Artur Pokropek
Polska Akademia Nauk
Instytut Filozofii i Socjologii

Ernest Pokropek
Division of Robotics, Perception, and Learning,
KTH Royal Institute of Technology

Wykorzystanie głębokich sieci neuronowych do wykrywania błędnych specyfikacji modeli statystycznych. Przypadek równoważności pomiarowej

Metodologiczne Inspiracje 2021
23 - 24 września, Jabłonna

This presentation has been prepared under the Scales Comparability in Large Scale Cross-country Surveys Project, which is funded by the Polish National Science Centre, as part of the grant competition Sonata 8 (UMO-2014/15/D/HS6/04934).

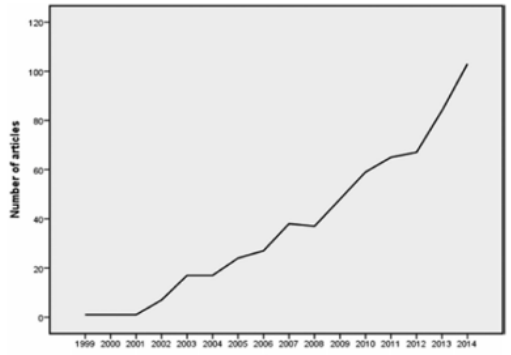
NATIONAL SCIENCE CENTRE POLAND

Kontekst

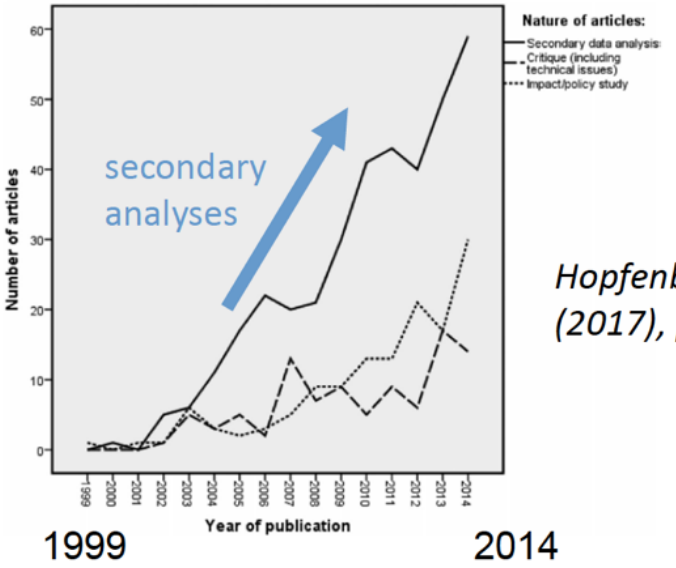
- W ostatnich latach odnotowujemy coraz większe zainteresowanie porównawczymi badaniami sondażowymi
- *European Social Survey* (**ESS**), *International Social Survey Programme* (**ISSP**), *World Values Survey* (**WVS**), *Programme for International Student Assessment* (**PISA**), *International Mental Health Stigma Survey* (**IMHSS**), *Gallup World Poll* (**GWP**), *International Adult Literacy Survey* (**IALS**), *Adult Literacy, and Life Skills Survey* (**ALL**), *Programme for the International Assessment of Adult Competencies* (**PIAAC**), *Eurobarometer, Demographic and Health Surveys* (**DHS**), *Multinational Time Use Study* (**MTUS**), *World Health Survey* (**WHS**) and *World Internet Project* (**WIP**), *Progress in International Reading Literacy Study* (**PIRLS**), *Trends in International Mathematics and Science Study* (**TIMSS**), *Teaching and Learning International Survey* (**TALIS**), *Gallup Global Wellbeing Study* (***GGWS**; 155 krajów*)
- … a lista ta odnosi się tylko do największych projektów.

Kontekst

- Duże zainteresowanie w środowiskach naukowych
- Ustanowiono nowe czasopisma poświęcone badaniom porównawczym (np. *Large-scale Assessments in Education*)
- Liczba publikacji wykorzystujących dane z badań porównawczych rośnie
- Przykład dla badania PISA (Hopfenbeck et al., 2017)



Hopfenbeck et al. (2017), p. 5



Hopfenbeck et al. (2017), p. 8

Kontekst

- W każdym badaniu mierzone są różne konstrukty i budowane różne skale:
  - ESS koncentruje się na wartościach, orientacjach ideologicznych, orientacjach kulturowych, narodowych i podstawowej strukturze społecznej społeczeństwa (Fitzgerald & Jowell, 2010).
  - Każda fala badania ISSP skupia się na różnych aspektach życia społecznego, w tym na roli rządu, sieci społecznych i systemów wsparcia, na nierównościach społecznych, rodzinie, bada zmienianie się ról płciowych, orientacji na pracę, religii, (…), (Skjak, 2010).
  - Międzynarodowe badania umiejętności, takie jak PISA, PIRLS, TIMSS, PIAAC, zostały zaprojektowane do oceny wiedzy uczniów i ich umiejętności na całym świecie (Rutkowski, Gonzalez, Joncas, & von Davier, 2010; von Davier, Gonzalez, Kirsch, & Yamamoto, 2013), mierzą też konstrukty psychologiczne (np. poczucie własnej skuteczności w PISA czy wielka piątka cech osobowości w przyszłym PIAAC).
  - Badania porównawcze skal psychologicznych (np. OPQ32 32 cechy osobowości związane z pracą; Bartram 2013).

(Potencjalne) problemy

- W badaniu ESS skonstruowano skalę mierzącą poziom religijności. Skala składa się z kilku pytań dotyczących ogólnej religijności i praktyk religijnych. Turcja okazała się krajem o najniższej średniej wartości tego wskaźnika. Szczególnie niski wskaźnik obserwowany był wśród kobiet.
  - Jedno z pytań dotyczyło częstości odwiedzania kościoła/meczetu. Z przyczyn kulturowych w Turcji kobiety nie chodzą regularnie do meczetów.

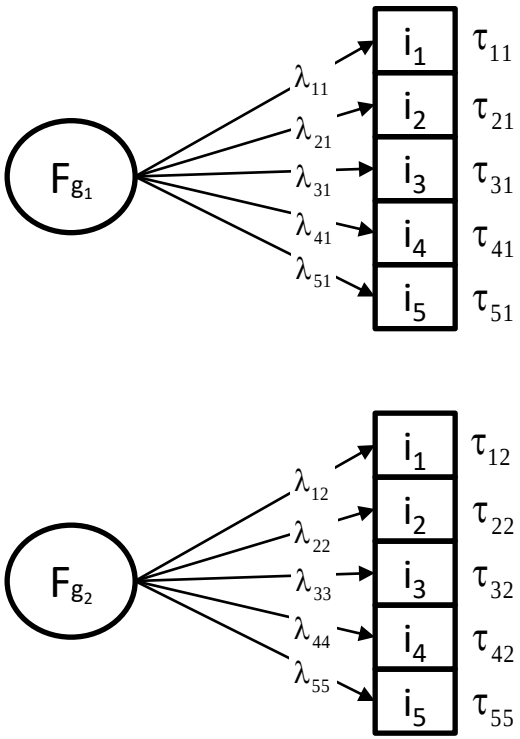
- W ESS zadaje się też serie pytań o stosunek do imigrantów. Danii w 2002 roku odnotowano szczególnie niski wskaźnik „niechęci” do imigrantów.
  - Problem jak się okazało tkwił w tłumaczeniu. Zamiast "punish immigrants that commit a crime", treść pytania przetłumaczono jako "punish immigrants who commit an offense"

Billiet J. 2013. Quantitative methods with survey data in comparative research. In A Handbook of Comparative Social Policy, ed. P Kennett, pp. 264–300. Cheltenham, UK: Edward Elgar. 2nd ed.

Prace metodologiczne

- Osiągnięto znaczny postęp w zapewnieniu pojęciowej porównywalności wskaźników postaw, wielu wskaźników pozycji społecznej, testów psychologicznych
- Jeszcze kilka lat temu problem porównywalności rzadko był analizowany po zebraniu danych, na poziomie modelowania statystycznego
- Sytuacja ta niekiedy doprowadzała do błędnych wniosków w niektórych wpływowych badaniach (Davidov 2009; Davidov, Meuleman, Cieciuch, Schmidt, and Billiet 2014; Ippel, Gelissen, and Moors 2014)
- Dziś, większość badaczy zgadza się z tym, że porównywalność wskaźników powinna być empirycznie testowana (przynajmniej zawsze tam, gdzie porównania stanowią oś badań – Davidov et al. 2014)

MG-CFA/IRT



...

- **Wielogrupowa konfirmacyjna analiza czynnikowa (MG-CFA):**

$$y_{ig} = \tau_{ig} + \lambda_{ig}\theta + e_{ig}$$

- **Wielogrupowy model odpowiedzi częściowej (MG-GR):**

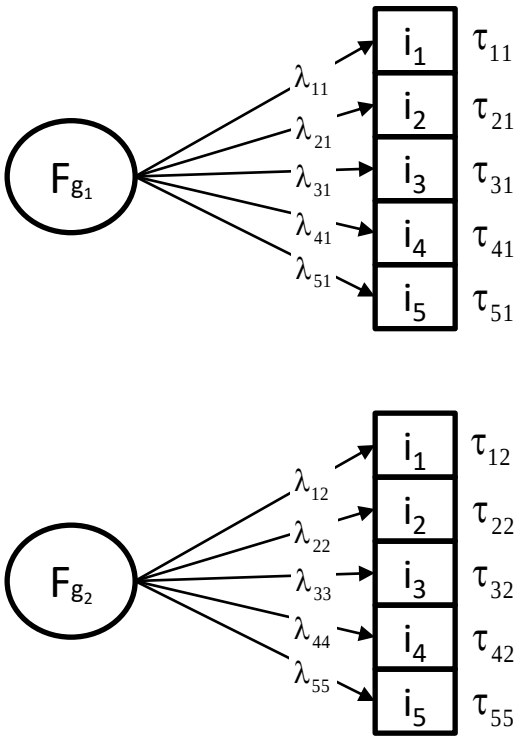
$$P(y_{ig} = c) = g[a_{ig}(\theta - b_{ig,c-1})] - g[a_{ig}(\theta - b_{ig,c})]$$

- **Uproszczenie w tej prezentacji:**

$$\tau_{ig} \approx b_{ig,c}$$

$$\lambda_{ig} \approx a_{ig}$$

Równoważność (niezmienność) pomiarowa

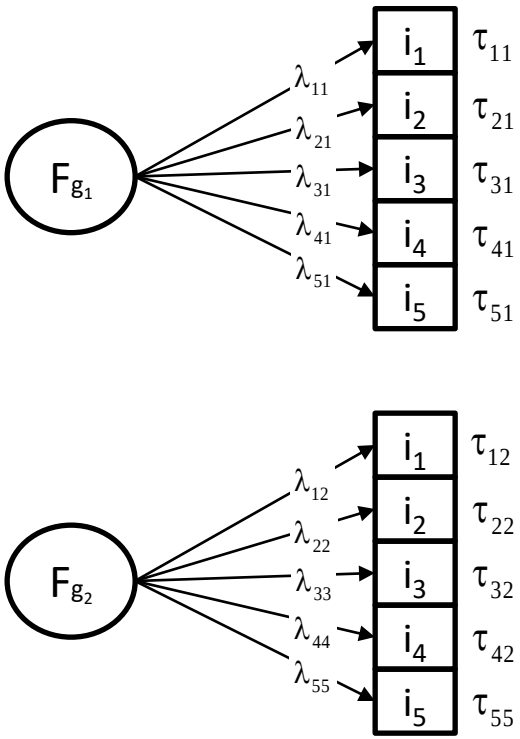


...

$$\lambda_{11} = \lambda_{12} \qquad \tau_{11} = \tau_{12}$$
$$\lambda_{21} = \lambda_{22} \qquad \tau_{21} = \tau_{22}$$
$$\lambda_{31} = \lambda_{32} \qquad \tau_{31} = \tau_{32}$$
$$\lambda_{41} = \lambda_{42} \qquad \tau_{41} = \tau_{42}$$
$$\lambda_{51} = \lambda_{52} \qquad \tau_{51} = \tau_{52}$$

Niestety, bardzo rzadko spotykana w warunkach naturalnych!

Brak równoważności pomiarowej

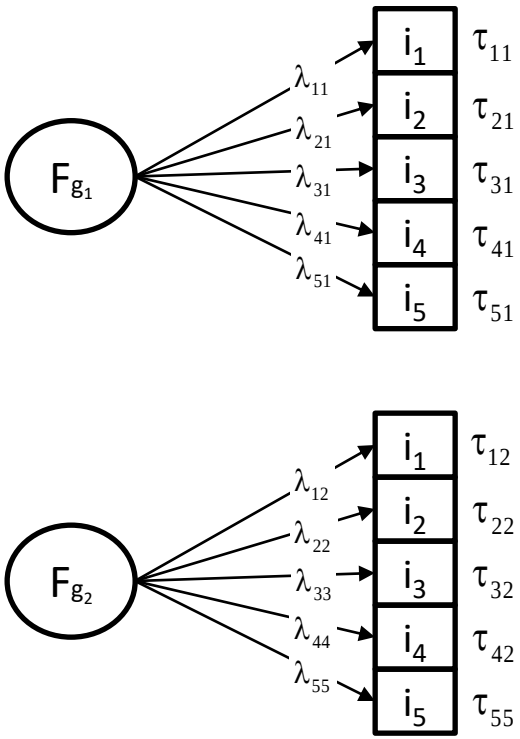


$$\lambda_{11} = \lambda_{12} \qquad \tau_{11} = \tau_{12}$$
$$\lambda_{21} \neq \lambda_{22} \qquad \tau_{21} = \tau_{22}$$
$$\lambda_{31} = \lambda_{32} \qquad \tau_{31} \neq \tau_{32}$$
$$\lambda_{41} = \lambda_{42} \qquad \tau_{41} = \tau_{42}$$
$$\lambda_{51} = \lambda_{52} \qquad \tau_{51} \neq \tau_{52}$$

•••

Pokropek, A., Davidov, E., & Schmidt, P. (2019). A Monte Carlo simulation study to assess the appropriateness of traditional and newer approaches to test for measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(5), 724-744.

Brak równoważności pomiarowej



$$\lambda_{11} = \lambda_{12}$$
$$\lambda_{21} \neq \lambda_{22}$$
$$\lambda_{31} = \lambda_{32}$$
$$\lambda_{41} = \lambda_{42}$$
$$\lambda_{51} = \lambda_{52}$$

$$\tau_{11} = \tau_{12}$$
$$\tau_{21} = \tau_{22}$$
$$\tau_{31} \neq \tau_{32}$$
$$\tau_{41} = \tau_{42}$$
$$\tau_{51} \neq \tau_{52}$$

...

Pokropek, A., Davidov, E., & Schmidt, P. (2019). A Monte Carlo simulation study to assess the appropriateness of traditional and newer approaches to test for measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(5), 724-744.

Metody wykrywania

1. Modification Index (MI) and the Power of the Test
   - Saris, Satorra, van der Veld (2009); Cieciuch, J., Davidov, E., Oberski, D.L., & Algesheimer, R. (2015)
2. Bayesian Structural Equating Modeling measurement invariance analysis
   - De Jong et al. (2007); Muthén & Asparouhov (2013:9)
3. Alignment optimization invariance analysis
   - Asparouhov & Muthén (2014:7); Muthén & Asparouhov, (2014)
4. Multilevel Confirmatory Factor Analysis
   - Fox, (2010); Hox et al. (2010); Muthén & Asparouhov (2013)
5. DIF (Mantel-Haenszel, Generalized Regression Spproach, IRT fit)
   - Mantel & Haenszel (1959); Swaminathan & Rogers (1990)
6. A Regularized Moderated Item Response Model for Assessing Dierential Item Functioning
   - Robitzsch & Lüdtke (2018)

(…)

Metody wykrywania

1. Większość metod zaprojektowana dla dwóch grup (np. Mantel–Haenszel test, regresja logistyczna)
2. W kontekście dwóch grup działają dobrze
3. Generalizacje na większą ilość grup (dodatkowo przy ograniczonej liczbie wskaźników) nie działają efektywnie
4. Metody klasycznie dostosowane do wielu grup również pozostawiają wiele do życzenia
5. Prace Lin (2020), Fincha (2016), Asparouhova o Muthéna (2014) pokazują, że dobra detekcja nierównoważnych parametrów znanymi narzędziami uzyskiwana jest jedynie gdy kraje nie różną się znacząco ze względu na badane cechy…

Sieci neuronowe

- Badania nad sztucznymi sieciami neuronowymi (ANNs) trwają już od kilkudziesięciu lat (Fukushima, 1975; Rumelhart i in., 1986), ale dopiero w ostatnich latach rozwój wydajnych algorytmów treningowych (Bengio i in., 2007; Hinton i in., 2006; Ranzato i in., 2007) oraz znaczący wzrost mocy obliczeniowej (Schmidhuber, 2015).
- Osiągają spektakularne sukcesy np. dziedzina rozpoznawania obrazów. Obecne rozwiązania DNN osiągają w takich zadaniach dokładność na poziomie **98 procent** (podczas gdy podejścia oparte na regresji i maszyny wektorów nośnych nie są w stanie osiągnąć więcej niż 75% dokładności (Mitchell, 2019)
- Znalazły również zastosowanie w przetwarzaniu obrazów, rozpoznawaniu mowy, tłumaczeniu maszynowym, robotyce, przetwarzaniu języka naturalnego, cyberbezpieczeństwie i wielu innych dziedzinach użytkowych (Schmidhuber, 2015).

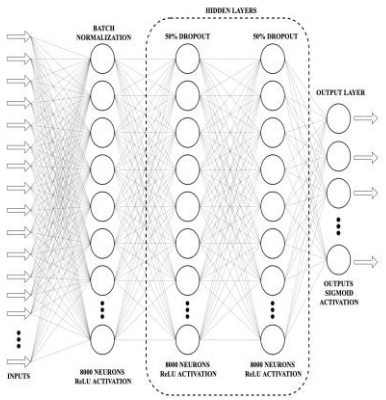
Co znaczy 98%?

- Entuzjaści mówią o przełomie, twierdząc, że modele wyprzedzają ludzi w zadaniach rozpoznawania obrazów.
- W typowym zadaniu predykcji algorytmy dostarczają uszeregowaną listę pięciu możliwych kategorii dla każdego obrazu. Jeśli wśród pięciu obiektów jest jeden, który faktycznie znajduje się na zdjęciu, predykcja jest uznawana za poprawną (Goodfellow i in., 2016, s. 23),
- Co dopuszcza wiele poważnych błędów, których ludzie zazwyczaj nie popełniają. Z pewnością ludzie wciąż są lepsi od DNN w złożonych zadaniach wizyjnych (Mitchell, 2019).

Sieci neuronowe w nauce

- **Fizyka:** analizy zderzeń cząstek subatomowych w wysokoenergetycznych zderzaczach cząstek - poprawiają metryki klasyfikacji aż o 8% w stosunku do najlepszych podejść, które były wcześniej używane (Baldi i inni; 2014) .
- **Fizyka:** Large Hadron Collider w CERN - do celów klasyfikacji i wykrywania anomalii w danych cząstek (Bhimji et al., 2018).
- **Badania farmaceutyczne** (Dahl et al., 2014; Wallach et al., 2015).
- **Matematyka stosowana:** problem trzech ciał. DNN mogą dostarczyć dokładnych rozwiązań przy stałym koszcie obliczeniowym i do 100 milionów razy szybciej niż konwencjonalne narzędzia (Breen et al., 2019).
- **Astronomia:** modelowania ruchów galaktyk (Ravanbakhsh et al., 2017) oraz klasyfikacji galaktyk (Kennamer et al., 2018).
- **Biologia:** klasyfikacjia sekwencji DNA (Bosco & Di Gangi, 2016)
- …
- **Socjologia?**

Aplikacja

Określenie warunków (inicjalizacja)	Generowanie danych Monte Carlo	Uczenie się głębokich sieci neuronowych	Oznaczanie braku równoważności
Zbadanie struktury i cech danych docelowych, czyli zbioru danych, w którym należy przetestować niezmienniczość pomiaru.	Symulacje Monte Carlo w celu wygenerowania danych odzwierciedlających sytuację pomiarową i potencjalne zakresy niezmienniczości (300 000)	Wygenerowane dane wykorzystywane do trenowania DNN. 	Wyznaczanie parametrów, które charakteryzują się brakiem równoważności

Weryfikacja symulacyjna

- Symulacje Monte Carlo i porównanie z alternatywnymi metodami:
  - CFA: Modification indices criteria (MI, *Lagrange multiplier* (Sarris, Satorra & Sorbom 1987, Sorbom, 1989).
  - CFA: Expected parameter change (EPC) indeks
  - CFA: Algorytm zaproponowany przez Asparouhova i Muthéna (2014)
  - Sekwencyjne porównania DIF za pomocą regresji logistycznej (procedura *ad hoc*)
- Dane wygenerowane zostały tak by odzwierciedlać strukturę danych ESS dla pytania o partycypację polityczną

	Warunek 1	Warunek 2	Warunek 3	Warunek 4
Grupy	4	10	4	10
Obserwacje	400	400	400	400
Wskaźniki	5	5	3	3

Warunek 3

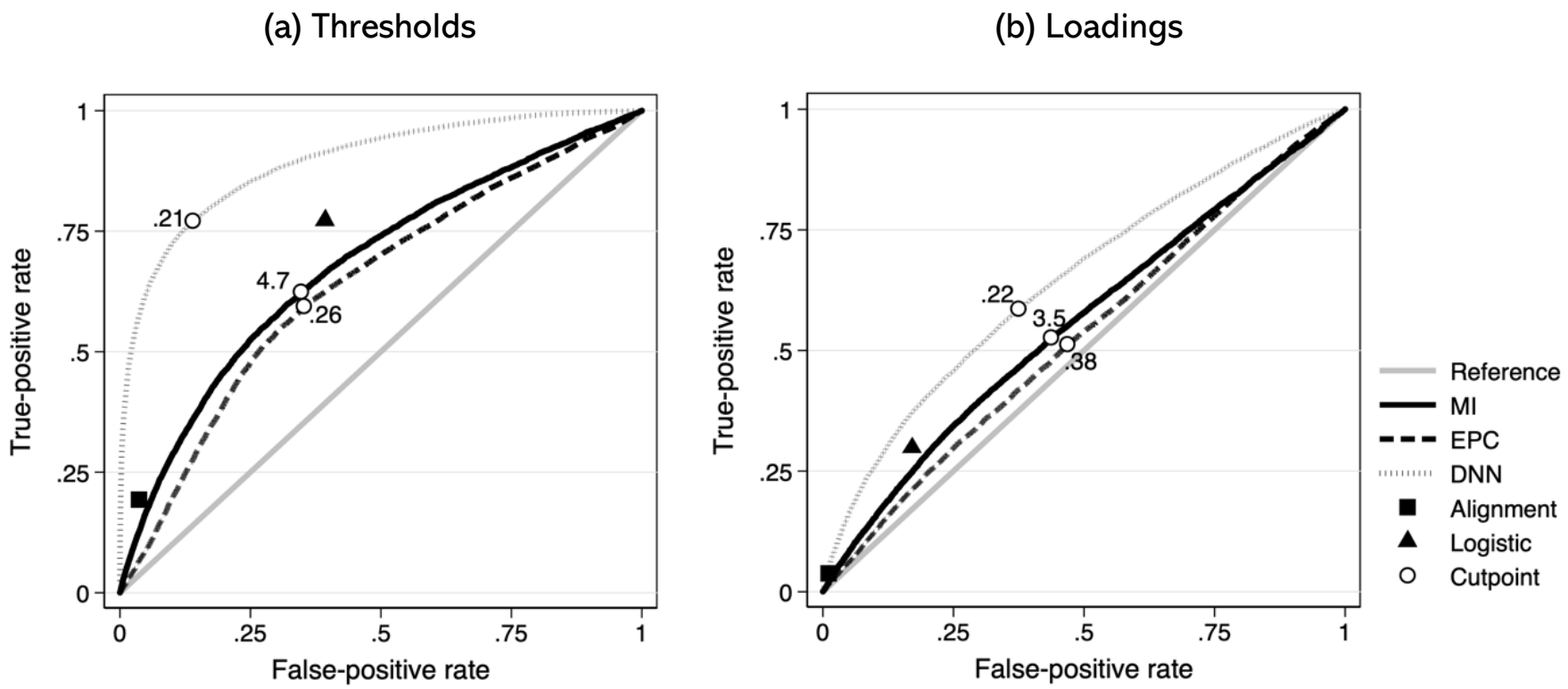


Figure 6. ROC curves for simulation study 3 (4 groups and 3 items)

$$y_{ig} = \tau_{ig} + \lambda_{ig}\theta + e_{ig}$$

Warunek 4

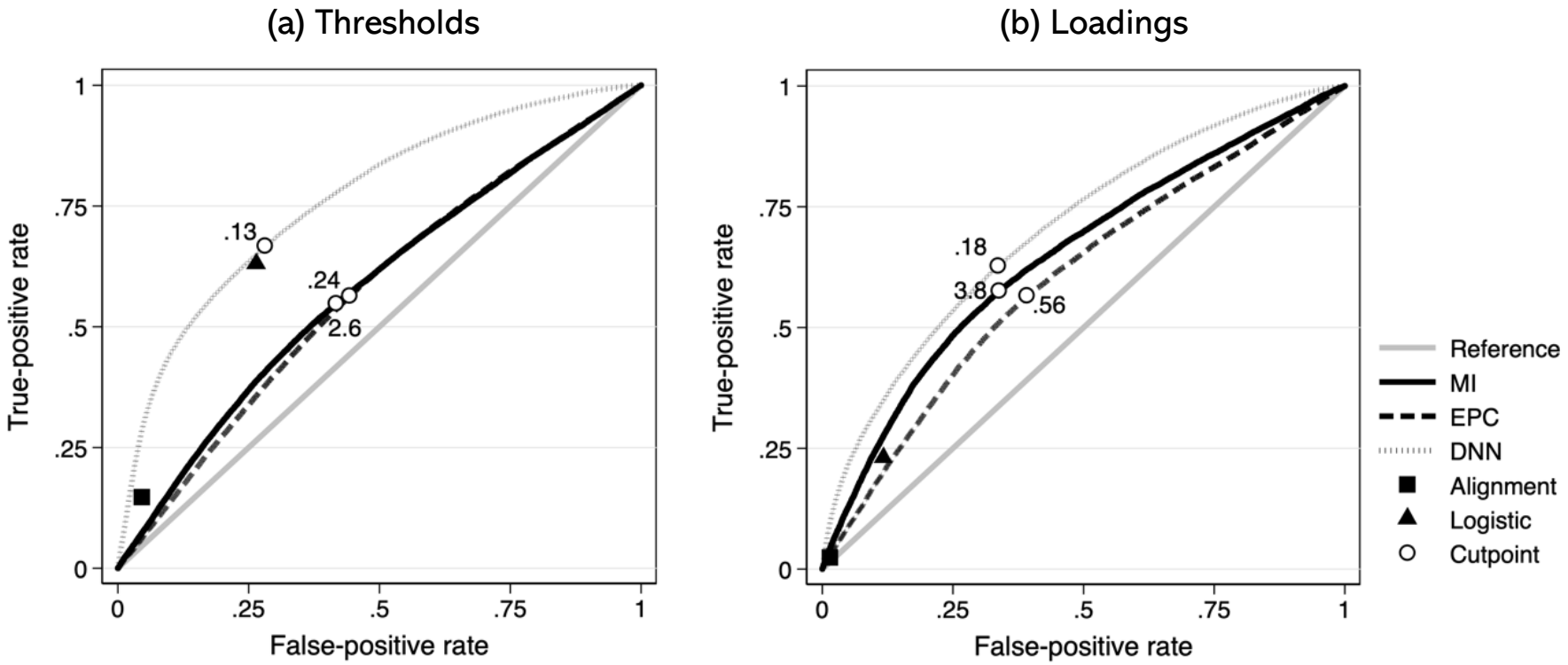


Figure 7. ROC curves for simulation study 4 (10 groups and 3 items)

$$y_{ig} = \tau_{ig} + \lambda_{ig}\theta + e_{ig}$$

Weryfikacja empiryczna

- Przykład empiryczny oparty jest na siódmej rundzie ESS (ESS 2014). Wykorzystaliśmy w nich pięć pozycji przeznaczonych do pomiaru uczestnictwa w życiu politycznym, w których respondenci byli pytani o to, czy:
  - Kontaktowali się z politykiem lub urzędnikiem państwowym w ciągu ostatnich 12 miesięcy (i1),
  - Działali w partii politycznej lub grupie działania w ciągu ostatnich 12 miesięcy (i2),
  - Działali w innej organizacji lub stowarzyszeniu w ciągu ostatnich 12 miesięcy (i3),
  - Podpisali petycję w ciągu ostatnich 12 miesięcy (i4)
  - Bojkotowali pewne produkty w ciągu ostatnich 12 miesięcy (i5).
- Użyliśmy siódmej rundy, ponieważ w jej ramach w słoweńskim kwestionariuszu popełniono błąd w tłumaczeniu, który został zgłoszony przez konsorcjum ESS - wyrażenie "**Pracował w innej organizacji lub stowarzyszeniu**" zostało przetłumaczone jako "**Pracował w innej politycznej organizacji lub stowarzyszeniu**".
  - Wyniki dla regresji logistycznej i DNN bardzo zbieżne
  - Obydwie metody wykryły nierównoważnoć stałych (*intercepts*) dla wskaźnika. Inne metody radziły sobie znacznie gorzej

Podsumowanie

- Procedury oparte na DNN i regresji logistycznej wydają się obiecujące i znacznie lepsze od dotychczasowych narzędzi
- DNN wydają się ciekawą propozycją dla metodologii i statystyki. Dalszy rozwój: style odpowiedzi (RS) i weryfikacji założeń różnych modeli (diagnozy modeli); badanie dopasowania
- Fazę generowania danych można uważać za słabą stronę tego podejścia, ale można również ją interpretować jako ciekawy sposób specyfikowania założeń narzędzia diagnostycznego (inne narzędzia też mają swoje założenia)
- Dostępność narzędzi opartych na DNN. Cała procedura jest bardzo wymagająca obliczeniowo ale wytrenowaną sieć można traktować jako pakiet statystyczny.

Dr hab. Artur Pokropek, Prof. IFiS PAN
Polska Akademia Nauk
Instytut Filozofii i Socjologii

Dziękuję za uwagę!

Pokropek, A. & Pokropek, E. (2021). Deep Neural Networks for Detecting Statistical Model Misspecifications. The Case of Measurement Invariance. *arxiv.* Preprint: https://arxiv.org/abs/2107.12757

This presentation has been prepared under the Scales Comparability in Large Scale Cross-country Surveys Project, which is funded by the Polish National Science Centre, as part of the grant competition Sonata 8 (UMO-2014/15/D/HS6/04934).