą i to by nie rozwiązywało problemu trwałości tych efektów. Czy efekty uczenia się zwalidowane w trakcie nauki, pozostają w zasobach jednostki w momencie certyfikacji? *Kultura – to co zostaje, kiedy zapomnisz wszystko, czego się nauczyłeś* (S. Lagerlöf). Odrębną sprawą jest podnoszona już kwestia, co jest przedmiotem walidacji: efekty uczenia się grupa czy jednostki? Świadectwo maturalne dostaje jednostka. O czym właściwie mówi tabela 4.31.? O tym jakie zadania-walidatory, czyli zadania uznane za takie na mocy przyjętych arbitralnie kryteriów, złożyły się na test maturalny w kolejnych latach?

S. 203 i dalsze. Sformułowanie problemu badawczego pozostawia wiele do życzenia. W rozprawie chodzi o rozwiązanie problemu praktycznego polegającego na stworzeniu systemu walidacji efektów uczenia się z matematyki. To jasne, jednak niektóre pytania szczegółowe są sformułowane w sposób trudny do zrozumienia. Na przykład *W jakim stopniu forma przeprowadzenia procesu osiągania efektów uczenia się z matematyki na szóstym poziomie Polskiej Ramy Kwalifikacji jest akceptowalna?* Co to znaczy *forma przeprowadzenia procesu osiągania efektów uczenia się?* Po pytaniach badawczych znajdujemy ściśle skorelowane z nimi hipotezy. Jednak formułowanie hipotez badawczych w wypadku badania stosowanego ma mało sensu. Jak zweryfikować główną hipotezę: *Budowanie i przeprowadzenie walidacji efektów uczenia się z matematyki na różnych poziomach Polskiej Ramy Kwalifikacji dokonuje się za pomocą zdefiniowanych, szczegółowych efektów uczenia się, a do ich pomiaru stosujemy zadania generatorowe.* Jak ją można sfalsyfikować? Czy określenia *dokonuje się* czy *stosujemy* to opis zamysłu badacza, stanu faktycznego czy ma - wbrew gramatyce - charakter normatywny? Te same wątpliwości budzą hipotezy szczegółowe. Na przykład hipoteza *Model efektu uczenia się powinien zawierać: odniesienie do poziomu Europejskiej Ramy Kwalifikacji, odniesienie do taksonomii Blooma i SOLO oraz wskazany i zbadany sposób jego pomiaru.* Jaką treść empiryczną ma ta hipoteza? To opis postulatu Autora a nie hipoteza badawcza.

Na zakończenie tego wątku oceny chciałbym wyjść poza tekst. Dla adekwatności wyboru zasobów teoretycznych ważne jest to, czy nie przeoczono takich, które świetnie pasowałby do problemu, a nie zostały uwzględnione. Mam na myśli rozwijane od ponad dwóch dekad

podejście zwane modelami diagnostycznymi. Pozwolę sobie przywołać fragment artykułu Artura Pokropka (2015) *Modele diagnostyczne (cognitive diagnostic models, CDM) to szeroka gama konfirmacyjnych modeli pomiarowych (…), które łączą założenie o wielowymiarowości cechy ukrytej z założeniem o jej nieciągłym charakterze. Modele te zawdzięczają swą nazwę praktycznym zastosowaniom, które najczęściej skupiają się na diagnostycznych, a nieróżnicujących aspektach pomiaru. Na przykład w badaniach edukacyjnych za pomocą modeli diagnostycznych próbuje się stwierdzić, jakie są słabe, a jakie mocne strony ucznia, jakie aspekty umiejętności już opanował, a z którymi nadal ma trudności. Modele te umożliwiają również analizę relacji między poszczególnymi aspektami wiedzy z konkretnych dziedzin, tym samym umożliwiają wyznaczenie warunków koniecznych dla możliwości przyswajania coraz bardziej złożonej wiedzy*[1]. Podejście to wydaje się świetnie pasować do rozwiązywanego przez Autora problemu badawczego, a zostało całkowicie pominięte.

Poprawność metodologiczna badań sprawdzających skuteczność zaproponowanego rozwiązania

S. 205 i dalsze. Charakterystyka metod badawczych zawiera fragment: *Wyniki egzaminu próbnego można także zaliczyć do dokumentów urzędowych, a zatem metodą badań będzie w takim przypadku analiza treści tych dokumentów. Narzędziem badawczym jest pomiar (walidacja) wybranych efektów uczenia się*. Zaliczenie analizy wyników testowania do analizy dokumentów, to dość karkołomna teza. Analiza dokumentów wiąże się z podejściem holistycznym, jakościowym, a test to metoda badań ilościowych. Poza tym wybór do walidacji efektów uczenia się danych z egzaminu próbnego jest nietrafne. Jeżeli walidujemy efekty uczenia się na 2 poziomie PRK to powinniśmy wykorzystać dane z egzaminu końcowego.

S. 208 i dalsze. Do analizy wyników egzaminu próbnego E8 z matematyki użyto podejścia IRT, jak można się domyślić - model Rascha. W tekście czytamy: *W rozdziale czwartym zostało opisane narzędzie do analizy uczącego się w odniesieniu do wybranego zadania w teście, tzw. model IRT. Analizę IRT zaleca się przeprowadzać pod warunkiem liczebności danych powyżej 1000.* Co oznacza zwrot *do analizy uczącego się w odniesieniu do wybranego zadania* trudno dociec. Ponowne przywołanie w tym fragmencie wymaganie n>1000, też

[1] https://www.researchgate.net/publication/279952422_Modele_diagnostyczne

wymaga komentarza. Gdyby do badania użyto testu o udokumentowanej trafności i rzetelności, to można by go poprawnie użyć do diagnozy nawet jednego ucznia. Natomiast, gdy niezgodnie z regułami sztuki, łączy się etap tworzenia narzędzia z badaniem właściwym, faktycznie duża próba jest wskazana. Jednak, kto wybiera taki *roller coaster*, musi się liczyć z tym, że test nie spełni oczekiwań i nie będzie mieć wystarczającej jakości. A to oznacza wyrzucenie zebranych danych do kosza. Niestety nie znajdujemy w opracowaniu wyników kluczowych w podejściu IRT charakterystyk psychometrycznych: brak analizy jednowymiarowości testu, brak parametrów dopasowania dla zadań (a wymagania w modelu Rascha są bardzo rygorystyczne), brak określenia globalnej rzetelności testu i krzywej informacyjnej testu (pozwala wyznaczać warunkowe błędy pomiaru). Zamieszczone w załączniku i na stronie 209. i 210. ICC dają podstawy do wątpliwości, co do dobroci dopasowania danych do modelu Rascha (albo jak kto woli, modelu Rascha do danych). Na wykresach zabrakło krzywych teoretycznych dla oszacowanych parametrach b lub oszacowanych wartości progów dla zadań *partial credit*. Problem dopasowania jest ważną przesłanką do zaufania do oszacowanych trudności zadań, co jest kluczowe dla użytego modelu walidacji. We wszystkich prezentacjach ICC brakuje krzywych teoretycznych, i co się z tym wiąże, miar dobroci dopasowania do modelu. Niestety bez potwierdzenia dobrego dopasowania wszystkie wnioski (przede wszystkim oszacowania parametru b) są problematyczne.

S. 205. Wyniki walidacji efektów uczenia na podstawie egzaminu próbnego z matematyki E8 Autor opisał na 6 stronach. W tabeli 6.3 wymieniono zadania, które przeszły pomyślnie walidację na podstawie wartości parametru b. Przyjęto kryterium -0,9. Nie znajdujemy w tekście uzasadnienia, dlaczego przyjęto ten model walidacji. Dziwi też przyjęte kryterium. Przypominam, że na s. 168 Autor napisał: *Uwaga! W przypadku walidacji efektu uczenia się, w sytuacji, gdy nie uczestniczy cała populacja, współczynnik trudności powinien spełniać bardziej rygorystyczny warunek: $b < -1{,}6$.* W badaniu, którego wyniki są analizowane, mieliśmy do czynienia z badaniem niepełnej populacji, więc dlaczego nie b<-1,6? Oba kryteria nie mają głębszego uzasadnienia, więc właściwie to nie ma większego znaczenia, ale warto było zachować spójność.

W tabeli 6.3. zabrakło informacji o oszacowaniach progów dla zadania 18. Jak można w przybliżeniu odczytać z wykresu 6.6. wartości te wynoszą odpowiednio -2,0 i 0,4. Zawarte w

tabeli 6.3. globalne oszacowanie b dla tego zadania wynosi -0,9. Ale w zadaniu *partial credit,* zgodnie z deklaracjami Autora w opisie dyskretnego modelu walidacji, powinno nas interesować pełne wykonanie zadania, a nie częściowe. Czyli nie globalny parametr b, a próg miedzy kodem 1 i 2, czyli 0,4. Czyli dla zadania 18. walidacja jest problematyczna.

S. 210. W podsumowaniu walidacji na podstawie analizy wyników egzaminu próbnego E8 czytamy:

Uczeń:

- *analizuje średnią,*
- *szacuje wartość kąta,*
- *oblicza pole powierzchni figury w sytuacjach praktycznych,*
- *konstruuje figury w symetrii osiowej i środkowej,*
- *analizuje sytuację problemową.*

Znów wracamy do problemu, czego dotyczy walidacja efektów uczenia się: jednostki czy populacji. Co to za magiczny „uczeń" analizuje średnią? Parametr b mówi nam o trudności danego zadania, czyli jaki poziom umiejętności na skali *theta* musi osiągnąć uczeń, by mieć 50% szansy na jego wykonanie. Czyli w zadaniu 2. uczeń o poziomie umiejętności na skali theta -0,88 ma 50% szansy na sukces. Jaki odsetek w badanej grupie uzyskał w całym teście wynik -0,88 lub wyższy/niższy, nie wiemy. Brak rozkładu wyników w badanej grupie. Nawet jakby rozkład był znany i tak nie wiadomo, co oznacza taka walidacja.

Inny problem to operacjonalizacja efektu kształcenia. Jak sam Autor dobitnie podkreślał, wybór zadania-walidatora nie jest ściśle wyznaczony treścią efektu uczenia się. Zatem inne zadanie, też dające się połączyć z tym efektem uczenia się, może dać inny wynik walidacji. Szczególnie przy ogólnikowo sformułowanym efekcie uczenia się jak np.: *analizuje sytuację problemową.*

S. 213. W podrozdziale 6.3. *Badanie uczniów: czwarty poziom Polskiej Ramy Kwalifikacji* znajduje się tabela 6.5. przedstawiająca charakterystykę próby ze względu na pochodzenie uczniów (wielkość miejscowości) na tle danych populacyjnych (analogiczna tabela znajduje się w podrozdziale 6.2). Tego typu analizy robi się w celu sprawdzenia, czy ze względu na uwzględnioną cechę próba jest podobna do populacji. W wypadku prób nielosowych i tak trudno mówić o reprezentatywności, ale można takie analizy traktować jako namiastkę. Czego dowiadujemy się z tabeli? Po pierwsze, wbrew nazwie kolumny, nie jest to

pochodzenie uczniów, a lokalizacja szkoły. Tak jest przynajmniej w danych populacyjnych CKE. Jakie informacje zbierano w badania, nie wiadomo. Po drugie, z tabeli dowiadujemy się, że struktura próby ze względu na cechę uwzględnioną w tabeli, istotnie różni się od składu populacji. Największa różnica dla miast powyżej 100 tysięcy mieszkańców wnosi -9,2 pp. Czyli ta namiastka badania reprezentatywności próby wskazuje, że niedoreprezentowani są maturzyści z dużych miast. Ponieważ wyniki egzaminów maturalnych na poziomie podstawowym są powiązane z lokalizacją szkoły, to uzyskane wyniki są obciążone (bardziej prawdopodobna negatywna walidacja).

S. 216 – 217. We fragmencie *Porównanie zadań generatorowych w próbnej maturze z zadaniami zamkniętymi na maturze w 2022 roku* podjęto bardzo ciekawy temat wpływu formatu pytania na jego wartość informacyjną (dyskryminację, pseduzgadywanie). W badaniu użyto tzw. zadań generatorowych, czyli wersji zadań tworzonych przez podstawianie różnych wartości/elementów tak, by podstawowa trudność (rozumiana potocznie) nie zmieniała się. Odpowiedź w nich ma charakter otwarty. Zestawiono charakterystyki psychometryczne tych zadań z analogicznymi zadaniami zamkniętymi z prawdziwej matury. Autor użył modelu Birnbauma z rodziny IRT. Jest to model 2PL (choć nie mam pewności, bo brak matematycznego opisu zastosowanego modelu). W modelu tym dla każdego zadania szacuje się parametr trudności i dyskryminacji (albo pseudozgadywania; parametr dyskryminacji jest powiązany pseudozgadywaniem - oszacowanie dolnej asymptoty krzywej logistycznej)ą Jednak do takich analiz jest dedykowany model 3PL. Wykres 6.14. sugeruje, że analizy wykazały, że dla zadań wariantowych parametr pseudozgadywania jest znacząco niższy niż dla analogicznych zamkniętych zadań maturalnych. Brak niestety miar dopasowania i analizy, jak zmienia się trudność zadań generatorowych przy różnych podstawieniach. Mimo tych niedostatków wyniki są bardzo ciekawe, choć wnioski za daleko idące: *Na podstawie tego badania można postulować, aby na maturze z matematyki odejść od wszystkich zadań zamkniętych.* Wniosek z wyników powinien być ostrożniejszy: w badaniach pilotażowych należy stosować model 3PL i zadania o wysokim parametrze psedozgadywania eliminować z bazy zadań.

S. 226. i dalsze. Jeden z opracowanych przez Habilitanta systemów walidacji efektów uczenia się poddano ewaluacji poprzez oceny studentów. W badaniu wzięło udział 38 z 90 studentów trzech grup, w których wprowadzono elektroniczny system walidacji efektów

uczenia się. Krótka ankieta ewaluacyjna wykazała, że ci, którzy ją wypełnili, w zdecydowanej większości pozytywnie ocenili system. Czytamy, że: *To pozwoliło na pełne wdrożenie Systemu walidacji efektów uczenia się w kolejnym roku akademickim.* Czyli na podstawie zadeklarowanej, pozytywnej oceny ok. 30 studentów wdrożono dużą innowację dydaktyczną. Na dodatek nie wiemy, jak była oceniana dotychczasowa metoda egzaminowania, nie znamy też kosztów tej innowacji, co jest kluczowe dla analizy *cost effectiveness*.

S. 228. i dalsze. Do weryfikacji hipotezy *Skonstruowano model efektu uczenia się* wykorzystano wyniki ankiety, w której wzięło udział 122 studentów (ze 140 uczestniczących w zajęciach w 5 wylosowanych grupach studenckich). Grupy były bardzo różnorodne, również niestacjonarne. Pierwsze wybrane do weryfikacji pytanie brzmiało: *Sposób przeprowadzenia egzaminu (zaliczenia) z algebry z geometrią analityczną lub analizy matematycznej przy komputerze sprawia, że chętniej przystępuję do uczenia się.* Odpowiadających podzielona na dwie grupy: tych, którzy zgodzili, się z tym stwierdzeniem (w jakim stopniu, nie wiemy) oraz tych, którzy się nie zgodzili. Liczebności tych grup to 122 *vs* 1. Następnie Autor opisuje wyniki testu statystycznego, który ma zweryfikować istotność statystyczną różnicy między odsetkiem „za" i „przeciw". To dość kuriozalna analiza. Zamiast napisać, że np. *Zaobserwowana różnica jest w oczywisty sposób istotna statystycznie (p<0,01)* znajdujemy szczegółowy opis całej procedury rodem z podręcznika statystyki. Jeżeli to miało dodać tekstowi naukowej powagi, to efekt jest odwrotny do zamierzonego. Następnie znajduje się w tekście funkcjonalnie tożsama analiza tylko z wykorzystaniem testu różnicy między medianami (skala porządkowa). I znów podręcznikowy opis na pół strony. To „pozwoliło" treść pozycji z ankiety przytoczyć *in extenso* pięć razy na przestrzeni niespełna strony. Lepiej byłoby zastanowić się na tym, na ile te deklarowane opinie są reprezentatywne. Przypadkowa heterogeniczność grupy (bynajmniej nie będąca efektem doboru warstwowego) każe wątpić, czy uzyskany wynik można uogólnić na jakąkolwiek populację. A jak Autor nie miał ambicji uogólniania, to po co test istotności? Te uwagi - niejako techniczne - nie są jednak najważniejsze. Kluczowe są dwie kwestie: (1) jaki sens respondenci nadawali zwrotowi *chętniej przystępuję do uczenia się* oraz (2) w jakim sensie odpowiedź na to pytanie, pozwala zweryfikować hipotezę, że *Skonstruowano model efektu uczenia się.* W dalszej części tekstu Autor podobną statystyczną zabawę przeprowadza dla

pozostałych pozycji ankiety: *Przeprowadzenie egzaminu (zaliczenia) z algebry z geometrią analityczną lub analizy matematycznej z wykorzystaniem komputera sprawia, że jest on mniej stresujący niż w tradycyjnej formie* (120 *vs* 3), *Dobrym pomysłem jest brak konieczności powtórnego rozwiązywania zadań z algebry z geometrią analityczną lub analizy matematycznej z wykorzystaniem komputera, które zostały poprawnie zweryfikowane we wcześniejszych terminach* (120 *vs* 3), *Forma egzaminu (zaliczenia) z wykorzystaniem komputera może przyczynić się do zmniejszenia zjawiska korepetycji* (113 *vs* 10), *Wcześniejsza znajomość zagadnień (efektów uczenia się), które będą oceniane na egzaminie (zaliczeniu) z algebry z geometrią analityczną lub analizy matematycznej, pomogła mi w uczeniu się* (122 *vs* 1), *Możliwość przećwiczenia przykładowych zadań egzaminacyjnych z wykorzystaniem komputera w trakcie semestru (przed przystąpieniem do zaliczenia/egzaminu) daje poczucia bezpieczeństwa, ponieważ zwiększa możliwość uzyskania pozytywnego wyniku* (122 *vs* 1), *Zastosowanie formuły egzaminu z wykorzystaniem komputera skraca czas przygotowania się do egzaminu (zaliczenia), ponieważ dokładnie wiadomo, czego należy oczekiwać* (120 *vs* 3).

Kolejne analizowane pytanie miało inną formę i dotyczyło tylko podgrupy n=89. Czytamy co następuje: *W badaniu wykorzystano ankiety studentów z grupy AB, N=89. Okazało się, że średnia ocena zgodności zadań rozwiązywanych na egzaminie (zaliczeniu) z wykorzystaniem komputera z kursu algebry z geometrią analityczną lub analizy matematycznej była wyższa od średniej oceny zgodności zadań rozwiązywanych na egzaminie (zaliczeniu) prowadzonym w tradycyjnej formie (nie przy komputerze) z kursu algebry z geometrią analityczną lub analizy matematycznej.* O jaką zgodność chodzi? Zgodność z czym? Z sylabusem? W dalszej części tekstu analizuje się rozkłady odpowiedzi dla kolejnych 6 pozycji ankiety sprawdzając, czy odpowiedzi korzystne dla innowacji przeważają nad niekorzystnymi. Do każdej z tych analiz mają zastosowanie krytyczne uwagi sformułowane w stosunku do pierwszej rozpatrywanej pozycji ankiety.

Dodatkowo chciałbym zwrócić uwagę, że akceptacja dla opinii *Dobrym pomysłem jest brak konieczności powtórnego rozwiązywania zadań z algebry z geometrią analityczną lub analizy matematycznej z wykorzystaniem komputera, które zostały poprawnie zweryfikowane we wcześniejszych terminach* jest problematycznym kryterium oceny systemu. Przecież niska trwałość efektów uczenia jest jedną z kluczowych bolączek szkolnej edukacji.

S. 240 - 248. Po analizie pytań zamkniętych na dziewięciu stronach ciągnie się analiza odpowiedzi na pytania otwarte. Monotonnie, bez zastosowania metod znanych z analizy danych jakościowych, Autor przytacza wpisy respondentów dzieląc je zwykle na dwie kategorie: pozytywne *vs* negatywne.

S. 248. Podrozdział *Analiza wyników i wnioski: Badanie części 3* to tylko niespełna 2 strony. Bardzo lapidarny opis badania. Zastosowano zdefiniowany wcześniej przez Autora model klasyczny walidacji. W podsumowaniu czytamy: *Badana grupa studentów D nie spełniła klasycznej walidacji dwóch efektów uczenia się: interpretuje definicję granicy ciągu i wyznacza długość krzywej za pomocą całki (rozwiązywalność 90% najlepszych studentów jest niższa niż 0,90).* Czy gdyby to były wyniki na podstawie danych z matury z matematyki na poziomie podstawowym, to należałoby odmówić wydania dyplomu całej grupie niezależnie od wyników indywidualnych? Cały czas powtarza się problem, kto jest podmiotem walidacji, grupa czy jednostka?

S. 249. Podrozdział 6.4. poświęcony jest taksonomii Blooma. Jak pisze Autor, poszukiwano odpowiedzi na pytanie badawcze *czy da się przypisać jednoznacznie kategorię Blooma do wybranego efektu uczenia się?* By poszukać odpowiedzi na to pytanie, dano do oceny 103 nauczycielom matematyki (członkowie SNM; sędziowie kompetentni?), dziewięć efektów uczenia się z prośba o przypisanie ich do kategorii zmodyfikowanej taksonomii Blooma. Efekty były sformułowane dość ogólnie, np. *Układa zadania i je rozwiązuje*. Nauczyciele mogli wskazać więcej niż jedną kategorię, nawet instrukcja ich do tego zachęcała. Jeżeli pytanie badawcze dotyczy jednoznaczności połączeń *efekt uczenia się -> kategoria Blooma*, to dlaczego instrukcja wręcz zachęcała do złamania tego założenia? Niektóre efekty uczenia się były sformułowane z kolei sugerująco. Na przykład dla efektu *Stosuje twierdzenie Pitagorasa* najczęściej przypisywano kategorię *Zastosowanie*. Czy zadziałało słowo *Stosuje*? Prawdopodobnie.

Kluczowe dla wartości tej analizy jest jednak to, czy tak ogólne sformułowanie efektów uczenia się w ogóle umożliwia jednoznaczne przypisanie i czy wiedza sędziów kompetentnych była wystarczająca. Nie ma informacji, czy uczestnicy badania byli wcześniej przeszkoleni w zakresie taksonomii Blooma.

Co o uzyskanych w tym eksperymencie mówi statystyka? W podaje się wartość statystyki chi^2. W opisie tabeli 6.32. czytamy: *Pytanie 1: Analiza testem χ2 zgodności i analiza częstości kryteriów efektu uczenia jako najskuteczniej weryfikujących danych efekt*. Gdy zsumujemy liczby przypisań, to wychodzi 261, czyli ok. 2,5 przypisania na respondenta. Jakiego rozkładu przypisań należałoby oczekiwać, gdyby były one w pełni losowe? Najprostsze założenie, to rozkład równomierny i taki przyjął Autor. Tabela zawiera wartości obserwowane i oczekiwane oraz ich różnicę, czyli reszty. Tylko dlaczego wartości reszt nie są różnicami wartości obserwowanych i oczekiwanych? Zastosowano jakiś bardziej złożony model, jaki? Wątpliwości budzi też model statystyczny, który zakłada niezależność przypisań, a ten warunek nie jest spełniony. Przypisania to klasyczna zmienna *multiple response* i wymaga innych modeli analizy (przypisania „zagnieżdżone" w osobach). Swoją drogą ciekawa byłaby analiza, jakie zestawy kategorii Blooma były przypisywane do poszczególnych efektów uczenia się.

S. 251. Do analizy taksonomii Blooma Autor zastosował wywiedzioną z teorii zbiorów rozmytych metodę „wektorową". Na podstawie przypisań nauczycieli i przyjętej, dość skomplikowanej metody, dla każdego z efektów uczenia się wyznaczył „wektor taksonomii Blooma". Na przykład dla *Stosuje twierdzenie Pitagorasa* przybrał on następująca wartość:

[0,50; 0,53; 0,85; 0,15;0,00; 0,00]

Poniższy wektor został wyznaczony przeze mnie na podstawie danych zawartych w pracy. Zawiera frakcje wyborów na podstawie tabeli 6.33. i tabel w załączniku 5.

[0,36; 0,74; 0,88; 0,31;0,07; 0,11]

Jak widać uporządkowanie wartości jest identyczne. Różnice w wartościach biorą się z arbitralnych parametrów wyznaczania przedziałów liczebności. Na czym polega przewaga złożonego modelu stworzonego przez Autora nad tym prostym rozwiązaniem? Brak w pracy argumentów. Niezależnie od wątpliwości dotyczących metody szacowania wektora przynależności danego efektu uczenia się do kategorii Blooma, pozostaje kwestia trafności danych źródłowych. Czy szacunki na podstawie wiedzy eksperckiej Autora (bez argumentacji) lub przypisania na podstawie zagregowanych danych pochodzących od nauczycieli bez weryfikacji ich wiedzy w zakresie taksonomii Blooma, można uznać za trafne dane?

Oddzielną kwestią jest stosowanie wnioskowania statystycznego. Na jaką populację będą uogólniane wyniki?

We wnioskach czytamy, że *Co do zasady nie można określić jednoznacznie kategorii Blooma dla danego efektu uczenia się.* Uzyskane wyniki są, zadaniem Autora, potwierdzeniem hipotezy HB5. W mojej ocenie określenie „co do zasady" jest empirycznie nieweryfikowalne. Można uzasadniać takie twierdzenia na drodze dedukcji, poprzez odwołanie się do ogólniejszego twierdzenia. Poza tym jeżeli w procedurze badawczej tak wiele zrobiono, by empiryczne wskazania były niejednoznaczne, to trudno uzyskany wynik traktować za potwierdzenie.

II. Badania podstawowe

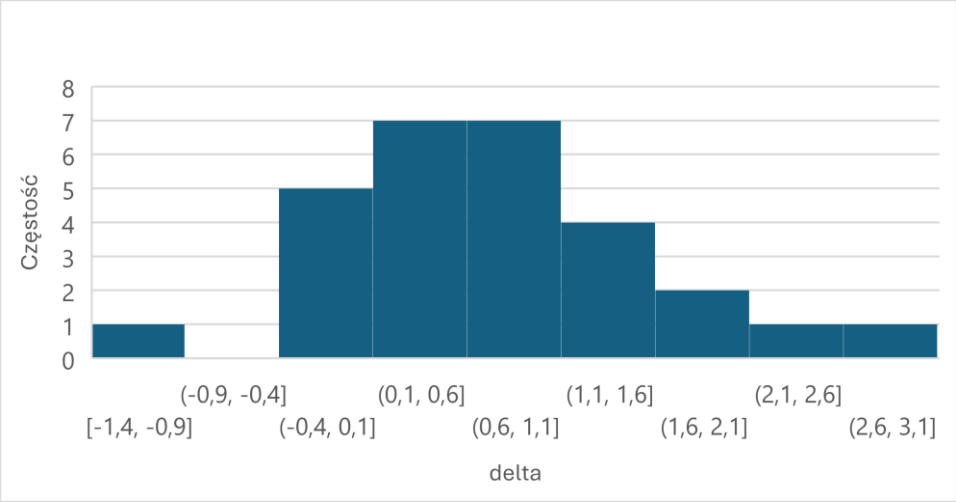
Poniżej prześledzimy fragmenty pracy poświęcone weryfikacji hipotez mających znamiona przynależności do badań podstawowych.

S. 254. Podrozdział 6.5. zaczyna się zdaniem: *Naukowcy starają się znaleźć zależności między różnymi zmiennymi w obszarze nauczania-uczenia się. Jedną z takich zmiennych jest na przykład walidacja efektów uczenia się, czas rozwiązania zadania lub liczba powrotów do rozwiązanego zadania.* W jakim sensie walidacja jest zmienną? Poza tym rzecz znacznie ważniejsza: brak przekonującego udokumentowania tej tezy i podstaw teoretycznych badania czasu wykonania i liczby podejść do zadania. Przegląd literatury jest skąpy a nieliczne przewołane artykuły zrelacjonowane są niedbale. Prawie połowę krótkiego akapitu, w którym zmieścił się przegląd badań, dotyczy związku wyników testów z ocenami szkolnymi. Brak pozycji literatur dotyczących liczby podejść/powrotów do zadania. Nie sposób rozwijać teorii bez rekapitulacji stanu wiedzy.

S. 254 - i dalsze. Komputerowa forma zadań dała możliwość łatwego zbierania danych o czasie wykonania poszczególnych zadań przez rozwiązujących. Autor odrzucił prosty model statystyczny, który pozwoliłby porównać rozkład czasu w grupach uczniów, którzy wykonali i nie wykonali poprawnie danego zadania. Czytamy: *W badaniach opisanych w literaturze przedmiotu, naukowcy wykorzystywali całą próbę badawczą. Zmiana autorskiej metodologii polega na tym, że dla każdego z zadań w próbnym egzaminie wyselekcjonowano tylko tych uczniów, którzy prawidłowo rozwiązali zadanie. Następnie uczniowie zostali podzieleni na tych, dla których wynik całościowy próbnego egzaminu był poniżej (grupa słabsza) albo*

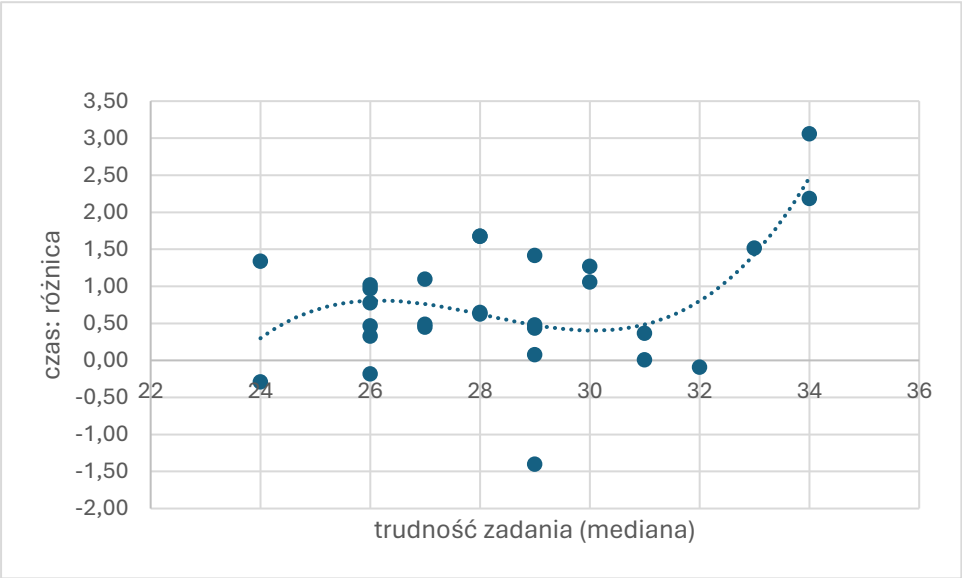
powyżej i równy medianie (grupa lepsza). Podobną zasadę zastosowano do liczby wejść do każdego z zadań (s. 254). Pomijając kwestie językowe i braku referencji należy zauważyć, że brak jest uzasadnienia przyjętego podejścia. W wypadku zadań kodowanych 0-1 rozwiązujący może opuścić zadanie (nie wiemy, czy była taka możliwość, załóżmy, że tak), rozwiązać niepoprawnie i rozwiązać poprawnie. Gdyby na podstawie danych dało się dla każdego zadania wyróżnić te trzy sytuacje, to analiza rozkładów czasu „pozostawania" w zadaniu pozwoliłaby odpowiedzieć na pytanie o związek czasu „podejścia" do zadania z jego rozwiązaniem. Oczywiście do analizy można by wprowadzać dodatkowe zmienne pełniące rolę moderatorów/mediatorów, np. ogólny wynik w teście (ew. z wykluczeniem danego zadania), trudność zadania, płeć itp. Dlaczego takie podejście zostało odrzucone? Nie wiadomo. Przyjęty model analizy prowadzi do redukcji liczebności próby i zróżnicowania tej liczebności w zależności od łatwości zadania (czyli różnej mocy statystycznej analiz dla różnych zadań/poziomów wykonania zadań) oraz do drastycznej redukcji wariancji zmiennej *ogólny wynik w teście* (sprowadzenie do zmiennej dychotomicznej). Co zyskiwano? Nie widzę przewag przyjętego sposobu analizy. Do takich analiz najbardziej poręczny byłby model regresji wielomianowej, logistycznej. Wyróżnione kategorie wykonania zadania byłyby wielomianową zmienną wyjaśnianą, czas zmienną wyjaśniającą. Do takiego modelu można by wprowadzać dodatkowe zmienne, np. ogólny wynik w teście. Można by testować też ciekawe interakcje np. czas x płeć.

S. 261. Wiele stron analiz danych z próbnej matury prezentowanych zarówno w formie tabelarycznej jak i graficznej (zdecydowanie nadmiarowa w stosunku do potrzeb analizy) prowadzi Autora do wniosku, że *Zmieniona przez autora metodologia badania, polegająca na wybraniu do badania tylko tej grupy uczniów, którzy prawidłowo rozwiązali zadanie, pozwoliła na sformułowanie wniosku: uczniowie lepsi potrzebują mniej czasu na poprawne rozwiązanie zadania.* To wniosek dość banalny, na dodatek nie w pełni wykorzystuje wyniki badania. Na przykład, używając danych tylko z tabel 6.38. (s. 258) można odkryć ciekawą zależność między trudnością zadania a różnicą między medianą czasu dla uczniów lepszych i słabszych (podział po medianie ogólnego wyniku w teście). Warto najpierw przyjrzeć się rozkładowi różnic w czasie dla zadań jednopunktowych:



Oszacowanie średniej dla różnicy wynosi 0,77 minuty (se=0,16). Czyli wniosek *Do prawidłowego rozwiązania zadań jednopunktowych prawie w każdym przypadku uczniowie, którzy osiągają lepsze wyniki, potrzebują mniej czasu na rozwiązanie zadania* wydaje się być trafny, ale różnica średnich nie jest duża: słabsi wykonują poprawnie zadania o średnio 46 sekund dłużej. Dla porządku należałoby oszacować ten wskaźnik różnicy przedziałowo. Jeżeli błąd standardowy wynosi 0,16, to 95% przedział ufności dla średniej będzie się rozciągał od 0,46 do 1,08. Czyli z dużą pewnością można potwierdzić zaobserwowaną korelację. Jednak wyniki analogicznych analiz dla zadań wielopunktowych są niekonkluzywne.

Ciekawe od czego w wypadku zadań 0-1 zależy różnica w czasach poprawnego wykonania zadania lepszych i słabszych matematyków? Sprawdźmy, czy wariancja różnicy da się wyjaśnić trudnością zadania (warunkowa mediana punktów w całym teście).



R^2 dla funkcji sześciennej wynosi 0,38.

Widzimy ciekawą, znaczącą statystycznie zależność. Jest ona znacznie bardziej złożona, niż korelacja wykryta przez Autora. Z pewnością w danych kryje się jeszcze wiele ciekawych odkryć.

S. 261. Równie obszerna analiza dla wyników E8 prowadzi Autora do wniosku: *Uczniowie lepsi potrzebują mniej czasu na rozwiązanie zadania, jednak różnice w stosunku do grupy słabszej są niewielkie*. Czyli analogicznie jak dla danych maturalnych. Ale w tym wypadku można mieć jeszcze dalej posunięte wątpliwości, co do trafności wniosku. Z tabeli 6.34. (zadania jednopunktowe) można odczytać, że dla 7 z 15 zadań różnice w średnich czasu poprawnego wykonania nie różnią się istotnie od zera. W trzech zadaniach średni czas jest dłuży w grupie słabszej, a w czterech - w grupie lepszej. W wypadku zadań dwupunktowych wykonanych w pełni dla dwóch zadań czas jest istotnie statystycznie dłuższy w grupie wyższych wyników, w jednym - w grupie słabszych. W zadaniach trzypunktowych (nie wiadomo przy jakim warunku) jeden wynik jest zgodny z wnioskiem, jeden nie, trzecia różnica bliska zeru.

S. 262. Analiza weryfikująca hipotezę dotyczącą liczby „wejść" do zadania nie przyniosła żadnych spektakularnych wyników. Autor wyciąga wnioski co do kierunku różnic, ale ogląd tabel 2.42.- 2.49 wykazuje, że różnica między medianami liczby wejść nie przekracza 1. W jednej tabeli (6.43) mediany są podawane z dokładnością do dwóch miejsc po przecinku i tu obserwujemy dla jednego zadania różnicę trochę większą. Dlaczego w tej tabeli podaje się mediany w innym formacie, nie wiadomo. Przeglądanie tabel pozwala też zauważyć, że w wielu przypadkach mediany mają tę samą wartość, a test statystyczny wykazuje istotność różnicy! Jest tak np. w tabeli 6.42, 6.46, 6.48. Trudno zatem zrozumieć wnioski: *uczniowie lepsi częściej wracają do ponownego rozwiązania zadania, aby zweryfikować jeszcze raz rozwiązanie lub aby je w pełni rozwiązać*. *Liczba powrotów do rozwiązania zadania, które mierzy określony efekt uczenia się, różnicuje uczniów lepszych i słabszych*. Gdyby powyższe wnioski przyjąć za dobrą monetę, to i tak nie bardzo wiadomo, jak przyczyniają się do rozwoju teorii. Związku z rozwiązaniem problemu praktycznego też nie widać.

S. 266. Podrozdział 6.6. nosi tytuł *Badanie mocnych i słabych stron uczniów w odniesieniu do walidacji efektów uczenia się.* W pierwszym paragrafie czytamy jednak*: Celem badania*

było ustalenie związku między wyborem trudnych i łatwych treści w zakresie matematyki, wykorzystaniem matematyki w codziennym życiu a wynikiem próbnych egzaminów. Krótki akapit po tej deklaracji zawiera odnośniki do prac z następujących zakresów: o wykorzystaniu poznanej w szkole matematyki w codziennym życiu, o pojęciu funkcji i związanych z nim trudnościami w uczeniu się oraz pomiarze umiejętności rozwiązywania problemów przy użyciu pytań matematycznych opartych na problemach. Nie wiadomo, jaki te artykuły mają związek z treścią podrozdziału.

S. 266. Przywołany powyżej podrozdział 6.6. przedstawia wyniki badania ankietowego przeprowadzonego przy okazji opisanych wcześniej pomiarów. W opisie metody badania czytamy: *Pytania 1 i 2 miały charakter zamknięty i pięć możliwych odpowiedzi. Po usunięciu odpowiedzi pustych (brak odpowiedzi na kwestionariusz ankiety), dane zostały pogrupowane.* Jak można usuwać dane z bazy? „Odpowiedzi puste" powinny być zakodowane jako braki danych. Baza danych powinna obejmować całą próbę założoną. Ta zasada pozwala analizować stopę realizacji badania i losowość „wypadnięcia" z próby i braków danych. Brak takich analiz nie pozwala oszacować ewentualnego obciążenia wyników.

S. 267-270. W analizach przedstawionych w podrozdziale 6.6.1. jako zmienne wyjaśniające wykorzystano wybrane pytania z ankiety. Dotyczyły one dodatkowych zajęć z matematyki i ocen z tego przedmiotu (pytania zamknięte) oraz treści matematycznych ocenianych jako przydatne w życiu i w pracy, trudnych zagadnień, ale uważanych za ważne oraz zagadnień, które są w ocenie respondenta jego mocną stroną (4 pytania otwarte).

W wypadku pytań zamkniętych przeprowadzono analizy, które Autor opisuje następująco: *Każde pytanie było rozpatrywane osobno. Następnie z ogólnej liczby punktów uzyskanej w teście policzone zostały mediany, średnie oraz liczebności próbek dla każdej z 5 możliwych odpowiedzi oraz dla całej próbki. Aby wykazać statystyczną różnicę między średnimi* (?) *w grupach, został wykonany test Manna-Whitneya dla każdych dwóch par grup.* Można się domyślać, że zmienną wyjaśnianą były wyniki próbnych egzaminów. Analizy istotności różnic między medianami wykonano dla wybranych par odpowiedzi (pominięto grupę: *Nie mam zdania*). Wyniki przedstawiono w tabeli 6.51. i 6.52. Dla E8 test Manna- Witneya wykazał istotność różnic między prawie wszystkimi podgrupami, ale znamienny jest brak różnicy

między skrajnymi grupami: *Nie chodziłem vs chodziłem cykliczne*. Zaobserwowane różnice były niewielkie i wynosiły jeden punkt.

Dla wyników próbnej matury wybrano inny schemat porównań parami, mianowicie w drugiej parze grupę nie chodzących na dodatkowe zajęcia zestawiono z pozostałymi. Wyniki dla trzech porównań opisano jako nieistotne (przyjąć H_0), jedynie porównanie *nieuczęszczających* z tymi, którzy *chcieliby uczęszczać*, wykazało wyższy wynik tych pierwszych. Niezrozumiałe jest uznanie różnic w trzecim i czwartym porównaniu jako nieistotnych, choć $p<0,05$. Na koniec we wnioskach czytamy: *W grupie maturzystów odnotowano różnicę między tymi, którzy chcieli chodzić na dodatkowe zajęcia, a tymi, którzy chodzili cyklicznie lub sporadycznie na dodatkowe zajęcia.* Czyli różnica, która została uznana za nieistotną została „wybita" we wnioskach, jako jedyna znacząca.

Uzyskane wyniki nie zostały odniesione do – przebogatej przecież – literatury przedmiotu.

S. 271-279. Na dziewięciu kolejnych stronach znajdujemy rozwlekłe analizy dla czterech kolejnych pytań ankiety (odpowiedzi na pytania otwarte zostały skategoryzowane). Niewiele się z nich dowiadujemy. Wnioski w rodzaju *Możemy stwierdzić, że oceny na świadectwie i wyniki z próbnych egzaminów są ze sobą kompatybilne* ani na jotę nie wzbogacają naszej wiedzy i na dodatek są zbyt mocnymi uogólnieniami. Zgodnie z SJP słowo *kompatybilny* w drugim znaczeniu oznacza: *odpowiadający czemuś lub przystosowany do czegoś pod każdym względem.* To oznacza, że korelacja między ocenami a wynikiem testu powinna być bliska 1. A tak przecież nie jest. Na dodatek obserwujemy znaczące różnice w tym zakresie ze względu na poziom nauczania, szkołę i nauczyciela.

Ciekawsze są wyniki dotyczące oceny przydatności różnych treści matematycznych w życiu z osiągnięciami. Wskazywanie przez uczniów SP i słabszych uczniów na bardziej podstawowe treści jako przydatne, to ciekawy trop. Jednak wszystkie te ustalenia i tak nie poszerzają naszej wiedzy, bo bez ich odniesienia do teorii i stanu badań ich wartość poznawcza jest bliska zeru. Na przykład badania nad dodatkowymi zajęciami/korepetycjami i ich wpływem na osiągnięcia, również w polskich realiach szkolnych, to setki studiów, wiele opracowań teoretycznych. Postęp w badaniach podstawowych wymaga nawiązywania do ustaleń poprzedników. Związek tych zagadnień z centralnym problemem praktycznym rozwiązywanym w tej pracy jest wątły. Również związek z tytułem podrozdziału jest trudno

dostrzegalny. W podsumowaniu czytamy: *Analizując odpowiedzi dotyczące mocnych stron wiedzy i umiejętności matematycznych uczniów, można zauważyć, że ósmoklasiści deklarujący chęć poznania w lepszym stopniu geometrii lub prawdopodobieństwa i statystyki osiągają lepsze wyniki niż uczniowie deklarujący pozostałe kategorie zagadnień. Z kolei uczniowie wskazujący arytmetykę jako swoją najmocniejszą stronę uzyskują istotnie niższe rezultaty niż uczniowie, którzy za swoje mocne strony uważają inne grupy pojęć, np. geometrię.* Trudno uznać to za wyczerpanie tytułowego zagadnienia.

S. 280 – 296. Rozdział *Badanie z wykorzystaniem narzędzia Google Trends* jest poświęcony poszukiwaniu odpowiedzi na pytanie *Czy i w jaki sposób pandemia COVID-19 zaktywizowała uczniów do poszukiwania treści matematycznych w Internecie?* W analizie wykorzystano ciekawą metodę analizy trendów czasowych. Metoda ta zmienną obserwowaną w planie podłużnym rozkłada na trzy składowe: trend, cykliczne (sezonowe) wahania i błąd (efekt losowy). Metoda ta zastosowana dla danych z *Google Trends* dla najpopularniejszych haseł związanych z matematyką dla Polski (hasła po polsku) i na świecie (hasła po angielsku; czy to faktycznie świat?) wykazała znaczący trend wzrostowy liczby ich wyszukiwań w okresie pandemii Covid-19 i spadek po jej wygaszeniu. Nie jest to jednak spadek do poziomu sprzed pandemii. Co ciekawe, spadek ten jest silniejszy w Polsce. Gdyby skorelować nasilenie trendów z liczbą dni zdalnej nauki w poszczególnych krajach, to interpretacja trendu byłaby pewniejsza. Niestety Autor nie wchodzi w pogłębione analizy i zadawala się potwierdzeniem dość oczywistej hipotezy. Brak odwołań do badań w tym zakresie czyni uzyskane wyniki ciekawostkami, a nie wiedzą naukową.

III. Błędy i usterki

Praca ma sporo usterek i błędów. Część z nich wymieniam poniżej.

S. 33. Jean Piaget stał się Johnem.

S. 39. Przy okazji omawiania znaczenia dla osiągnięć akademickich samodyscypliny Autor pisze:

Z badań Angeli L. Duckworth i Martina E.P. Seligmana (2005) wynika, że główną przyczyną, dla której uczniowie nie wykorzystują swojego potencjału intelektualnego, jest brak

samodyscypliny. W tym badaniu okazało się także, że samodyscyplina miała dwukrotnie większą wariancję niż testy IQ, oceny końcowe i frekwencja w szkole.

W przywoływanym artykule czytamy: *Samodyscyplina mierzona jesienią odpowiadała za ponad dwukrotnie większą wariancję ocen końcowych niż IQ …. No comments.*

S. 122. Szwankuje etyka języka: *wyniki badań przeprowadzonych na nauczycielach*

S. 136-137 Przy okazji prezentacji różnych wskaźników psychometrycznych Autor pisze: „Wskaźnik *dyskryminacji testu* obliczamy jako średnią arytmetyczną wskaźników trudności poszczególnych zadań." Tak zdefiniowana dyskryminacja testu to termin nieznany psychometrii. Można mówić o testach służących dyskryminacji, ale w tym wypadku raczej przejęło się mówić o testach różnicujących. Jaki związek miałaby taka miara z różnicowaniem badanej grupy ze względu na poziom osiągnięć? Można mówić o mocy różnicującej zadania, czy w obrębie IRT o parametrze dyskryminacji pozycji testowej.

S. 138. Dwukrotny błąd w nazwisku: *współczynnik α Crombacha*.

S. 167. Niezrozumiały akapit (nad „Walidacja probabilistyczna")

S. 191. Niepodpisany wykres, sam wykres nie pasuje do przykładu.

S. 255 i dalsze. A tabelach prezentujących wyniki analiz statystycznych (test różnicy median) znajduje się kolumna *p wartość dla hipotezy kierunkowej*. Ten sposób szacowania jest stosowany, gdy mamy podstawy do sformułowania hipotezy co do kierunku różnicy. Dla poszczególnych zadań kierunek tej hipotezy jest inny. Na jakiej podstawie ustalano kierunek? Przyjęcie *ad hoc* hipotezy kierunkowej nie jest poprawną procedurą.

IV. Struktura pracy

Struktura pracy jest mało przemyślana. Mnóstwo miejsca poświęcono na kwestie, z punktu wiedziana problemu badawczego, drugorzędne lub zbędne. Po co we wstępnych rozdziałach bardzo obszernie przedstawiono proces boloński, ERK, KRK, PRK, RS, ZSK? Poświęcono tym czysto odtwórczym opisom aż 27 stron. Kolejne 49 stron to przedstawienie koncepcji efektów uczenia się, głównie w zakresie kompetencji społecznych, które nie są przedmiotem badań w omawianej pracy. Po co na ponad 8 stronach opisany został CEFR, a na kolejnych czterech - kształcenie językowe w Polsce? Na s. 143 i dalszych znajdziemy opis

prób wdrożenia w Polsce testowania on-line i e-oceniania. Tematy te słabo wiążą się z tematem pracy, a już z pewnością niepotrzebnie przedstawia się szczegóły organizacyjne, tym bardziej, że nie analizuje się ich z punktu widzenia jakości testowania. Treści przy omawianiu typów zadań testowych powtarzają się. Warto było skorzystać z bardziej autorytatywnych źródeł. Opis kategorii zadań generatorowych jest zupełnie zbytecznie opisany za pomocą matematycznych formalizmów, na dodatek ten typ zadań nie jest innowacją (na s. 153 czytamy: *Dyskutowane poniżej zadania generatorowe i zadania sterowane strumieniem danych stanowią zagadnienie nowe/innowacyjne/aktualne w konstruowanych testach egzaminacyjnych online*). Prace nad generatorami zadań testowych trwają przynajmniej od 15 lat (np. w ETS).

Podsumowanie

Jak zapowiedziałem we wstępie, przedstawioną do oceny pracę recenzowałem dwutorowo: wątek badania stosowanego i oddzielnie elementy badań podstawowych. Pora na ocenę zbiorczą.

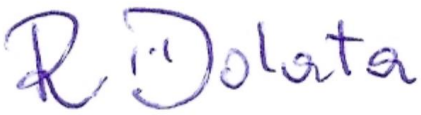
Czy praca badawcza dr Jacka Stańdo pozwoliła rozwiązać ważny problem praktyczny? W mojej ocenie nie. Samo sformułowanie problemu jest dalekie od precyzji, a zasoby teoretyczne nauk społecznych ze szczególnym uwzględnieniem pedagogiki nie zostały wykorzystane właściwie. Badania prowadzące do sprawdzenia skuteczności proponowanego rozwiązania wielokrotnie łamią zasady metodologii naukowej.

Oceniając realizację wątków badań podstawowych muszę z całą mocą podkreślić, że nie spełniają one podstawowego warunku wyjściowego, czyli przedstawienia stanu badań danego zagadnienia. Nie dziwi zatem, że hipotezy mają bardzo niski poziom zaawansowania teoretycznego. Do spełnienia kluczowego kryterium, czyli przyczynienia się do rozwój teorii danego zjawiska lub zaproponowanie nowej, lepszej teorii, Habilitant nawet się nie zbliżył. Brak zakorzenienia teoretycznego i liczne błędy metodologiczne nie pozwalają pozytywnie ocenić przeprowadzonych w tym wątku pracy badań.

Przedstawiona do oceny praca nie wzbogaciła wiedzy w dziedzinie pedagogika.

Konkluzja

Bardzo niska ocena przedstawionej monografii, mimo znaczącego dorobku publikacyjnego (choć w dużej mierze spoza pedagogiki) i niezwykłej aktywności naukowej, nie pozwala mi pozytywnie zaopiniować wniosku dr Jacka Stańdo o nadanie stopnia doktora habilitowanego w dziedzinie nauk społecznych, w dyscyplinie *pedagogika*.



Warszawa, dnia 24.03.2024 roku