Dr hab. Roman Dolata
zam. 01-231 Warszawa, ul. Płocka 17 m. 99
adres e-mailowy: rdolata@uw.edu.pl

14.03.2021 r.

Recenzja rozprawy doktorskiej Szymona Borsicha

Zmiany zachodzące w zakresie wybranych podmiotowych czynników tworzących kontekst osiągnięć szkolnych uczniów klas IV-V i ich uwarunkowania

W recenzji zajmę się kolejno: wstępnym sformułowaniem problemu badawczego, przeglądem literatury, finalnym sformułowaniem problemu, metodologią jego rozwiązania, analizą wyników i językiem rozprawy.

Wstępne sformułowanie problemu badawczego

Na s. 5. i 6. znajduje się wstępny opis tematyki i charakterystyka badania. Czytamy:

„Czynnikami podmiotowymi, które uwzględniono w tej rozprawie, są poczucie bezradności na lekcjach języka polskiego i matematyki, poczucie zagrożenia stereotypem niskich zdolności na lekcjach języka polskiego i matematyki oraz styl wyjaśniania przyczyn własnych sukcesów i porażek w nauce. Spośród wielu czynników związanych ze środowiskiem rodzinnym wzięto pod uwagę status socjoekonomiczny rodziny, w której wychowuje się badany uczeń. Miarą osiągnięć szkolnych były oceny szkolne. (…) Celem poznawczym rozprawy jest diagnoza zmian zachodzących w zakresie wybranych czynników podmiotowych, tworzących kontekst osiągnięć szkolnych w toku nauki w klasie IV i V. (…) Zaplanowano ich trzykrotny pomiar – na początku i pod koniec klasy IV oraz na koniec roku szkolnego w V klasie – w ośmiu losowo wybranych zespołach klasowych z trzech grup szkół wyodrębnionych ze względu na zróżnicowanie wyników osiąganych w sprawdzianach na zakończenie klasy VI.”

By ten wstępny opis uzupełnić, należy dodać, że SES rodziny ucznia i jego płeć miały w badaniu pełnić funkcję po trosze czynnika wyjaśniającego wariancję charakterystyk psychologicznych i osiągnięć uczniów, ale głównie miały być moderatorem trendów w trzech kolejnych pomiarach.

Takie wstępne sformułowanie tematu pracy jest zadawalające, jednak pod warunkiem, że po przeglądzie literatury i syntezie stanu wiedzy zostanie sformułowany oryginalny problem badawczy.

Przegląd literatury

Przegląd literatury dotyczy głównie trzech kluczowych dla pracy konstruktów: zagrożenia stereotypem niskich zdolności, wyuczonej bezradności i stylu wyjaśniania przyczyn własnych sukcesów i porażek w nauce. W mniejszym stopniu przegląd zawiera problematykę SES, płci i osiągnięć szkolnych. Bibliografia zawiera około 200 pozycji.

W podrozdziale 1.1. Autor przygotowuje teoretyczny grunt przedstawiając charakterystykę etapu rozwoju, w którym znajdują się badani. Niestety podrozdział pozbawiony jest struktury, brak teoretycznej linii wywodu. Czytelnik napotyka dużo powtórzeń, literatura źródłowa jest skromna i zbyt często teorie i badania są omawiane z drugiej ręki. Rozdział autorstwa Karoliny Apelt i podręcznik Helen Bee urastają do rangi filarów wiedzy w zakresie psychologii rozwojowej. Brak wielu klasycznych i współczesnych prac dotyczących samooceny i związków samooceny z osiągnięciami szkolnymi.

Nadużywanie niektórych pozycji nie ogranicza się do tych dwóch prac. Książka recenzenta jest przywoływana kilkanaście razy (Dolata 2012 – z błędną datą, poprawna 2008) i to głównie jako źródło omówień z drugiej ręki. W tym wypadku mogłem też wyłapać problem polegający na zbyt obfitym, nieudokumentowanym czerpaniu z cytowanego tekstu, mianowicie na ss. 91. i 92. całe fragmenty są przytoczone z minimalnymi zmianami. Oczywiście czasami takie czerpanie z tekstu innych autorów może być usprawiedliwione, ale tylko gdy jest to podporządkowane analizowanemu problemowi badawczemu i /lub wynika z potrzeby krytyki naukowej. W tym wypadku nic takiego nie zachodzi. Przywoływane rozważania nic nie wnoszą do pracy.

W opisach teorii i badań znajdujemy czasami fragmenty wskazujące nie niepełne rozumienie omawianej materii. Na s. 10. w opisie teorii J. Piageta Doktorant pisze: „Natomiast posługiwanie się zasadą pojemności kategorii pojawia się w wieku 7-8 lat. Dzieci zaczynają rozumieć, że podkategorie mieszczą się w jednej większej, głównej kategorii, „na przykład jamniki w kategorii psów." Gdyby nie podany przykład, czytelnik nie miałby szans zrozumieć, o co chodzi. Czy „zasada pojemności kategorii" dotyczy pojęcia stałości objętości, czy też doskonalenia się umiejętności klasyfikacyjnych i rozumienia relacji zawierania się kategorii podrzędnej w nadrzędnej (inkluzja klas).

W przeglądzie literatury rzuca się w oczy zdecydowanie zbyt częste powoływanie na teksty, które zdecydowanie nie mogą być uznane za klasyczne dla omawianych zagadnień. Na przykład wspominany już w recenzji tekst Apelt (2015). *Wiek szkolny. Jak rozpoznać potencjał dziecka?* jest przywoływany ponad 20 razy i w większości wypadków wskazane byłoby użycie tekstów bardziej klasycznych.

Inny przykład pochodzi ze Wstępu. Doktorant definiując pojęcie „kontekst dydaktyczny" (s. 5) powołuje się pracę Krystyny Szmigel. Naprawdę, mimo wielkiego uznania dla prac dr Szmigel w zakresie rozwoju systemu egzaminacyjnego w Polsce, nie sposób przywoływanego tekstu traktować jako źródła standardów definicyjnych w tym zakresie.

Kolejne podrozdziały poświęcone są kluczowym dla rozprawy pojęciom. W ich teoretycznej charakterystyce brakuje czytelnej struktury i linii narracyjnej. Rzeczy ważne mieszają się z peryferyjnymi, nie widomo, co przygotowuje grunt do sformułowania problemów badawczych, a co jest ogólnym wprowadzeniem do danej problematyki.

W omówieniu badań czytelnik może też napotkać dość ważne niejasności. Np. na s. 29. znajdujemy taki opis badania nad zagrożeniem stereotypem: „Badanie to wykazało istotne statystycznie różnice ($p<0,03$) pomiędzy wynikami białych studentów a Afroamerykanami, ale w sytuacji, w której podkreślano diagnostyczną rolę testu w zakresie zdolności intelektualnych, interakcja czynników „rasa*warunek eksperymentu" okazała się nieistotna statystycznie ($p<0,19$). Wyniki badania wskazywały więc, że to nie rasa jest samoistnym czynnikiem warunkującym wystąpienie zagrożenia stereotypem. Z uwagi bowiem na fakt, że stereotyp odnosi się do zdolności intelektualnych, aby wywołać zagrożenie stereotypem konieczne jest także stworzenie szczególnych warunków, w których jednostki mają przekonanie, że to właśnie ich zdolności intelektualne są poddawane diagnozie." Jeżeli w badaniu podkreślano diagnostyczną rolę testu, to jak rozumiem warunek sformułowany w ostatnim zdaniu przytoczonego fragmentu został spełniony. Czyli jest to badanie wykazujące, że efekt zagrożenia stereotypem nie wstępuje. Takiej konstatacji w rozprawie nie ma.

Przytoczony fragment dobrze ilustruje też problem spójności narracji. W opisie badań Autor czasami podaje szczegóły metodologiczne i statystyczne, czasami nie? W powyższym fragmencie Doktorant podaje wartości p, ale w wielu innych opisach tego brak. Oczywiście jakieś kluczowe badania mogą być szczegółowo opisane, ale tym musi rządzić jakaś reguła. W recenzowanym doktoracie jest to całkiem losowe. Na stronie 40. znajdujemy na przykład opis badania Oswald i Harveya z 2000 z podaniem szczegółowych statystyk opisowych i wartości statystyk F. Czemu to służyło, nie wiadomo. Wygląda to na proste wklejenie wcześniej przygotowanej notatki z lektury.

W opisie badań zdarzają się też powtórzenia. Na przykład badania Bosson i in. z 2004 przywoływane są na s. 28 i 31.

Ważnym mankamentem przeglądu literatury jest omawianie teorii i badań z „drugiej ręki". Oczywiście, przy trudno dostępnym źródle, sporadycznie może się to zdarzać, ale w pracy Doktoranta to prawie reguła. W większości wypadków jest to wprost dokumentowane, ale to nie załatwia sprawy.

Z trochę innym problem mamy do czynienia w wypadku tabeli 1. przedstawiającej przegląd badań nad występowaniem zjawiska zagrożenia stereotypem wśród dzieci. Co prawda pod tabelą znajduje się źródło, jednak samo umieszczenie takiej tabeli w tekście sugeruje „zasługę" Autora rozprawy.

Jak już pisałem, opisy poszczególnych konstruktów nie mają wspólnej struktury ani linii narracyjnej. Na przykład przy opisie stylu atrybucji porażek i sukcesów szkolnych przykładowy kwestionariusz (s. 85) pojawia się jak „królik z kapelusza". Skąd pochodzą polskie *itemy* tego narzędzia? Czy to tłumaczenie Autora, czy pozycje z polskiej adaptacji kwestionariusza Seligmana?

W przeglądzie literatury Doktorant sporo miejsca poświęca problemom definicyjnym. Jest to ze wszech miar pożądany elemnt rozprawy doktorskiej, ale czasami wątpliwości budzą przyjęte konwencje. Na przykład fragment poświęcony definicjom wyuczonej bezradności brzmi tak:

„Wyuczona bezradność jest *poddaniem się, zaprzestaniem działania wynikającym z przekonania, że cokolwiek się zrobi, nie będzie to miało żadnego znaczenia* (Seligman, 1993, s. 32). Podobnie definiuje zjawisko Barbara Ciżkowicz (2009 s. 12) jako *niemożność realizacji własnych potrzeb poprzez własne zachowania* w aspekcie obiektywnym lub jako postrzeganie przez osobę braku związku między jej zachowaniem a jego przewidywanymi konsekwencjami w aspekcie subiektywnym, określając zjawisko mianem poczucia bezradności."

Staranna lektura tego fragmentu pozwala dostrzec niespójność przytaczanych podejść do wyuczonej bezradności, doktorant pisze natomiast, że definicje są podobne. Drugie podejście (przynajmniej na podstawie przywołanych fragmentów) wydaje się gubić clou konstruktu oryginalnego, czyli tego, że chodzi o bezradność **wyuczoną.**

Niestety przeglądu literatury nie wieńczy próba krytycznej syntezy stanu wiedzy w eksplorowanym temacie i brak identyfikacji obszarów poznawczej niepewności. To źle rokuje dla teoretycznej dojrzałości problemu badawczego.

Ostateczne sformułowanie problemu badawczego

Na s. 7. czytamy:

„Przeprowadzone badania nie pozwalają na generalizację wniosków, jednak wzbogacają wiedzę o czynnikach podmiotowych, które mogą utrudniać funkcjonowanie uczniom w środowisku szkolnym, a dzięki badaniom podłużnym możliwe będzie jednoczesne przeanalizowanie zmian ich poziomu w czasie w zależności od płci i statusu socjoekonomicznego rodziny ucznia (dalej: SES) oraz zmian zachodzących w szkolnych osiągnięciach badanych."

Jeżeli schemat badania, a w szczególności dobór próby, nie pozwalają na generalizację wyników z próby na jakąś szerszą populację, to nie można sensownie twierdzić, że wyniki takiego badania wzbogacą naszą wiedzę. Jedynie w badaniach stosowanych, ewaluacyjnych lub diagnostycznych, nie jest konieczna generalizacja. Faktycznie, gdy dowiemy się w badaniu ewaluacyjnym, czy działanie X w danej sytuacji było skuteczne, to rozwiążemy ważny problem praktyczny. Analogicznie, jeżeli w badaniu diagnostycznym poznamy przyczyny wystąpienia w danym przypadku jakiegoś niekorzystnego zjawiska X, to też przyczynimy się do rozwiązania problemu praktycznego. Jednak podjęta przez Doktoranta tematyka w żadnym razie nie prowadzi do sformułowania problemu badawczego o charakterze praktycznym. Wszystkie przeprowadzone analizy są charakterystyczne dla badań podstawowych. Zatem kwestia generalizacji wyników z próby na populację jest absolutnie kluczowa. I żadne zabiegi czysto werbalne tego nie zmienią.

Znajdujemy bowiem w pracy odwołania do podręcznika metodologii K. Konarzewskiego i deklarację, że opisywane badanie ma charakter praktyczny. Na s. 104. czytamy:

„Takie badania Krzysztof Konarzewski (2000, s. 13) określa jako badania praktyczne. Praktyczne, bo podejmowane *nie po to by stworzyć lub udoskonalić jakąś teorię, leczy by dostarczyć impulsów do rozwoju pewnej dziedziny praktyki społecznej*."

Natomiast na s. 100. Czytamy:

„Celem naukowym planowanych badań jest **wzbogacenie teorii edukacji** (pogrubienie RD), dzięki lepszemu poznaniu czynników mających znaczenie dla angażowania się uczniów w naukę, na które może mieć wpływ nauczyciel, takich jak: zagrożenie stereotypem i poczucie bezradności na lekcjach języka polskiego i matematyki, przyjmowanego przez nich stylu wyjaśniania przyczyn własnych porażek i sukcesów w nauce."

Całość pracy nie pozostawia wątpliwości: praca mieści się w kategorii badań podstawowych i powinna być oceniana przede wszystkim z perspektywy rozwoju teorii badanych zjawisk. Praktyczność prezentowanego badania można jedynie uzasadniać tym, że badana próba pełniła funkcję grupy

kontrolnej w eksperymencie pedagogicznym „Niebieskoocy" w naszej szkole – przezwyciężanie stereotypów drogą do tworzenia uczniom lepszych warunków rozwoju w klasie szkolnej". To oczywiście żartobliwa uwaga, ale przy okazji pragnę zwrócić uwagę, że pochodzenie danych wykorzystanych w badaniu zostało opisane w tekście zdecydowanie zbyt późno i zbyt zdawkowo.

Próba naginania klasyfikacji Konarzewskiego i przypisania opisywanego badania do kategorii praktycznych badań rozpoznawczych jest nie do przyjęcia. Przedstawione badanie to typowe badanie podstawowe, tylko słabo osadzone w teorii, bez sformułowania oryginalnego problemu badawczego. Podłączenie do kategorii badań rozpoznawczych jest też próbą zamaskowania słabości teoretycznej pracy - braku krytycznej syntezy stanu wiedzy i znalezienia oryginalnego problemu badawczego.

Powtórzmy, problem badawczy w baniach podstawowych powinien być wynikiem krytycznej syntezy dotychczasowego stanu wiedzy i identyfikacji obszarów niepewności. Problem badawczy (pytania, ew. hipotezy) powinien, jeżeli uda się go rozwiązać, przyczynić się do rozwoju teorii.

W rozdziale 2.3. *Problemy badawcze* znajdujemy ostateczną postać problemów badawczych. Doktorant wskazuje na 5 głównych. Przykładowo pierwszy z nich brzmi:

„1. W jakim stopniu w poszczególnych pomiarach (na początku oraz na końcu roku szkolnego w IV klasie oraz pod koniec V klasy) występuje wśród uczniów zjawisko zagrożenia stereotypem niskich zdolności oraz poczucia bezradności na lekcjach języka polskiego i matematyki oraz jaki jest styl wyjaśniania przyczyn własnych sukcesów i porażek w nauce, które dotyczą uczniów, a także jakie są wyniki w nauce osiągane przez nich na lekcjach języka polskiego i matematyki w poszczególnych semestrach dwóch lat nauki?"

Pamiętając z wcześniejszych deklaracji Doktoranta, że wyniki badania nie będą mogły być generalizowane na szerszą populację, nasuwa się pytanie, co właściwie chce zbadać? Czy chodzi tylko o grupę przebadanych uczniów (max 150)? To jaką lukę w naszej wiedzy to zapełnia? Jak to pytanie/problem ma się do dotychczasowych teorii i wyników badań? Jak to rozwija teorię?

Patrząc na problem od strony pomiarowej powstaje wątpliwość, czy jesteśmy w stanie zinterpretować wyniki dla takich zmiennych jak zagrożenie stereotypem niskich zdolności czy wyuczonej bezradności? Czy teoria podpowiada jakieś kryteria interpretacji nasilenia? Czy mamy szansę na odniesienie wyników do norm populacyjnych? Niestety nie znajdziemy odpowiedzi na te pytania w pracy Doktoranta.

Przyjrzymy się bardziej obiecującemu teoretycznie problemowi nr 3:

„3. Jakie znaczenie dla poziomu zmiennych w poszczególnych pomiarach oraz dla zachodzących zmian mają płeć i status socjoekonomiczny rodziny ucznia?”

I w tym wypadku razi ateoretczny język sformułowania problemu badawczego. Ale sam problem wydaje się ciekawy, bo jest mniej wrażliwy na ograniczania generalizacji wyników z małych, nierepre-zentatywnych prób. O co w tym pytaniu chodzi? Jak można rozumieć, przypuszcza się, że trendy roz-wojowe w nasileniu badanych konstruktów są moderowane przez SES rodziny ucznia i jego płeć. Szczególnie ta pierwsza zależność byłaby bardzo ciekawa. Doktorant zresztą ma tego świadomość i o związku moderującego znaczenia SES dla trendów w zakresie zagrożenia stereotypem niskich zdol-ności, wyuczonej bezradności i stylów atrybucji porażek i sukcesów szkolnych z problematyką nie-równości edukacyjnych wspomina. Jednak nie rozwija tego wątku w dojrzałą postać problemu nau-kowego. A sądzę, że mógłby to być samodzielny problem naukowy w zupełności wystarczający na dobrą pracę doktorską. Jeszcze do tego wątku w recenzji powrócę.

Dopełnieniem sformułowania przez Doktoranta problemu badawczego są opisane w podrozdziale 2.4 hipotezy. Niestety, podobnie jak pytania badawcze, nie mają one wiele wspólnego z myśleniem teoretycznym. Samo ich sformułowanie w języku techniczno-statystycznym jest dziwaczne. Na przy-kład H3.2.1:

„Przynajmniej w jednym pomiarze istnieją istotne statystycznie różnice w średnim poziomie nastę-pujących zmiennych: poczucie zagrożenia stereotypem niskich zdolności oraz poczucie bezradności na lekcjach języka polskiego i matematyki, styl wyjaśniania przyczyn własnych sukcesów i porażek w nauce, średnia ocen cząstkowych z języka polskiego i matematyki, pomiędzy co najmniej dwoma wy-odrębnionymi z uwagi na SES grupami.”

Formułowanie hipotez "zbiorowych" nie ma sensu, również hipotezy nie mówiące o kierunku za-leżności są teoretycznie bezwartościowe. Niektóre hipotezy prowadzą natomiast do skrajnie banal-nych przewidywań. Np z H3.2.1. wynika, że oczekujemy zależności SES i ocen szkolnych.

Metoda rozwiązania problemu

W pracy znajdujemy dużo odwołać do literatury metodologicznej i statystycznej. Takie nawiązania są celowe, gdy badacz rozważa jakieś dylematy przed którymi stanął i uzasadnia przejęty sposób jego rozwiązania. Warto też wskazywać źródła w wypadku korzystania z nowych lub mało znanych metod badawczych i modeli statystycznych. Odwoływanie się jednak do literatury w wypadku „elementarza”

metodologicznego czy standardowych, prostych metod statystycznych, jest tylko „nabijaniem stron" psującym jakość tekstu. Najbardziej kuriozalny przykład znajdujemy na s. 195:

„Związek statystyczny (korelacyjny) polega na tym, „że określonym wartościom jednej zmiennej odpowiadają ściśle określone wartości drugiej zmiennej" (Stanisz, 2006, s. 290). Korelacja dodatnia występuje wtedy, gdy wzrostowi wartości jednej cechy odpowiada wzrost średnich wartości drugiej cechy, natomiast ujemna, gdy wzrostowi wartości jednej cechy odpowiada spadek średnich wartości drugiej cechy (Stanisz, 2006, s. 290)."

Ciekawe, że Doktorant uznał za stosowne zacytować autora podręcznika, zamiast napisać to samodzielnie.

Dobór próby

Jednym z kluczowych elementów metody badawczej jest dobór próby. Ponieważ budzi on w pracy Doktoranta spore wątpliwości, pozwolę sobie na obszerny cytat z niewielkimi tylko skrótami.

„Klasy do badań zostały wylosowane z trzech warstw szkół wyodrębnionych na podstawie średniego wyniku sprawdzianu przeprowadzanego przez Centralną Komisję Egzaminacyjną na zakończenie klasy VI, uzyskanego przez uczniów danej szkoły w ciągu trzech ostatnich lat przed rozpoczęciem badań: 2016, 2015, 2014. Umiejscowienie szkół w poszczególnych latach nieco się zmieniało, dlatego ostateczny wynik uśredniono. Szkoły, które znalazły się w przedziale od 1 do 3,5 stanina uznano za warstwę 1, szkoły z wynikami w przedziale od 3,6 do 6,5 warstwę 2, a szkoły z wynikami od 6,6 do 9 stanina – warstwę nr 3. Wykorzystując formułę programu Excel losowano klasy z utworzonych według tych kryteriów trzech grup szkół równocześnie. Na tym samym arkuszu ustawiono także formułę losowania klasy w zależności od liczby klas w danej szkole. W przypadku wylosowania danej szkoły, sprawdzano ile obecnie jest w niej klas IV i sprawdzano, jaka klasa (a, b, c, d…) została wylosowana. W kolejnym kroku wszystkim wylosowanym klasom w poszczególnych grupach nadano numery 1-15 pierwsza grupa, 16-30 druga grupa, 31-45 trzecia. (…) Kolejność ich wylosowania pozwalała na wybór dziewięciu klas do projektu, a pozostałych pozostawienie jako rezerwowych. (…) W drodze opisanej procedury losowej do grupy kontrolnej projektu, zakwalifikowano dziewięć klas IV, którym zagwarantowano anonimowość. Z uwagi na fakt zagubienia numerów kodowych w jednej szkole oraz podłużny charakter badań, została ona wyłączona z prezentowanych analiz."

Pominę analizę klarowności tego opisu i przejdę do meritum. W próbie (grupa kontrolna w eksperymencie) znalazło się zatem 8 oddziałów klasy IV. Z danych podanych w innych miejscach pracy

dowiadujemy się, że dane zgromadzono dla od 103 (wskaźnik ISES) do 131 uczniów (II pomiar bezradności). Niestety zabrakło w pracy tabeli, która by zbiorczo przedstawiała liczebności uczniów dla poszczególnych pomiarów. Brak też analizy losowości braków danych, a nawet prostej informacji, ilu uczniów było w tych 8 oddziałach i z której warstwy szkoła wypadła.

Jednak nie to jest najbardziej niepokojące. Główny grzech metodologiczny polega na tym, że w analizach statystycznych nie uwzględniono pogrupowania danych i próbę traktowano jako próbę prostą. Wiadomo, że w wypadku pogrupowaniu danych, należy ten fakt uwzględnić w modelach analizy, a szczególnie w szacowaniu błędów standardowych. Gdy dane są pogrupowane, to możliwe jest, że podobieństwo jednostek w obrębie grupy (oddział klasowy) jest większe niż przeciętne podobieństwo jednostek w całej próbie/populacji. Gdyby jednostki w obrębie klastra były identyczne, to efektywnie wielkość próby równałaby się liczbie wylosowanych grup. W wypadku analizowanego badania jest wysoce prawdopodobne, że korelacja wewnątrzklasowa (ICC) dla głównych zmiennych była znacząca. Wynika to z tego, że w losowaniu do stworzenia warstw wykorzystano średnie wyniki sprawdzianu szóstoklasisty w każdej szkole, a kluczowe zmienne w badaniu są z tymi wynikami znacząco skorelowane. Należy oczekiwać dość wysokiego ICC też z tego powodu, że zastosowane warstwowanie połączone ze sposobem alokacji szkół w warstwach „wyostrzyło" wariancję międzyszkolną czyli zmniejszyło wariancję wewnątrzszkolną (z niewiadomych powodów wybierano po 3 szkoły z wyróżnionych warstw, zamiast 2-5-2 lub 3-4-3). Na podstawie wiedzy o ICC w populacji szkół podstawowych w Bydgoszczy i konsekwencji zastosowanej alokacji szkół w warstwach oraz faktu, że finalnie wybierano oddziały klasowe (czyli wchodzi jeszcze czynnik segregacji wewnątrzszkolnej) można ICC ostrożnie szacować na 0,20.

Jaka jest efektywna wielkość próby przy takim schemacie doboru? Prosta, podręcznikowa formuła wygląda następująco:

$$N_e=(k \times m)/(1 + ICC \times (m-1))$$

gdzie k-liczba klastrów; m-średnia liczebność klastra; ICC – wsp. korelacji wewnątrzklasowej

Po podstawieniu wartości k=8, m=19 i ICC=0,20 otrzymujmy wielkość efektywną próby ok. 63. Gdy weźmiemy pod uwagę braki danych, wartość ta nie przekroczy dla niektórych analiz 50.

Co to znaczy? Oznacza to, że stosowanie w analizach statystycznych modeli zakładających, że mamy do czynienia z próbą prostą, prowadzi do znacznego zaniżenia wielkości błędów standardowych oszacowań a tym samym do zaniżania wartości p. Tak mała efektywna wielkość próby oznacza

też, że np. w wypadku prostego testu t-Studenta, będziemy w stanie wykrywać efekty rzędu d=1. To praktycznie uniemożliwia weryfikację jakiejkolwiek sensownej hipotezy dotyczącej badanej problematyki.

Pomiar zmiennych

W wypadku pomiaru kluczowych zmiennych brak przemyślanej strategii wyboru narzędzi, skalowania wyników i badania ich jakości. Choć – poza jednym – w badaniu stosuje się istniejące już testy, to analizy ich trafności i rzetelności są tak rozbudowane, jakby to było głównym celem badania (ok. 40 stron). W dalszej ocenie pominę kwestię teoretycznego uzasadnienia wyboru narzędzi, bo jest to efekt słabego zakorzenienia teoretycznego problemu badawczego.

Ocenę tego fragmentu pracy zacznę od przyjętego modelu skalowania. Zasadniczo mamy do czynienia z podejściem zgodnym z klasyczną teorią testu wzbogaconym analizami z wykorzystaniem modeli głównych składowych i konfirmacyjnych analiz czynnikowych. Gdyby przyniosło to dobre efekty, to anachronizm KTT nie byłby wystarczającym powodem do krytyki. Jednak znając problemy, na które napotkał Doktorant, można stwierdzić, że podejście teorii odpowiedzi na zadanie testowe (IRT) byłoby znacznie bardziej użyteczne.

Zacznijmy od początku. W pracy analizie narzędzi poświęcono dziesiątki stron, a mimo to zabrakło miejsca na przedstawienie rozkładów odpowiedzi w poszczególnych pozycjach testowych. To bardzo ważna analiza, bo pozwala wybrać adekwatny model skalowania. Poza tym może być pomocna przy interpretacji statystyk opisowych dla wskaźników sumarycznych. Niestety nic o tych rozkładach nie wiemy, a prawdopodobnie w wypadku wielu pozycji mamy rozkłady bardzo nierównomierne, być może skoncentrowane na skrajach skal odpowiedzi.

Kolejny krok to skalowanie wyników. Doktorant mimo wykonania dziesiątek analiz struktury testów, też z użyciem CFA, wskaźniki sumaryczne dla testów oblicza jako średnią z arbitralnie przypisanych wag do pozycji na skalach szacunkowych. Za każdym razem starannie sprawdza spełnienie założenia o normalności rozkładu i – zwykle - z powodu jego niespełnienia wybiera nieparametryczne modele analizy. Jednak kompletnie traci z pola uwagi fakt, że licząc średnie arytmetyczne z odpowiedzi respondentów potraktował dane wejściowe jako pomiar interwałowy! Skośność rozkładu jest oczywiście istotną przeszkodą do stosowania wielu modeli statystycznych, ale można było tego uniknąć stosują inny model skalowania. Gdyby potraktować odpowiedzi respondentów w testach jak dane porządkowe i do skalowania wykorzystać któryś z modeli IRT, być może można by było bez

problemów stosować bardziej efektywne modele statystyczne analizy zależności. Zresztą pozostając na gruncie KTT można było wskaźniki surowe znormalizować i przełożyć na jakąś skalę standardową.

W analizach struktury testów brak adekwatnej strategii analizy. Eksploracyjna analiza czynnikowa może być narzędziem pomocniczym, głównym powinna być jednak konfirmacyjna analiza czynnikowa. CFA jest przez Doktoranta stosowana, ale w trudnych do przewidzenia sytuacjach wyskakuje jak filip z konopi. Natomiast analizy EFA nie wiadomo dlaczego są robione głównie metodą analizy głównych składowych z rotacją Varimax. Założenie ortogonalności czynników po rotacji jest kompletnie nieuzasadnione, czego Autor zresztą sam dowodzi w przeprowadzonych analizach SEM, np. w wypadku skali poczucie bezradności na lekcjach języka polskiego korelacje między podwymiarami skali wynoszą nawet blisko 0,8.

Podstawową przewagą CFA nad EFA jest możliwość estymacji miar dopasowania. To podstawowe statystyki pozwalające argumentować, że przyjęty model pomiarowy znajduje potwierdzenie w danych, miary te pozwalają też wybierać spośród alternatywnych modeli ten najlepszy. W analizach CFA przeprowadzonych przez Doktoranta tych miar nie znajdziemy. Autor pisze, że założenia zastosowanych modeli CFA nie były spełnione i miar dopasowanie nie można estymować. Jednak jeżeli założenia nie są spełnione, to model w całości daje niewiarygodne wyniki. Prawdopodobnie problem wynika z wykorzystania modelu zakładającego, że zmienne wejściowe są interwałowe. Trzeba było wykorzystać inny model zakładający, że dane na poziomie itemów są porządkowe i wykorzystać algorytmy programu MPlus.

Wątpliwości budzi też strategia badania trafności kryterialnej pomiarów. Dobór kryteriów jest ateoretyczny. Dlaczego na przykład zagrożenie stereotypem niskich zdolności miałoby korelować z relacjami z rówieśnikami? Trzeba też zwrócić uwagę, że w badaniu trafności zabrakło miejsca na trafność kryterialną różnicową, a zaprezentowane wyniki wskazują, że problem jest. Na przykład korelacja zagrożenia stereotypem z bezradnością na poziomie 0,7 to już zagrożenie dla trafności różnicowej.

Ważny pomiar w badaniu to osiągnięcia szkolne. Na s. 122 Doktorant charakteryzuje pomiar tej zmiennej następująco:

„Realizacja wymagań programowych (szkolnych) to stopień opanowania przez ucznia wymaganej na danym etapie nauczania wiedzy i umiejętności, które określa podstawa programowa kształcenia ogólnego (Stróżyński, 2000, s. 35), w tym przypadku dla IV i V klasy szkoły podstawowej. Pełni ona rolę informacji o stopniu opanowania wymaganej wiedzy dla samego ucznia, jego rodziców oraz

nauczyciela, wskazującej deficyty w celu ich eliminowania lub pokazującej obszary rozwijania zainteresowań."

Takie ujęcie stopni szkolnych w pracy naukowej jest niedopuszczalne. Naukowych opracowań na ten temat jest mnóstwo, natomiast przywoływany tekst z pewnością nie należy do klasyki. Bez rozważenia ograniczeń tej miary, nie wiemy, co liczymy.

W pracy zabrakło też opisu platform testowych.

Modele analizy danych

Ocenę zastosowanych modeli analizy danych trzeba zacząć od głównego problemu, czyli braku uwzględnienia w analizach złożonego schematu losowania (pogrupowania danych). Oznacza to, że wszystkie oszacowania błędów standardowych statystyk są prawdopodobnie znacznie zaniżone, a zatem zależności statystyczne uznane za istotne, prawdopodobnie takimi nie są. Oczywiście dobra praca naukowa może kończyć się wnioskiem, że nie ma dostatecznych dowodów, że jakaś zależność istnieje. Jeżeli badanie jest poprawne metodologiczne, to taki wynik wzbogaca naszą wiedzę. Jednak prezentowane badania nie spełniają tego warunku. Co więcej, brak pewnych zależności (patrz dalsza część tej części recenzji) wskazuje, że mamy do czynienia z „dziwnym" zestawem danych.

Przejdźmy do podrozdziału 3.1. Podaje się w nim statystyki opisowe zmiennych. Prawie na 20 stronach na różne sposoby analizuje się rozkłady! Cały ten rozdział jest irytująco rozwlekły. Moim zdaniem trzeba było użyć narzędzi graficznych i tabel zbiorczo przedstawiających parametry. Pozwoliłoby to radykalnie skrócić ten podrozdział.

Poważniejszy problem widzę jednak w strategii analiz. Ma ona znamiona „ateoretycznej kombinatoryki statystycznej": 7 zmiennych zależnych, 3 pomiary, dwa moderatory, jeszcze na dokładkę analizy związku moderatorów ze zmiennymi zależnymi. Jak się zrobi 100 analiz na losowym szumie, to przy $p<0,05$, pięć da "istotne" wyniki. Wszystko to wygląda jak pomysł na generowanie miliona tabelek i wykresów, z których nic nie wynika, bo nie może ze względu na niewielką próbę i małą efektywność nieparametrycznych modeli statystycznych.

Słabość strategii analizy danych jest częściowo pochodną praktycznie ateoretycznego formułowania problemów badawczych. Gdyby skupić się na jednym, dobrze teoretycznie ugruntowanym problemie i wykorzystać podłużny charakter badania, to być może dałoby się zebrane dane wykorzystać do wstępnego przetestowania jakiegoś oryginalnego pomysłu analitycznego. Jak trafnie zauważa Doktorant, moderujący wpływ SES na trendy w poziomie zagrożenia stereotypem niskich zdolności

czy wyuczonej bezradności może się wiązać z problematyką nierówności edukacyjnych. Z pewnością zdecydowanie zbyt mało wiemy o mechanizmach odpowiedzialnych za korelację statusu przypisanego z osiągnięciami szkolnymi. Trochę tę „czarną skrzynkę" otwierają badania z zakresu genetyki zachowania, teorii oczekiwań, czy badania nad wpływem rówieśników, ale niewiele jest prób wyjaśniania statusowej determinacji przez takie procesy psychologiczne jak zagrożenie stereotypem czy wyuczona bezradność. Gdyby sprawdzić, czy związek SES -> osiągnięcia szkolne jest mediowany przez zagrożenie stereotypem niskich zdolności czy przez wyuczoną bezradność (dane podłużne świetnie się do tego nadają), to byłby to realny wkład do naszej wiedzy. Gdyby jeszcze potwierdził się (zakładany w pracy) moderujący wpływ SES na trendy w zakresie tych procesów, to moglibyśmy dowiedzieć się czegoś naprawdę ważnego. Nawet gdyby mała próba nie dała konkluzywnych wyników, byłby to ciekawy test podejścia.

Jednak to tylko mrzonki recenzenta. Jak już wspomniałem w pracy testuje się „hipotezę" o istnieniu związku SES z ocenami szkolnymi. To oczywiście bezsensowny zamiar z rodzaju dowodzenia, że ziemia jest okrągła. Ale tu zaskoczenie. W badanej próbie analizy wykazały, że „ziemia jest płaska" - brak istotnego statystycznie efektu SES zarówno dla ocen z polskiego jak i matematyki. Czy to oznacza, że faktycznie trzeba wątpić w determinizm statusowy osiągnięć szkolnych? Co prawda Doktorant poprzez modyfikacje modelu analizy „znalazł" w końcu szukany efekt, ale jego minimalna siły dowodzi, że to z próbą lub danymi coś jest nie tak. Już nie tylko chodzi o to, że próba jest malutka, ale jest mocno nietypowa. Jeżeli brak znaczącej korelacji SES – osiągnięcia szkolne, trudno oczywiście testować te wszystkie przepuszczenia sformułowane przeze mnie w poprzednim paragrafie.

Wraca więc problem próby. Niestety procedura badawcza jest jak wieloprzęsłowy most: jego wytrzymałość nie zależy od najmocarniejszego przęsła, nie jest też średnią wytrzymałości elementów składowych przeprawy. Zależy od wytrzymałości najsłabszego przęsła! Brak typowego efektu korelacji SES rodziny ucznia z osiągnięciami wskazuje w sposób dobitny, że nie tylko mamy problem małej próby, ale na dokładkę jest ona niereprezentatywna ze względu na bardzo ważną w badaniach edukacyjnych zależność.

Na koniec oceny zastosowanych modeli statystycznych zajmijmy się analizą dotycząca *ad hoc* sformułowanego pytania o związek średniego SES ze średnią ocen w oddziale klasowym. Szacowanie efektu składu oddziału za pomocą modelu jednopoziomowego jest niepoprawne. Trzeba było użyć modelu dwupoziomowego (uczeń, oddział). Choć nie rozwiązałoby to problemu dla tej analizy kluczowego: n dla tej analizy wynosi 8! Dodatkowo w wypadku zmiennej wyjaśnianej, można mieć

wątpliwości co do wiarygodności międzygrupowej wariancji ocen nauczycielskich. Skąd się bierze tak silna korelacja? Można by szukać przyczyn, ale bardziej zasadne jest uznanie, że dla takich mikroskopijnych prób statystyka jeszcze nie działa.

Język rozprawy i staranność edycji

Język naukowy rozprawy jest nierówny, choć generalnie poprawny. Czasami zdarzają się błędy logiczne. Np. na s. 25. znajdujemy fragment: „Jako desygnat pojęcia stosuje się także termin zagrożenie samooceny, z towarzyszącym mu niepokojem wynikającym z potwierdzenia negatywnego stereotypu w oczach innych lub własnych, prowadzącym do zachowania zgodnego ze stereotypem…" W jakim sensie termin może być desygnatem pojęcia? Jak terminowi może towarzyszyć niepokój? Zdarzają się też błędy ortograficzne, ale nieliczne. Na str. 91 znajdziemy: „W Polskiej klasyfikacji z …". Natomiast w pracy jest sporo niestaranności. I tak w tytule rysunku 12. znajduje się informacja o tym, że znajdą się na nim statystyki opisowe. Nie ma ich tam. Z kolei w opisie ISES czytamy: „Wartości indeksu mieszczą się w przedziale 10 do 70." Z tabeli wynika, że 14-70. W opisie testu do pomiaru zagrożenia stereotypem czytamy: „Badania właściwe zostały poprzedzone badaniami pilotażowymi przeprowadzonymi w dziewięciu klasach (dwie klasy – IV, dwie – V oraz dwie – VI) w trzech bydgoskich szkołach (N=139)." To w 9 czy w 6? W tabeli 55: w dwóch polach pojawiają się korelacje ujemne? To raczej błąd, powinny być dodatnie. Na rysunku 49: błąd w legendzie (3 razy IV klasa).

Konkluzja

Z pewnością mocną stroną pracy jest rozległy przegląd literatury i bardzo bogaty arsenał zastosowanych narzędzi statystycznych. Rozprawa potwierdza również orientację Doktoranta w dyscyplinie. Niestety sformułowane przez niego problemy badawcze są bardzo słabo zakorzenione w teorii a metodologia ich rozwiązania ma nieusuwalne wady. Nie mogę zatem wnioskować o dopuszczenie mgr Szymona Borsicha do dalszych etapów przewodu doktorskiego.