

Program mednarodne primerjave dosežkov učencev PISA

**ZBORNİK PRISPEVKOV O METODOLOŠKIH
VIDIKIH RAZISKAVE PISA**

Zbornik prevodov prispevkov iz publikacije

*Journal of Applied Measurement – Special issue: Programme for
International Student Assessment, Volume 8, Number 3, 2007*

Pedagoški inštitut

Uredila: Mojca Štraus

Izdal: Nacionalni center PISA, Pedagoški inštitut, Ljubljana

Prevod in priredba besedila: Tjaša Rajič

Strokovni pregled: Gašper Cankar

Oblikovanje naslovnice: Ana Korenini in Matej Perše

Oblikovanje besedila: Mateja Jagodič

Tisk: Tiskarna Pleško d.o.o.

Prva izdaja, prvi natis

Izid publikacije je del projekta Evalvacija vzgoje in izobraževanja v Republiki Sloveniji. Aktivnosti v okviru projekta omogoča sofinanciranje Evropskega socialnega sklada Evropske unije in Ministrstva za šolstvo in šport.

Objavljeno z dovoljenjem uredništva Journal of Applied Measurement.

Ljubljana, avgust 2008

CIP - Kataložni zapis o publikaciji
Narodna in univerzitetna knjižnica, Ljubljana

371.26(100)(082)

PROGRAM mednarodne primerjave dosežkov učencev PISA : zbornik prispevkov o metodoloških vidikih raziskave PISA / [uredila Mojca Štraus ; prevod in priredba besedila Tjaša Rajič]. - 1. izd., 1. natis. - Ljubljana : Pedagoški inštitut, 2008

ISBN 978-961-6086-63-9
1. Štraus, Mojca, 1968-
240513280

PREDGOVOR K SLOVENSKEMU PREVODU

Raziskava PISA je odmevna mednarodna raziskava držav članic Organizacije za ekonomsko sodelovanje in razvoj in držav partnerk. Do sedaj so bili izvedeni trije cikli raziskave v letih 2000, 2003 in 2006. V letu 2006 je bilo poleg Slovenije, ki je sodelovala prvič, v raziskavo vključenih že 57 držav. Program mednarodne primerjave dosežkov učencev PISA, kot raziskavo imenujemo v Sloveniji, je mednarodna raziskava o bralni, matematični in naravoslovni pismenosti 15-letnikov v sodelujočih državah. Raziskava išče odgovore na vprašanja o pripravljenosti mladih ob zaključku obveznega izobraževanja na vstop v njihovo nadaljnje izobraževanje, na vstop na trg delovne sile in uspešno vključevanje v družbo nasploh.

Krovno vlogo pri zasnovi raziskave PISA ima mednarodni svet raziskave, ki je sestavljen iz predstavnikov držav članic OECD in držav partnerk. Mednarodni svet raziskave odloča o njeni konceptualni zasnovi in ustreznosti načrta izvedbe in nadzoruje kakovost poteka izvedbe. Za koordinacijo izvedbe med sodelujočimi državami je v vsakem ciklu določen mednarodni center. Vsi trije do sedaj končani cikli raziskave so potekali pod koordinacijo združenja organizacij pod vodstvom avstralskega inštituta za raziskovanje v izobraževanju ACER (*Australian Council of Educational Research*). Raziskovalci inštituta ACER so priznani mednarodni strokovnjaki za izvajanje takšnih raziskav. Svoje izkušnje pri delu na raziskavi PISA so strnili v prispevke, ki so bili jeseni 2007 objavljeni v reviji *Journal of Applied Measurement*. Prispevki se nanašajo na metodološke vidike raziskave in analize ustreznosti posameznih pristopov pri objavi rezultatov na mednarodni ravni.

V nacionalnem centru raziskave PISA smo z dovoljenjem uredništva *Journal of Applied Measurement* pripravili prevod teh prispevkov in ga objavljamo v pričujočem zborniku. Upamo, da ga bomo na ta način lahko približali širši strokovni javnosti v Sloveniji. Ob tej priliki se Tjaši Rajić, Gašperju Cankarju, Ani Korenini, Mateju Perše in Mateji Jagodič zahvaljujem za strokovno in kvalitetno sodelovanje pri pripravi zbornika.

Mojca Štraus

nacionalna koordinatorica raziskave PISA, Pedagoški inštitut

PREDGOVOR K IZVIRNIKU

Noah in Eckstein v svoji publikaciji *Towards a Science of Comparative Education* iz leta 1967 ugotavljata, da je sistematična analiza mednarodnih podatkov o dosežkih učencev ključna za primerjalne študije na področju izobraževanja.

V štirih desetletjih od izdaje njune knjige je zbiranje in analiziranje mednarodnih podatkov o dosežkih učencev postajalo vse pogostejše, s študijami Mednarodne zveze za evalvacijo dosežkov v izobraževanju (International Association for the Evaluation of Educational Achievement – IEA) in Organizacije za ekonomsko sodelovanje in razvoj (Organisation for Economic Cooperation and Development – OECD) pa tudi vse bolj raznoliko. Program mednarodne primerjave dosežkov učencev PISA (Programme for International Student Assessment) poteka pod okriljem OECD in se osredotoča na 15-letne učence, ki zaključujejo obvezno šolanje, načrtuje in izvaja pa ga mednarodno združenje različnih organizacij pod vodstvom Avstralskega inštituta za raziskovanje v izobraževanju (Australian Council for Educational Research – ACER).

Vsak mednarodni preizkus znanja mora ustrezno pokrivati področja preverjanja, določena v izhodiščih raziskave. Raziskava PISA ne meri toliko doseganja določenih kurikularnih ciljev, temveč bolj ugotavlja, kako znajo učenci pridobljeno znanje uporabiti v različnih vsakodnevnih situacijah. V preizkus je treba vključiti ogromno snovi, zato so naloge povezane v sklope, vsak sklop pa se pojavi v več delovnih zvezkih. Vsakemu učencu je dodeljen naključno izbran delovni zvezek. Lestvice dosežkov učencev, ki omogočajo različne primerjave podatkov, so v raziskavi PISA sestavljene z uporabo Raschevega matematičnega modela teorije odgovora na postavko (oz. logit multinomialnega modela mešanih koeficientov). Model na podlagi odgovorov učencev hkrati določi oceno težavnosti posameznih nalog in zmožnosti učencev, ki so naloge reševali, in tako naloge kot učence razvrsti vzdolž lestvice dosežkov po lastnosti, ki jo želimo izmeriti. Podatke o zmožnostih učencev lahko nato povežemo z različnimi spremljajočimi dejavniki, kot sta spol in socialno-ekonomski položaj. S pomočjo ocen povprečnih vrednosti za populacijo in ostalih statističnih podatkov lahko primerjamo različne skupine učencev znotraj ene vzorčne populacije ali med posameznimi vzorci raziskave PISA. Tako zasnovana raziskava omogoča tudi spremljanje trendov dosežkov od enega cikla raziskave do drugega.

Ob izvajanju raziskave PISA se je porodila vrsta ključnih organizacijskih in psihometričnih vprašanj. Članki v tej publikaciji obravnavajo dve vrsti psihometričnih vprašanj, ki so pomembna za vsako primerjalno raziskavo o dosežkih učencev. Prva se nanašajo na primerljivost merskih instrumentov med državami: ali je preverjanje znanja enako veljavno v različnih nacionalnih kontekstih? Zlasti pomembna je primerljivost prevodov nalog v različne jezike, vpliv kulture na reševanje vprašalnikov o stališčih učencev in koliko se učenci iz posamezne države potrudijo pri reševanju preizkusa. Drugi niz vprašanj se nanaša na uporabo podatkov iz različnih ciklov raziskave, da bi lahko dolgoročno spremljali trende dosežkov. Pomembni vprašanji sta učinkovitost izenačevanja lestvic dosežkov s pomočjo veznih nalog in ocena napake pri ugotavljanju trendov dosežkov.

Grisay in soavtorji (2007) se lotevajo vprašanj o primerljivosti preizkusov v različnih jezikih in nacionalnih kontekstih. V svojem članku natančno opišejo postopke prevajanja in prirejanja nalog v raziskavi PISA ter določitve kriterijev za preverjanje njihove primerljivosti. Njihov zaključek je, da je preverjanje kognitivnih sposobnosti v raziskavi PISA (in v ostalih mednarodnih študijah) kulturno in jezikovno bližje učencem iz zahodnih držav, ki predstavljajo večino sodelujočih, kakor pa njihovim vrstnikom iz ostalih držav. Ocene pristranosti, ki jih navajajo, avtorje spodbudijo k podrobnejši analizi vzrokov in narave razlik v delovanju nalog. Walker (2007) preuči razlike v naravnosti učencev k izbiranju »skrajnih« odgovorov pri vprašanjih Likertovega tipa, ki ugotavljajo stališča učencev, in zaključí, da morebitne razlike med državami niso velike. Butler in Adams (2007) ugotavljata, da se učenci iz različnih držav pri reševanju preizkusa raziskave PISA podobno potrudijo, četudi se zdi, da so dosežki posameznih učencev odvisni od truda, ki ga vložijo. Navedene ugotovitve kažejo, da pri razvoju preizkusov znanja, ki bi bili primerljivi glede na težavnost, nekatera vprašanja ostajajo odprta, razlike v dosežkih med državami pa niso posledica večjega ali manjšega truda učencev ali naravnosti k izbiranju skrajnih odgovorov na vprašanja o njihovih stališčih.

Vprašanja o spremljanju trendov dosežkov s pomočjo mednarodnih primerjalnih raziskav so obravnavana v dveh člankih. Kazalniki trendov za oblikovalce politik predstavljajo pomembno informacijo o spremembah v ravneh dosežkov pri učencih posamezne države v primerjavi z ostalimi državami, pa tudi o zmanjševanju ali povečevanju razlik znotraj ene države, na primer tistih med skupinami z različnim socialno-ekonomskim položajem. Dva članka v tej publikaciji razkrivata, kako velik izziv je preučevanje kazalnikov trenda za raziskovalce. Gebhardt in Adams (2007) pokažeta, da so raziskovalni zaključki lahko odvisni od izbrane metodologije izenačevanja lestvic dosežkov (*equating methodology*), regresijskih modelov in načina povezovanja postavk. Uporaba pogojnih namesto marginalnih ocen lahko privede do drugačnih zaključkov o trendih dosežkov, tako na ravni države kakor tudi znotraj nje. Z uporabo pogojnih srednjih vrednosti

izravnamo razlike v značilnostih vzorcev (še zlasti razlike v trudu učencev) iz posameznih ciklov raziskave, vendar pa je to oblikovalcem politik težje razložiti.

Monseur in Berezner (2007) se ukvarjata z oceno napake pri izenačevanju lestvic dosežkov v tistih mednarodnih primerjalnih študijah, kjer so naloge povezane v sklope in razporejene tako, da se vsak sklop pojavi v več preizkusih, vsak učenec pa rešuje naključno izbran preizkus. Takšna ocena pokaže, ali so spremembe v dosežkih realne ali pa gre le za nihanja, ki jih povzroči vzorčenje, vračunavanje (*imputation*) in izenačevanje lestvic dosežkov. Z upoštevanjem napake izenačevanja, ki jo je v nekaterih primerih treba prišteti k napakam vzorčenja in vračunavanja, raziskava PISA prinaša pomembno novost za podobno obsežna preverjanja znanja. Ocenjevanje napake izenačevanja pa je kompleksen postopek, ki terja dodatne raziskave o učinkih hierarhične strukture nalog, končnih populacijah, politomnih nalogah in enakomernosti napake izenačevanja dosežkov.

Projekt PISA zbira pomembne mednarodne podatke o dosežkih učencev, s pomočjo katerih lahko preučujemo izobraževalne politike in prakse in dolgoročno spremljamo trende dosežkov s področja bralne, matematične in naravoslovne pismenosti učencev. Vse tovrstne raziskave pa se soočajo tudi s številnimi tehničnimi izzivi in vprašanji, ki lahko ogrozijo veljavnost mednarodnih primerjav. V tej publikaciji so zbrane poglobljene študije o nekaj tovrstnih tehničnih izzivih.

Geoffrey N. Masters

Avstralski inštitut za raziskovanje v izobraževanju

(*Australian Council for Educational Research*)

Viri

Noah, H.J. in Eckstein, M.Z. (1967). *Towards a science of comparative education*. New York: McMillan.

KAZALO

PROGRAM MEDNARODNE PRIMERJAVE DOSEŽKOV UČENCEV: KRATEK PREGLED Ross Turner, Raymond J. Adams	1
ENAKOVREDNOST PREVODOV ZA RAZISKAVO PISA V RAZLIČNIH DRŽAVAH Aletta Grisay, John H.A.L. de Jong, Eveline Gebhardt, Alla Berezner, Beatrice Halleux-Monseur	17
KAKO SE PRI MERJENJU STALIŠČ IZOGNITI KULTURNO POGOJENI NARAVNANOSTI K IZBIRANJU SKRAJNIH ODGOVOROV? Maurice Walker	43
VPLIV TRUDA UČENCEV NA IZIDE MEDNARODNIH RAZISKAV Jayne Butler, Raymond J. Adams	61
VPLIV METODOLOGIJE IZENAČEVANJA LESTVIC NA TRENDE, UGOTOVLJENE V RAZISKAVI PISA Eveline Gebhardt, Raymond J. Adams	99
IZRAČUN NAPAK IZENAČEVANJA LESTVIC DOSEŽKOV V MEDNARODNIH RAZISKAVAH NA PODROČJU IZOBRAŽEVANJA Christian Monseur, Alla Berezner	125

PROGRAM MEDNARODNE PRIMERJAVE DOSEŽKOV UČENCEV: KRATEK PREGLED

Ross Turner

Avstralski inštitut za raziskovanje v izobraževanju

Raymond J. Adams

Avstralski inštitut za raziskovanje v izobraževanju

Univerza v Melbournu

V tem prispevku bomo na kratko predstavili Program mednarodne primerjave dosežkov učencev (PISA), mednarodno primerjalno raziskavo o dosežkih 15-letnih učencev. Raziskava PISA poteka pod okriljem Organizacije za ekonomsko sodelovanje in razvoj (OECD), za obdobje 1998–2010 pa jo je zasnovalo in izvedlo mednarodno združenje različnih organizacij pod vodstvom Avstralskega inštituta za raziskovanje v izobraževanju (*Australian Council for Educational Research – ACER*).

Kaj je PISA?

Program mednarodne primerjave dosežkov učencev (PISA) je kot dolgoročno, periodično, mednarodno primerjalno raziskavo o matematični, naravoslovni in bralni pismenosti 15-letnih učencev v poznih 90. letih prejšnjega stoletja spodbudila Organizacija za ekonomsko sodelovanje in razvoj (OECD). Raziskavo PISA izvajajo in vodijo države članice OECD v sodelovanju z vse več državami *partnerkami*, ki niso članice OECD.

Podatki se v raziskavi PISA zbirajo vsake tri leta. Prvi cikel raziskave je bil izveden leta 2000, drugi pa leta 2003. Izsledki obeh raziskav so bili objavljeni v začetnih poročilih (OECD, 2001, 2003a, 2004b) in v vrsti tematskih in tehničnih poročil.¹ Tretji cikel raziskave je bil izveden leta 2006, naslednji pa je načrtovan za leto 2009. Pri vsakem ciklu raziskave je poudarek na eni izmed pismenosti (matematični, naravoslovni ali bralni); znanje in spretnosti učencev na ostalih dveh področjih se takrat preverjajo manj temeljito. V raziskavi leta 2000 je bil poudarek na bralni, leta 2003 na matematični in leta 2006 na naravoslovni pismenosti. Leta 2009 se bo ponovno podrobneje preverjala bralna pismenost.

Za zajem podatkov v letih 2000, 2003, 2006 in 2009 je odgovorno mednarodno združenje različnih organizacij pod vodstvom Avstralskega inštituta za raziskovanje v izobraževanju (*Australian Council for Educational Research – ACER*), ki ga vodi Ray Adams, mednarodni direktor raziskave PISA. Poleg Avstralskega inštituta za raziskovanje v izobraževanju pri izvedbi raziskave PISA na mednarodni ravni sodelujejo še nizozemski Nacionalni inštitut za merjenje znanja (*National Institute for Educational Measurement; Cito Group*), Nemški inštitut za mednarodno raziskovanje v izobraževanju (*DIPF*), Služba za eksperimentalno pedagogiko (*Service de Pédagogie Expérimentale*) z univerze v Liègeu, belgijska Družba za nadzor jezikovne kakovosti *CAPSTAN*, ameriški organizaciji *Westat* in Služba za testiranje v izobraževanju (*Educational Testing Service – ETS*) ter japonski Nacionalni inštitut za raziskovanje izobraževalnih politik (*National Institute for Educational Policy Research – NIER*).

Raziskava PISA temelji na izbiri učencev glede na starost, saj preverja znanje 15-letnikov, ki obiskujejo sedmi ali višji razred. V večini sodelujočih držav se ti učenci bližajo zaključku obveznega šolanja, prav tako pa so v skoraj vseh državah OECD takorekoč vsi učenci te starosti vključeni v redni izobraževalni sistem.

Raziskava PISA se osredotoča na področje *pismenosti* in na to, v kolikšni meri znajo učenci znanja in spretnosti, ki so jih pridobili v šoli, uporabiti v različnih situacijah in pri razreševanju problemov. PISA tako preverja, kako znajo učenci svoje spretnosti branja uporabiti za razumevanje in interpretacijo različnih besedil iz vsakdanjega življenja; kako se s pomočjo matematičnega znanja in spretnosti soočajo z različnimi izzivi in problemi, ki terjajo matematično znanje; in kako znajo

¹ Glej www.pisa.oecd.org.

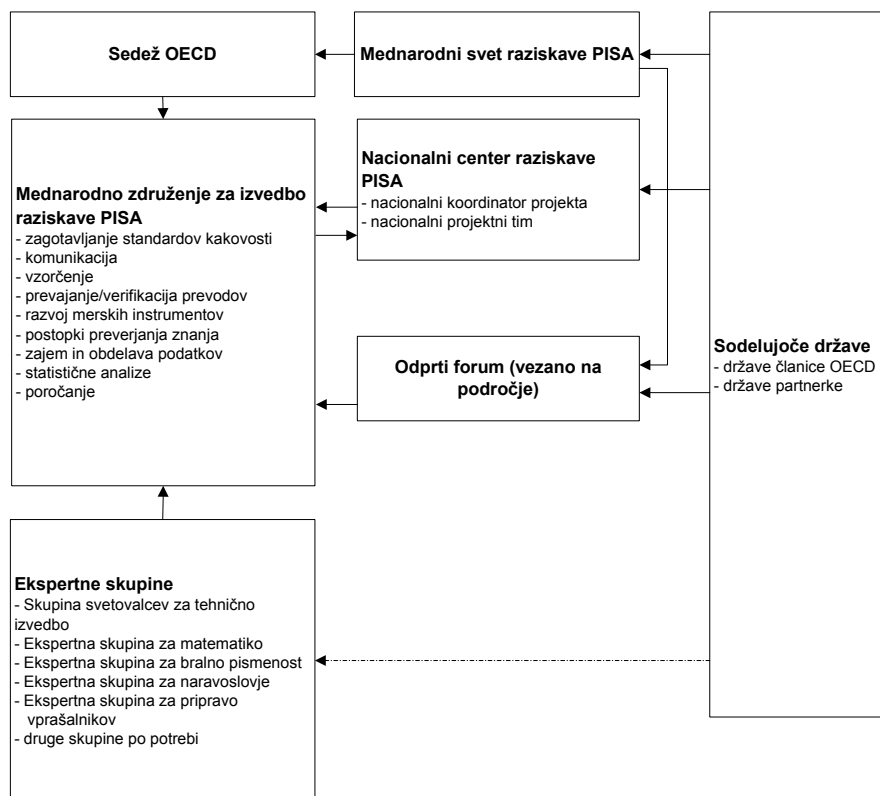
svoje naravoslovno znanje in spretnosti uporabiti za razumevanje, razlago in razreševanje različnih situacij in problemov s področja naravoslovja.

Raziskava PISA sodelujočim državam omogoča tudi občasno preverjanje drugih kurikularnih kompetenc po lastni presoji. Cikel raziskave 2003 je tako vključeval tudi preverjanje splošnih kompetenc učencev za reševanje problemskih nalog. V raziskavi PISA se poleg tega zbirajo tudi podatki o različnih vidikih domačega, družinskega in šolskega okolja učencev ter podatki šol o različnih vidikih organizacije in izobraževalnega procesa na šolah. Takšne informacije omogočajo podrobne študije raznih dejavnikov znotraj posameznih držav in med njimi, ki so povezani z različnimi ravni bralne, matematične in naravoslovne pismenosti 15-letnih učencev iz vsake države.

Organizacijske značilnosti raziskave PISA

Raziskavo PISA si lahko predstavljamo kot projekt, ki deluje na več ravneh. Različne ravni in ključne udeležence prikazuje shema 1.

Shema 1: Ključni udeleženci v projektu PISA



Raziskavo PISA nadzoruje sedež OECD v Parizu. Izhodišča projekta postavlja mednarodni svet raziskave PISA, ki deluje kot komisija OECD, sestavljena iz

predstavnikov in opazovalcev iz sodelujočih držav. Sestane se približno dvakrat letno. Raziskavo PISA torej vodijo izkušeni poznavalci izobraževalnih sistemov posameznih držav, ki med seboj sodelujejo.

V vsaki sodelujoči državi je vzpostavljen nacionalni center, vse aktivnosti na nacionalni ravni pa vodi nacionalni koordinator projekta (*National Project Manager – NPM*). Koordinator pri vprašanjih šolske politike, izvedbe raziskave in analiz ter poročanja o izidih raziskave, ki so za državo še posebej pomembni, ponavadi tesno sodeluje s predstavnikom svoje države v mednarodnem svetu raziskave PISA. V večini nacionalnih centrov majhna skupina ljudi skrbi za razvoj in potek projekta ter za poročanje o raziskavi na nacionalni ravni.

Na mednarodni ravni za izvedbo projekta skrbi pogodbeni izvajalec, ki ga imenuje OECD. Izvajalca se za vsak cikel raziskave izbere v odprtem javnem razpisu. V vseh štirih ciklih raziskave PISA doslej je bilo za izvajalca izbrano mednarodno združenje organizacij pod vodstvom Avstralskega inštituta za raziskovanje v izobraževanju in Raya Adamsa kot mednarodnega vodje raziskave. Mednarodni center je odgovoren za vse vidike izvedbe raziskave, vključno z:

- razvojem standardov kakovosti na vseh ravneh projekta;
- razvojem postopkov za zagotavljanje doseganja enakih standardov kakovosti pri vseh udeležencih ter mehanizmov za spremljanje kakovosti izvedbe projekta;
- spodbujanjem odprte in učinkovite komunikacije med vsemi udeleženci v projektu;
- tesnim sodelovanjem s predstavniki nacionalnih centrov, da bi lahko zbrali informacije o razvoju in izvajanju projekta v posameznih državah;
- razvojem izhodišč za preverjanje znanja;
- razvojem raziskovalnih in merskih instrumentov;
- razvojem in načrtovanjem vzorčenja;
- razvojem postopkov preverjanja znanja in pripravo dokumentacije;
- usposabljanjem ključnih zaposlenih v nacionalnih centrih za izvedbo raziskave;
- razvojem postopkov za zajem podatkov;
- zajemom in obdelavo podatkov preverjanja znanja;
- analizo rezultatov in pripravo gradiv, ki OECD pomagajo pri pripravi potrebnih poročil;
- podporo sedežu OECD s pregledom postopkov in načrtovanjem izboljšav za prihodnje cikle raziskave.

Da bi lahko izvršilo vse navedene naloge, mora mednarodno združenje za izvedbo raziskave PISA najprej imenovati široko skupino strokovnjakov z različnih področij. Nekatera od teh področij lahko pokrijejo strokovnjaki, ki delujejo v partnerskih organizacijah združenja. Poleg njih z mednarodnim združenjem

sodelujejo še številne druge ključne skupine in posamezniki: vrsta strokovnih skupin, v katerih delujejo mednarodno priznani strokovnjaki s področij, kot so specifični kognitivni vidiki preverjanja znanja, vzorčenje, statistične analize, razvoj vprašalnikov, in izkušeni prevajalci. Po potrebi se lahko vključi tudi strokovnjake z drugih področij.

Najbolj pomembno pa je, da mednarodno združenje za izvedbo raziskave PISA tesno sodeluje s posamezniki, ki delujejo na nacionalni ravni. Povezanost nacionalnih koordinatorjev projekta z mednarodnim združenjem je nujna za uspeh raziskave. Nacionalni centri seznanjajo mednarodno združenje s pogoji in ovirami pri izvedbi raziskave, dajejo povratne informacije o razvojnih rešitvah in svetujejo, na kakšen način bi raziskavo v določeni državi najbolje izvedli. Mednarodno združenje pa nacionalne centre obvešča o zahtevah projekta, jim pošilja delovno gradivo v pregled ter nudi usposabljanja in gradiva, ki jim olajšajo izvedbo raziskave.

Od časa do časa konzultacije potekajo tudi nekoliko drugače – npr. v obliki naravoslovnega foruma v raziskavi PISA 2006. Šlo je za odprt forum, v okviru katerega so sodelujoče države imele možnost imenovati svoje strokovnjake, ki so lahko neposredno zagovarjali interese in poglede svoje države pri nekaterih tehničnih vidikih raziskave. Na tak način je bilo strokovno razpravo mogoče zasnovati mnogo širše kakor le v okviru ekspertnih skupin mednarodnega izvajalca raziskave. Skupina naravoslovnega foruma je preučila prioritete in nekatera vprašanja o tem, kako zasnovati izhodišča za merjenje naravoslovne pismenosti, in pomembno prispevala k razvoju gradiv za preverjanje naravoslovne pismenosti in odnosa učencev do naravoslovja. Udeležence foruma so predlagali predstavniki posamezne države v mednarodnem svetu raziskave PISA. Podoben forum je bil organiziran tudi za matematiko v raziskavi PISA 2003; občasno se sestane forum za pripravo vprašalnikov, trenutno pa poteka forum za bralno pismenost v raziskavi PISA 2009.

Kdo sodeluje?

Vključno z raziskavo PISA 2009 je v eni ali več raziskavah PISA doslej sodelovalo 65 držav. Te države so našteje v tabeli 1.

Tabela 1: Sodelujoče države v raziskavi PISA

Ime nacionalnega centra	2000 ¹	2003	2006	2009 ²	Ime nacionalnega centra	2000 ¹	2003	2006	2009 ²
Albanija	✓			✓	Latvija	✓	✓	✓	✓
Argentina	✓		✓	✓	Liechtenstein	✓	✓	✓	✓
Avstralija	✓	✓	✓	✓	Litva			✓	✓
Avstrija	✓	✓	✓	✓	Luksemburg	✓	✓	✓	✓
Azerbajdžan			✓	✓	Makedonija	✓			
Belgija	✓	✓	✓	✓	Mehika	✓	✓	✓	✓
Brazilijska	✓	✓	✓	✓	Nizozemska	✓	✓	✓	✓
Bolgarija	✓		✓	✓	Nova Zelandija	✓	✓	✓	✓
Kanada	✓	✓	✓	✓	Norveška	✓	✓	✓	✓
Čile	✓		✓	✓	Peru	✓			
Kitajska (Šanghaj)				✓	Poljska	✓	✓	✓	✓
Hong Kong - Kitajska	✓	✓	✓	✓	Portugalska	✓	✓	✓	✓
Macao - Kitajska			✓	✓	Katar			✓	✓
Kitajski Tajpej			✓	✓	Južna Koreja	✓	✓	✓	✓
Kolumbija			✓	✓	Romunija	✓		✓	✓
Hrvaška			✓	✓	Rusija	✓	✓	✓	✓
Češka	✓	✓	✓	✓	Škotska	✓	✓	✓	✓
Danska	✓	✓	✓	✓	Črna Gora			✓	✓
Dominikanska republika				✓	Srbija		✓	✓	✓
Estonija			✓	✓	Singapur				✓
Finska	✓	✓	✓	✓	Slovaška		✓	✓	✓
Francija	✓	✓	✓	✓	Slovenija			✓	✓
Nemčija	✓	✓	✓	✓	Španija	✓	✓	✓	✓
Grčija	✓	✓	✓	✓	Švedska	✓	✓	✓	✓
Madžarska	✓	✓	✓	✓	Švica	✓	✓	✓	✓
Islandija	✓	✓	✓	✓	Tajska	✓	✓	✓	✓
Indonezija	✓	✓	✓	✓	Trinidad in Tobago				✓
Irska	✓	✓	✓	✓	Tunizija		✓	✓	✓
Izrael	✓		✓	✓	Turčija		✓	✓	✓
Italija	✓	✓	✓	✓	Velika Britanija	✓	✓	✓	✓
Japonska	✓	✓	✓	✓	ZDA	✓	✓	✓	✓
Jordanija			✓	✓	Urugvaj		✓	✓	✓
Kirgizistan			✓	✓					

¹ Vključno s sodelovanjem v raziskavi PISA plus (OECD 2003a), ki je ponovitev raziskave PISA 2000 v letu 2001.

² Seznam sodelujočih držav za leto 2009 je okviren.

Tehnične značilnosti raziskave PISA

Tehnične značilnosti raziskave PISA vključujejo številne elemente. Sestava preizkusa znanja in njegove značilnosti sta ključna pri razvoju preizkusov za raziskavo PISA. Drugi pomembni element je načrt in postopek vzorčenja, tako šol kakor učencev znotraj posamezne šole. Veliko tehničnih vprašanj je povezanih tudi z večjezikovno naravo preizkusov, kar vključuje pravila in postopke za zagotavljanje primerljivosti preizkusov v različnih jezikih, in sicer tako med državami kot znotraj posamezne države, pri čemer je treba upoštevati tudi različne kulturne kontekste teh držav. Pozornost je treba nameniti tudi številnim operativnim postopkom, kot so priprava na izvedbo preverjanja znanja, zajem in obdelava podatkov ter mehanizmi za zagotavljanje kakovosti, s pomočjo katerih v vseh državah pridobimo primerljive podatke. Sledijo tehnična opravila sestavljanja lestvic dosežkov in analiz podatkov ter poročanje o rezultatih. Lestvice dosežkov v

raziskavi PISA so zasnovane po metodologiji Raschevih modelov; prav tako tudi uporaba lestvic dosežkov kot osnovnega orodja za poročanje o rezultatih raziskave PISA. V nadaljevanju bo vsako od navedenih tehničnih področij na kratko predstavljeno. Podrobnejše informacije o tem so na voljo v tehničnih poročilih raziskave PISA iz let 2000 in 2003 (Adams in Wu, 2002, OECD 2005a).

Standarde kakovosti pri vsakem od navedenih tehničnih področij se definira, nadzoruje in zagotavlja z upoštevanjem vrste tehničnih standardov, ki jih potrdi mednarodni svet raziskave PISA in so osnova in zagotovilo za kakovostno izvedbo projekta v vsaki od sodelujočih držav.

Sestavljanje in razvoj preizkusov znanja

V raziskavi PISA so bili doslej uporabljeni le klasični preizkusi na papirju (*pencil-and-paper tests*).² Kognitivni del preverjanja znanja s področij bralne, naravoslovne in matematične pismenosti traja dve uri. V raziskavi PISA 2006 je kognitivni del preizkusa vključeval tudi vprašanja, ki so ugotavljala odnos učencev do naravoslovja. Poleg tega učenci izpolnijo tudi kratek vprašalnik, s katerim se zbirajo spremljajoči podatki o osebnih okoliščinah učenca, njegovih mnenjih, preferencah in željah, o njegovem domačem in družinskem okolju, pa tudi o šolskem okolju. Vprašalnik učenci izpolnijo v 20–30 minutah. Kratek splošni vprašalnik o šolskem okolju izpolnijo tudi ravnatelji.

Kognitivni del preizkusa mora na ustrezen način pokrivati vsa področja, ki se preverjajo, s preizkusom pa je treba pridobiti primerne podatke skladno z izhodišči in namenom preverjanja (OECD 2006). Povedano drugače, razvoj merskih instrumentov za preverjanje znanja se začne z nizom natančno opredeljenih ciljev, ki jih želimo preveriti s posameznimi postavkami preizkusa. Cilji pokrivajo različne vidike učne snovi ter vrsto učnih procesov in kontekstov, v katerih so prikazane naloge preizkusa, tako da nobena tema ne prevladuje in so učenci z različnimi interesi in izkušnjami ter iz različnih nacionalnih okolij v enakopravnem položaju. Preizkusi v raziskavi PISA so sestavljeni tako, da ugotavljajo, v kolikšni meri znajo učenci pri reševanju problemov v različnih situacijah uporabiti znanja in spretnosti, ki so jih pridobili v šoli. Poudarek ni na preverjanju določenih kurikularnih ciljev, temveč na ugotavljanju, kako znajo učenci pridobljeno znanje uporabiti v vsakodnevnih situacijah.

V vsakem ciklu raziskave je pripravljenih ogromno nalog, kar zagotavlja ustrezno pokritost in uravnoteženost izhodišč vsakega področja in vidika, ki se preverja. Posamezne naloge tvorijo en sklop, vsak sklop pa se pojavi v več delovnih zvezkih. Vsakemu učencu je dodeljen naključno izbran delovni zvezek.

Preden so naloge dokončno sestavljene in vključene v preizkus, se v postopku priprave preizkusa preveri njihove karakteristike s pomočjo Rascheve analize (*Rasch analysis*). Leto pred vsakim ciklom glavne raziskave se v obširni

² V raziskavi PISA 2006 je bilo mogoče naravoslovno pismenost preveriti tudi s pomočjo računalnika. To možnost so izkoristile tri države.

predraziskavi izvede pretestiranje vseh nalog, ki naj bi bile vključene v banko nalog za glavno raziskavo. Odgovore učencev se analizira v aplikaciji ConQuest (Wu, Adams in Wilson, 1998). S pomočjo klasičnih statističnih podatkov o nalogah v aplikaciji ConQuest, podatkov o ujemanju rezultatov preverjanja z Raschevim modelom in različnih kazalnikov se preveri, kako se obnese posamezna naloga. Ti podatki vključujejo indekse diskriminativnosti in ujemanja z modelom, točkovne biserialne korelacije, povprečne zmožnosti učencev po različnih kategorijah odgovorov, pregled razvrstitve različnih kategorij odgovorov pri nalogah, kjer se upošteva tudi delno pravilne odgovore, konsistentnost indeksov po državah, pričakovane in dejanske krivulje dosežkov po spolu in državah ter pričakovane in dejanske karakteristične krivulje nalog glede na različne kategorije odgovorov.

Navedene statistične analize so osnova za vključitev posameznih nalog v glavno raziskavo, uporabijo pa se tudi za določitev nalog, ki so zaradi določenih karakteristik neprimerne za preverjanje znanja v raziskavi PISA, ali nalog, ki jih je treba dodatno pregledati (na primer zato, ker podatki o dosežkih kažejo, da je bil prevod naloge neustrezen). Posebej pomembni so podatki o težavnosti nalog, saj morajo preizkusi vsebovati različno težke naloge iz vsakega področja preverjanja.

V raziskavi PISA so uporabljena vprašanja različnih tipov: vprašanja izbirnega tipa, odprta vprašanja s kratkim odgovorom in odprta vprašanja z dolgim odgovorom. Vprašanja izbirnega tipa so bodisi klasična, z določenim številom možnih odgovorov (navadno s štirimi), izmed katerih učenci izberejo najustreznejšega, ali kompleksna vprašanja izbirnega tipa, pri katerih je navedenih več trditev, za vsako izmed njih pa učenci izberejo enega od dveh ali več odgovorov (drži/ne drži, pravilno/nepravilno itd.). Odprta vprašanja s kratkim odgovorom so lahko zaprtega tipa, kjer so učenci pri oblikovanju svojih odgovorov zelo omejeni (na primer matematične naloge, kjer je odgovor le število), ali vprašanja, kjer rešitve zapišejo z besedo ali kratko povedjo. Slednje so podobne vprašanjem s kratkim odgovorom odgovorov zaprtega tipa, le da je tukaj nabor možnih odgovorov večji. Pri odprtih vprašanjih z dolgim odgovorom je možnih pravih odgovorov veliko več. Navadno zahtevajo več pisanja, prikaz izračuna ali sklepanja, učenci pa morajo svoje odgovore pogosto tudi razložiti ali utemeljiti.

V prejšnjih ciklih raziskave PISA so sestavljavci preizkusov in strokovnjaki pokazali, da lahko z vprašanji odprtega tipa merimo veliko več različnih kognitivnih procesov kakor le z vprašanji zaprtega tipa, kot so vprašanja izbirnega tipa ali vprašanja s kratkim odgovorom. Podobne so bile tudi ugotovitve na podlagi povratnih informacij iz posameznih držav v fazi priprave nalog. Mednarodni svet raziskave PISA zastopa stališče, da večja raznolikost podatkov, ki jih pridobimo z uporabo nalog različnih tipov, upraviči dodatne stroške kodiranja in vrednotenja odgovorov. Do podobnih ugotovitev so prišli tudi v študiji o uporabi različnih tipov vprašanj na podlagi podatkov iz raziskave PISA (Routitsky in Turner, 2003), ki je pokazala, kako pomembno je v preizkus vključiti različne tipe vprašanj, če želimo z njimi izmeriti znanje učencev z različnimi zmožnostmi, ki sodelujejo v raziskavi PISA.

Odgovore učencev pri več kakor polovici vprašanj kognitivnega dela preizkusa je bilo v glavni raziskavi PISA 2006 mogoče računalniško obdelati. Pri preostalih 80 vprašanjih od skupno 185 (tj. 43 odstotkov) je bilo treba pred tem izvesti še kodiranje odprtih vprašanj.

Za vprašalnike o spremljajočih dejavnikih je bil izbran skupni nabor vprašanj. Z vprašalniki se ugotavlja socialne, kulturne, ekonomske in izobraževalne dejavnike, ki so povezani z dosežki učencev. Tako je mogoče raziskati povezave med dosežki učencev pri kognitivnem delu preizkusa in različnimi dejavniki v zvezi z učencem ali šolo ter ugotoviti, kako se ti dejavniki razlikujejo v različnih izobraževalnih sistemih, državah in med leti.

Vzorčenje v raziskavi PISA

Vzorčenje v raziskavi PISA poteka v dveh korakih, in sicer tako, da je verjetnost, da bo učenec iz ciljne skupine vključen v vzorec, znana vnaprej. Določitev mednarodne populacije omogoča pripravo izhodišč za vzorčenje, ki vključuje vse 15-letne učence, ki obiskujejo sedmi ali višji razred. Najprej se opravi naključno vzorčenje šol, pri čemer je verjetnost, da bo neka šola izbrana, sorazmerna z velikostjo šole. V drugem koraku se na vsaki izbrani šoli naključno izbere še vzorec 35 učencev. Tako je naključni vzorec izbranih učencev za raziskavo PISA zares reprezentativen in v vsaki od sodelujočih držav dobro pokriva populacijo 15-letnikov, vključenih v redno izobraževanje.

V vsaki državi je izbranih vsaj 150 šol (ali vse šole, če je manj kot 150 takšnih, ki jih obiskujejo učenci, ki ustrezajo pogojem raziskave). Načrtovana velikost vzorca 35 učencev na šolo pomeni, da bo v raziskavo vključenih vsaj 5250 učencev iz vsake države, preizkus pa bo predvidoma opravljalo vsaj 4500 učencev. Če je v več šolah manj kot 35 učencev, ki izpolnjujejo pogoje za sodelovanje v raziskavi, se vključi še dodatne šole, da je zajet dovolj velik vzorec učencev.

Standardi zagotavljajo: i) ustrezno pokritost ciljne populacije (vključno s strogimi pravili, katere šole ali učence je mogoče izločiti iz raziskave); ii) primerno točnost in natančnost ocen na podlagi vzorca (vključno s strogimi pravili o zahtevani velikosti vzorca); in iii) ustrezno odzivnost šol in učencev (vključno s strogimi pravili in postopki za morebitno zamenjavo neodzivnih šol, da bi tako dosegli zadostno odzivnost, ter s pravili o upoštevanju ali neupoštevanju podatkov učencev glede na njihovo odzivnost).

Prevajanje in kulturna ustreznost gradiv v raziskavi PISA

PISA je ena največjih tovrstnih mednarodnih študij doslej. Raziskava PISA 2006 je na primer vključevala preverjanje znanja na vsaj 150 šolah v vsaki od 57 sodelujočih držav, pripravljenih pa je bilo 81 nacionalnih različic merskih instrumentov v 42 jezikih. Pri takšnih projektih je zagotavljanje primerljivosti vseh različic preizkusov zahtevna naloga. Da bi bili pri tem uspešni, moramo najprej zagotoviti, da so gradiva za preverjanje znanja primerna za uporabo v kulturno tako raznolikih okoljih ter da so vse različice tega gradiva enakovredne.

Če želimo zagotoviti kulturno ustreznost gradiv, moramo uporabiti raznolik material, ki predstavlja različne kulturne izkušnje in okolja, gradivo pa je treba nato obdelati in izpopolniti na tak način, da so različna interesna področja enakomerno zastopana in da so gradiva primerna za učence iz vseh držav. Mehanizmi, ki zagotavljajo, da so gradiva za raziskavo PISA kulturno ustrezna, so:

- Gradiva za preizkuse (vprašalnik in naloge) so izbrana iz kar največ različnih virov, potrdijo pa jih vse sodelujoče države. Postopke za pripravo preizkusov nadzirajo mednarodni eksperti z določenega področja, vodijo pa jih strokovnjaki za sestavljanje preizkusov iz različnih držav in kulturnih okolij.
- Sodelujoče države imajo v postopku priprave večkrat možnost pregledati naloge in podati svoje mnenje. Gradivo pregledajo tudi mednarodne ekspertne skupine za posamezno področje preverjanja.
- Pri pripravi gradiv izvedemo t.i. »kognitivni laboratorij« in druge pilotne aktivnosti z učenci iz več različnih držav.
- V široko zasnovani predraziskavi se na učencih iz vseh sodelujočih držav preveri, kako se obnašajo naloge, ki jih razvijamo.
- Statistične analize podatkov iz predraziskave omogočajo, da se gradiva za izvedbo raziskave v vseh sodelujočih državah empirično ovrednoti in da se morebitna odstopanja v dosežkih posameznih držav pravočasno odkrijejo.
- Vsa gradiva se ponovno pregleda v luči informacij, ki jih pridobimo z vsakim od navedenih mehanizmov, v glavno raziskavo PISA pa so vključene le dovolj kvalitetne naloge in ostala gradiva.
- Končno različico nalog in ostalih gradiv pregledajo nacionalni koordinatorji raziskave, potrdi pa jo mednarodni svet raziskave PISA.

Vzporedno z razvijanjem kulturno ustreznih gradiv za merske instrumente raziskave PISA poteka tudi priprava originalnih gradiv v angleščini in francoščini, na podlagi katerih nastanejo ustrezne nacionalne različice. Iz originalnega gradiva nacionalni centri v vsaki sodelujoči državi sami pripravijo nacionalne izvode preizkusov, vendar se morajo pri tem držati strogo nadzorovanih postopkov dvojnega neodvisnega prevajanja, strokovnega usklajevanja obeh različic prevodov ter neodvisne mednarodne verifikacije končnega prevoda.

Prevedena gradiva vsake države se preizkusi v obsežni predraziskavi, enakovrednost prevodov v različne jezike pa se z analizo podatkov predraziskave tudi empirično preveri. Podatke o značilnostih posameznih nalog v različnih jezikih se pridobi z že omenjenimi Raschevimi analizami. Posebej pomembni so podatki o značilnostih naloge po posameznih državah (oblika analize razlik v delovanju nalog/DIF oz. *differential item functioning analysis*), ki razkrijejo naloge, katerih značilnosti se razlikujejo glede na jezik ali kulturo, v katerih so bile predstavljene. Instrumenti za merjenje znanja v glavni raziskavi vsebujejo le naloge, s pomočjo katerih bomo lahko določili mednarodno enotne parametre težavnosti.

S pomočjo analiz podatkov iz predraziskave se opravi pregled gradiv, končna različica gradiva za glavno raziskavo PISA pa je določena tako, da so izbrane le brezhibne naloge ne glede na državo ali jezik, v katerega so bile prevedene. Ko je ta postopek zaključen, je vsaka nacionalna različica gradiva jezikovno in psihometrično karseda enakovredna originalnemu gradivu v angleščini ali francoščini, s čimer je moč določiti mednarodno enotne parametre nalog.

Izvedbeni postopki

Drugi način, kako zagotoviti primerljivost rezultatov preverjanja med tako raznolikimi državami in okolji, je standardizacija postopkov izvedbe preverjanja znanja. Razviti in zapisani so bili številni postopki, ki sodelujočim državam pomagajo izvesti preverjanje na tak način, da so dobljeni podatki mednarodno primerljivi.

V priročniku za nacionalne koordinatorje projekta so opisani tudi vsi postopki, ki jih morajo razviti in izvesti nacionalni centri raziskave PISA, vključno s posvetovanjem in pregledovanjem postopkov, izvedbo vzorčenja, postopki priprave in distribucije gradiva za preverjanje znanja, iskanjem in usposabljanjem sodelavcev, ki izvajajo preverjanje na šolah, in sistemi nadzora njihovega dela. Priročnik vsebuje tudi smernice za izvedbo postopkov spremljanja kakovosti, iskanja in usposabljanja koderjev odprtih odgovorov, vodenje postopkov kodiranja in vnosa podatkov, postopkov zajema podatkov in priprave vseh podatkov raziskave PISA za posredovanje mednarodnemu centru ter sodelovanje pri pripravi analiz in poročil.

V posebnih priročnikih so opisani postopki za izvedbo vzorčenja, prevajanje gradiv, izvedbo in koordinacijo raziskave na šolah, kodiranje odprtih odgovorov v preizkusih, obdelavo podatkov in napotki za uporabo posebne programske opreme za zajem podatkov. Poleg tega mednarodni center raziskave PISA organizira številne sestanke nacionalnih koordinatorjev projekta in drugih ključnih sodelavcev nacionalnih centrov, kjer jih dodatno usposablja in jih seznaja s postopki izvedbe preverjanja in z ostalimi vidiki izvedbe raziskave v posameznih državah.

Mednarodni center raziskave PISA z dokumentiranjem vseh izvedbenih postopkov, rednimi izobraževalnimi in informativnimi sestanki ter stalno komunikacijo preko elektronske pošte in telefona omogoča sodelujočim državam, da dosledno izpeljejo postopke preverjanja znanja. Nadzor izvedbe preverjanja v posameznih državah opredeljujejo postopki za spremljanje kakovosti, ki so opisani v nadaljevanju.

Spremljanje kakovosti

Izvedbo projekta PISA določa vrsta standardov kakovosti, ki opredeljujejo določanje ciljne populacije, vzorčenje, jezik preizkusa, pripravo preizkusov in priročnikov, izvedbo preverjanja znanja, kakovost tiska gradiv za preverjanje, tajnost gradiv itd.

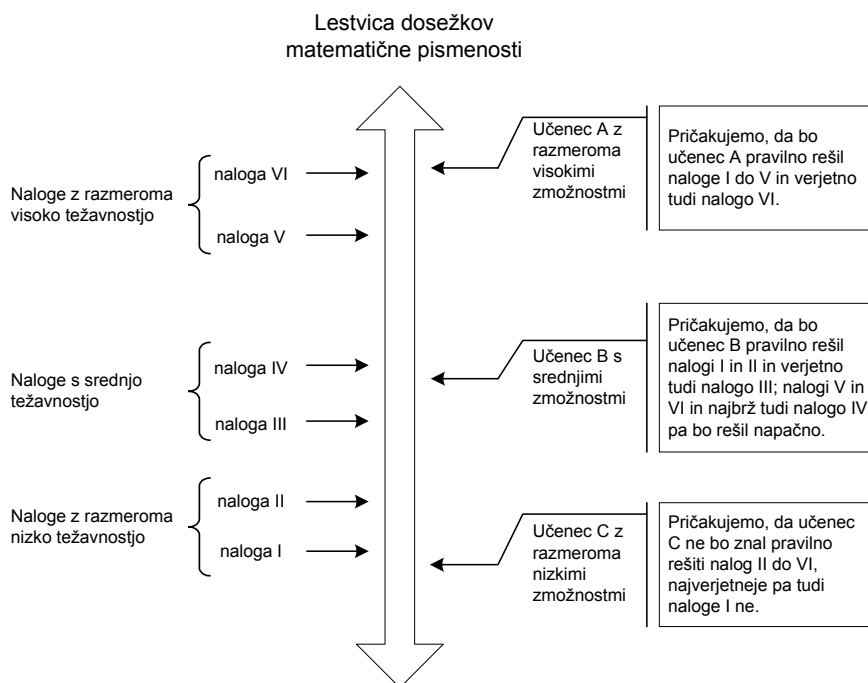
Za vsako od navedenih področij je v posebnih načrtih zapisano, kako sodelujoče države zadostijo postavljenim standardom. Nacionalni centri po posebnih kontrolnih postopkih stalno preverjajo, ali je izvedba raziskave skladna s postavljenimi standardi kakovosti. Po pravilih zunanjega nadzora lahko mednarodni center posreduje v primeru suma, da določena država ne dosega postavljenih standardov, ali takrat, ko je neupoštevanje standardov v neki državi že potrjeno – v tem primeru ugotovi, ali je zbrane podatke kljub temu mogoče uporabiti.

Mednarodni center tudi presodi, ali so podatki za neko državo primerni za vključitev v mednarodno bazo podatkov. Če so podatki neustrezni, se posamezno državo lahko izključi iz nekaterih ali celo vseh mednarodnih poročil o rezultatih raziskave PISA. V postopku presoje o primernosti podatkov sodelujejo izkušeni člani mednarodnega združenja za izvedbo raziskave PISA, skupina svetovalcev za tehnično izvedbo raziskave, vključno s strokovnjakom za vzorčenje, in sedež OECD, končno odločitev pa na podlagi njihovega mnenja sprejme mednarodni svet raziskave PISA.

Oblikovanje lestvic dosežkov in analiza podatkov

Oblikovanje lestvic dosežkov v raziskavi PISA temelji na preprostih predpostavkah: da ena ali več lastnosti, ki jih želimo izmeriti (te lastnosti so opredeljene v izhodiščih merjenja dosežkov na različnih področjih pismenosti), tvorijo kontinuum ali lestvico; da lahko s primernimi nalogami izmerimo vsako od teh lastnosti (glede na to, v kolikšni meri morajo učenci posedovati to lastnost, da lahko nalogo rešijo); in da lahko iz učenčevih odgovorov razberemo, v kolikšni meri je ta lastnost pri njem izražena. Lestvico dosežkov si lahko predstavljamo kot črto, na kateri lastnost, ki jo merimo, v eno smer narašča in v drugo upada. Naloge lahko razvrstimo vzdolž te črte glede na to, koliko neke lastnosti ali zmožnosti zahtevajo od učenca pri reševanju. Na enak način lahko vzdolž lestvice razvrstimo tudi učence glede na to, v kolikšni meri je neka lastnost ali zmožnost pri njih izražena. S pomočjo Rascheve analize lahko oblikujemo intervalne lestvice dosežkov. Shema 2 prikazuje lestvico dosežkov matematične pismenosti; na enak način se lestvice oblikujejo tudi za ostala področja pismenosti.

Shema 2: Lestvica dosežkov pismenosti v raziskavi PISA



Lestvice dosežkov učencev, s pomočjo katerih lahko OECD primerja podatke o dosežkih in o njih poroča, so v raziskavi PISA sestavljene z uporabo Raschevega matematičnega modela teorije odgovora na postavko. Gre za logit multinomialni model mešanih koeficientov (*mixed coefficients multinomial logit model*), ki je posplošena različica Raschevega modela. Model na podlagi odgovorov učencev hkrati določi oceno težavnosti posameznih nalog in zmožnosti učencev, ki so naloge reševali, in tako naloge kot učence razvrsti vzdolž lestvice dosežkov po lastnosti, ki jo želimo izmeriti. Model in metode oblikovanja lestvic dosežkov so natančneje opisani v tehničnem poročilu PISA 2000 (*PISA 2000 Technical Report*; Adams in WU, 2002) in nekaterih člankih (Adams in Wu, 2006; Adams, Wu in Carstensen, 2006).

Z uporabo opisanega modela lahko v raziskavi PISA odgovore učencev oziroma njihove zmožnosti, ki jih iz odgovorov razberemo, z regresijskimi enačbami povežemo tudi z različnimi spremljajočimi dejavniki, kot so spol, socialno-ekonomski položaj ipd. S takšnimi analizami lahko dobimo ocene povprečnih vrednosti za populacijo in ostale statistične podatke, ki omogočajo primerjave različnih skupin učencev znotraj ene vzorčne populacije ali med posameznimi vzorci raziskave PISA. Primerjave skupin učencev, kjer njihove dosežke na kognitivnem delu preizkusa povezujemo z različnimi spremljajočimi spremenljivkami, raziskovalcem šolske politike omogočajo vpogled v dejavnike, ki lahko vplivajo na dosežke učencev, in spodbudijo razmislek o potrebnih ukrepih vodstev šol in drugih organizacij v šolstvu, o praksah poučevanja in učenja ipd.

Merski instrumenti za raziskavo PISA, vključno z oblikovanjem lestvic dosežkov in analizami podatkov, so oblikovani tako, da omogočajo spremljanje trendov pri dosežkih učencev po različnih ciklih raziskave. Nekaj nalog iz raziskave PISA 2000 je bilo zato uporabljenih tudi pri preverjanju v letu 2003, del nalog iz raziskave PISA 2003 pa se lahko ponovi tudi v ciklu raziskave leta 2006. Na podlagi rezultatov pri teh nalogah lahko ugotovljamo, kako se dosežki učencev spreminjajo iz leta v leto. Pričakovane razlike med enim in naslednjim ciklom raziskave so sicer zanemarljive, v daljšem časovnem obdobju pa izboljšave v šolski politiki in tehnikah poučevanja lahko pripeljejo do opaznejših nihanj dosežkov. V prvem ciklu raziskave leta 2000 je bil poudarek na bralni pismenosti. Če so rezultati tega cikla ali kakšni drugi dejavniki spodbudili spremembe v šolski politiki in praksi, lahko tudi v raziskavi PISA 2009, ko bo glavno področje preverjanja ponovno bralna pismenost, pričakujemo spremembe pri dosežkih učencev.

Poročanje o rezultatih raziskave PISA

Po vsakem ciklu raziskave OECD pripravi obsežno poročilo o dosežkih učencev na mednarodni ravni. Doslej sta bili objavljena dve takšni poročili, in sicer za raziskavo PISA 2000 (OECD, 2001) in za raziskavo PISA 2003 (OECD, 2004b).

Glavna mednarodna poročila vsebujejo kratko predstavitev projekta PISA, ključne značilnosti področja pismenosti, na katerem je poudarek (bralna pismenost v poročilu 2000, matematična v poročilu 2003), in smernice za interpretacijo rezultatov glede na postavljena izhodišča, navzkrižne primerjave dosežkov glavnega področja pismenosti med državami in njihovo povezanost s spremljajočimi dejavniki. V poročilo je vključen tudi pregled dosežkov na ostalih dveh področjih pismenosti in pregled podatkov o spremljajočih dejavnikih na ravni učencev in šol ter razlike v spremljajočih dejavnikih med državami. Poročila vsebujejo tudi mednarodne primerjave učnega in šolskega okolja in razmišljanja o tem, kakšen vpliv ima lahko raziskava PISA na izobraževalno politiko.

V poročilih o dosežkih na glavnem področju pismenosti je poudarek na profilih dosežkov učencev v vsaki državi glede na skupno lestvico dosežkov in podrobnejših lestvicah dosežkov, kot je opredeljeno v izhodiščih. Pri matematični pismenosti so na primer rezultati prikazani na skupni lestvici matematičnih dosežkov, poleg tega pa tudi na podrobnejših lestvicah dosežkov na vsakem od štirih vsebinskih področij matematične pismenosti. Za prikaz profila dosežkov učencev glede na raven, ki jo na lestvici dosežajo, je ključnega pomena opis pričakovanih zmožnosti učencev, uvrščenih na določeno raven. Opisi matematičnih zmožnosti jasno povedo, kakšna znanja morajo učenci uporabiti pri reševanju nalog, in večjo matematično pismenost pojasnjujejo z večjo sposobnostjo učencev, da prikažejo in uporabijo svoje znanje. V mednarodnem poročilu OECD je izpostavljen zlasti relativni delež učencev, ki se v posamezni državi umeščajo na določeno raven lestvice dosežkov.

Pri poročanju o spremljajočih dejavnikih (osebne, družinske in šolske okoliščine) se upošteva razlike med državami na teh področjih, z analizo povezave

med spremljajočimi dejavniki in dosežki učencev pa se poskuša ugotoviti, kakšno vlogo imajo spremljajoči dejavniki v posameznih državah.

Poleg glavnih mednarodnih poročil OECD pripravlja še vrsto drugih. Ob zaključku prvega cikla raziskave so bila objavljena izčrpna poročila o dosežkih bralne pismenosti (OECD, 2002c, 2003b). OECD je izdala tudi vrsto drugih poročil o nekaterih vidikih raziskave PISA 2000, na primer o odzivu učencev, o njihovih učnih pristopih, o šolskih okoliščinah in kakovosti pouka (OECD, 2003b, 2004a, 2005b) idr. Poleg tega OECD izdaja tudi vrsto tehničnih dokumentov, kot so tehnično poročilo, priročnik za uporabo baze podatkov, vzorce nalog in dokumente o izhodiščih projekta (OECD 2002a, 2002b). Podobna poročila so bila ali še bodo pripravljena tudi o rezultatih cikla 2003. Publikacije raziskave PISA so dostopne na spletni strani OECD <http://www.pisa.oecd.org/>.

Mnoge sodelujoče države pripravljajo tudi lastna nacionalna poročila, v katerih so dosežki države podrobneje prikazani in ovrednoteni, pri tem pa upoštevajo tudi številne dejavnike, značilne za njihovo državo.

Viri

Adams, R. J. in Wu, M. L. (2002). *PISA 2000 technical report*. Pariz: OECD Publications.

Adams, R. J. in Wu, M. L. (2006). The mixed-coefficient multinomial logit model: A generalized form of the Rasch model. V: M. v. Davier and C. H. Carstensen (ur.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 57-76). New York: Springer Verlag.

Adams, R. J., Wu, M. L. in Carstensen, C. H. (2006). Application of multivariate Rasch models in international large scale educational assessment. V: M. v. Davier and C. H. Carstensen (ur.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 271-280). New York: Springer Verlag.

OECD (2001). *Knowledge and skills for life: First results from PISA 2000*. Pariz: OECD Publications.

OECD (2002a). *Manual for the PISA 2000 database*. Pariz: OECD Publications.

OECD (2002b). *Sample tasks from the PISA 2000 assessment reading mathematical and scientific literacy*. Pariz: OECD Publications.

OECD (2002c). *Reading for change: Performance and engagement across countries*. Pariz: OECD Publications.

OECD (2003a). *Literacy skills for the world of tomorrow: Further results from PISA 2000*. Pariz: OECD Publications.

OECD (2003b). *Student engagement at school, a sense of belonging and participation: Results from PISA 2000*. Pariz: OECD Publications.

- OECD (2004a). *What makes school systems perform: Seeing school systems through the prism of PISA*. Pariz: OECD Publications.
- OECD (2004b). *Learning for tomorrow's world: First results from PISA 2003*. Pariz: OECD Publications.
- OECD (2005a). *PISA 2003 technical report*. Pariz: OECD Publications.
- OECD (2005b). *School factors related to quality and equity: Results from PISA 2000*. Pariz: OECD Publications.
- OECD (2006). *Assessing scientific, reading and mathematical literacy: A framework for PISA 2006*. Pariz: OECD Publications.
- Routitsky, A. in Turner, R. (April, 2003) *Item format types and their influence on cross-national comparisons of student performance*. Presentation given to the annual meeting of the American Educational Research Association. Chicago.
- Wu, M. L., Adams, R. J. in Wilson, M. R. (1998). *ACER ConQuest: Generalized item response modeling software* [Computer program]. Melbourne: ACER Press.

ENAKOVREDNOST PREVODOV ZA RAZISKAVO PISA V RAZLIČNIH DRŽAVAH

Aletta Grisay
Univerza v Liègeu, Belgija

John H.A.L. de Jong
Služba za jezikovno testiranje (*Language Testing Services*), Nizozemska

Eveline Gebhardt
Avstralski inštitut za raziskovanje v izobraževanju, Melbourne

Alla Berezner
Avstralski inštitut za raziskovanje v izobraževanju, Melbourne

Beatrice Halleux-Monseur
HallStat SPRL, Belgija

Število držav, ki sodelujejo v mednarodnih primerjalnih raziskavah, kot sta TIMSS in PISA, vse bolj narašča, zato je zagotavljanje jezikovne in kulturne enakovrednosti vrste nacionalnih različ merskih instrumentov vse večji izziv. V glavni raziskavi PISA 2006 je na primer sodelovalo 58 držav. Za vsako državo je bilo gradivo za preverjanje znanja treba prevesti v jezik, v katerem na vzorčnih šolah poteka pouk. Vse nacionalne različice, namenjene petim ali več odstotkom ciljne populacije (kar je skupaj 77 različic v 42 jezikih), so mednarodno verificirane. V postopku verifikacije se preveri, ali prevodi ustrezajo izvirnim različicam gradiva v angleškem in francoskem jeziku, ki ju pripravi mednarodno združenje organizacij za izvedbo raziskave PISA. Na podlagi verifikacije prevodov in empiričnih analiz se presodi, ali so gradiva vseh sodelujočih držav dovolj kakovostna in jezikovno enakovredna.

V tem prispevku so na kratko opisani postopki raziskave PISA za zagotavljanje kakovosti prevajanja in prirejanja gradiva, osredotoča pa se tudi na oblikovanje kazalnikov, na podlagi katerih lahko presojamo o enakovrednosti prevodov. Za prikaz analiz in glavnih ugotovitev smo uporabili podatke iz predraziskave PISA 2005.

Uvod

V prvem ciklu raziskave PISA leta 2000 so sodelovale vse države članice OECD razen Turčije in Slovaške, vseh 30 pa je bilo vključenih tudi v raziskavi PISA 2003 in PISA 2006 (to so Avstralija, Avstrija, Belgija, Češka, Danska, Finska, Francija, Grčija, Irska, Islandija, Italija, Japonska, Južna Koreja, Kanada, Luksemburg, Madžarska, Mehika, Nemčija, Nizozemska, Norveška, Nova Zelandija, Poljska, Portugalska, Slovaška, Španija, Švedska, Švica, Turčija, Velika Britanija in ZDA). Naštete države se geografsko, jezikovno in kulturno sicer precej razlikujejo, a so zlasti ekonomsko in socialno najbrž precej bolj homogene kakor države, ki sodelujejo v drugih podobnih mednarodnih raziskavah, kot je na primer TIMSS.

Od prvega cikla raziskave PISA do danes pa se je projektu pridružilo veliko držav partnerk, ki niso članice OECD (5 držav v raziskavi PISA 2000, 8 v ponovitvi raziskave PISA 2000 v letu 2001, 11 v raziskavi PISA 2003 in 27 v raziskavi PISA 2006).¹ Posledično se je tudi število jezikov, v katere je bilo treba prevesti merske instrumente, povečalo s 25 na 42, število nacionalnih različic gradiva, ki so terjale neodvisni pregled za potrditev njihove skladnosti z izvirnima različicama, pa z 41 na 77. Raznolikost gradiva je tako postala precej večja, s tem pa se je povečal tudi izziv, kako zagotoviti pravično preverjanje in primerljivost merskih instrumentov v različnih državah.

PISA s strogimi mehanizmi zagotavlja pripravo kakovostnih prevodov merskih instrumentov v vse jezike, v katerih poteka pouk v sodelujočih državah, z neodvisno verifikacijo pa preverja skladnost prevodov z izvirnimi različicami gradiva, ki jih pripravlja mednarodni center raziskave PISA.

Postopek priprave gradiva za preverjanje poteka tako, da se najprej sestavi dve enaki izvirni različici (v angleščini in francoščini), nato pa je priporočljivo, da vsaka država neodvisno pripravi po dve nacionalni različici (eno na podlagi angleškega in eno na podlagi francoskega izvirnika), ki ju nato uskladi in združi v končno nacionalno različico gradiva (Grisay, 2003a).

Tako v angleški kot v francoski različici, ki ju pripravi mednarodni center raziskave PISA, je poleg besedila nalog tudi vrsta opomb za prevajalce, ki pomagajo pri razreševanju morebitnih težav pri prevajanju in prirejanju gradiva. Pri vsaki nalogi, ki preverja bralno razumevanje, je na primer zapisan tudi cilj naloge. S tem se poveča verjetnost, da bo tudi prevedena naloga od učencev

¹ Države partnerke v raziskavi PISA 2000 so bile Brazilija, Hong Kong, Indonezija, Latvija, Rusija in Tajsko, v ponovitvi te raziskave v letu 2001 pa še Albanija, Argentina, Bolgarija, Čile, Izrael, Makedonija, Peru in Romunija. V raziskavi PISA 2003 je sodelovalo 11 držav partnerk: Brazilija, Hong Kong, Indonezija, Latvija, Liechtenstein, Macao, Rusija, Srbija, Tajsko, Tunizija in Urugvaj. Države partnerke v raziskavi PISA 2006 pa so bile: Argentina, Azerbajdžan, Brazilija, Bolgarija, Čile, Estonija, Hong Kong, Hrvaška, Indonezija, Izrael, Jordanija, Katar, Kirgizistan, Kolumbija, Latvija, Liechtenstein, Litva, Macao, Romunija, Rusija, Slovenija, Srbija in Črna Gora, Tajsko, Tajvan, Tunizija in Urugvaj.

zahtevala enaka znanja in spretnosti (McQueen in Mendelovits, 2003). Poleg tega vsaka država dobi tudi obsežna navodila s priporočili za izvedbo postopkov prevajanja in natančnimi navodili za prevajanje in prirejanje gradiva, s katerimi se skupina glavnih prevajalcev iz vsake države seznani na posebnem usposabljanju.

Mednarodna skupina strokovnjakov, ki jih imenuje in usposobi mednarodni center raziskave PISA, nato v postopku verifikacije pregleda vse nacionalne različice gradiva in ugotovi, ali so skladne z izvirniki. Skupino za verifikacijo prevodov sestavljajo profesionalni prevajalci za angleški in francoski jezik, ki so poleg tega tudi naravni govorci ciljnega jezika. Sodelujoče države nato pregledajo popravke, ki jih je predlagala skupina za verifikacijo prevoda, po skupnem posvetu včasih kakšnega tudi zavrnejo ali najdejo alternativno rešitev, po vnosu popravkov pa osnutke bodočih delovnih zvezkov za učence ponovno pošljejo v končni pregled skupini za verifikacijo. Ta ponovno preveri ustreznost popravkov, oblikovanja in slikovnega gradiva.

Vsaka država mora imenovati tudi nacionalno ekspertno skupino, ki preverja, ali so izvirne različice gradiva primerne za njihove 15-letne učence (še posebej je pozorna na morebitna neskladja s kurikuli). Skupina pomaga prevajalcem pri terminoloških in vsebinskih vprašanjih ter pregleda in potrdi končno različico gradiva.

V fazi predraziskave in glavne raziskave mednarodni center vsaki državi pošlje izčrpno poročilo o analizi podatkov te države po nalogah, zlasti »seznam vprašljivih nalog«, ki opozarja na izstopajoče naloge v tej državi ali druga odstopanja v njihovih podatkih. Nacionalni centri pregledajo takšne naloge in poiščejo verjetne razlage za odstopanja, nato pa naloge popravijo (v fazi predraziskave), včasih pa tudi izločijo iz analiz (v fazi glavne raziskave).

Ko je analiza podatkov v glavni raziskavi vsakega cikla zaključena, ekspertna skupina za pregled podatkov (*Data Adjudication expert panel*), sestavljena iz analitikov raziskave PISA, svetovalcev iz skupine za tehnično izvedbo raziskave in strokovnjakov s posameznih področij oceni, ali je vsaka sodelujoča država dosegla tehnične standarde in ali se njene rezultate lahko vključi v poročila o raziskavi PISA. Publikacijo, kjer so ti standardi natančno opredeljeni, potrdi mednarodni svet raziskave PISA in jo posreduje vsem nacionalnim centrom (OECD, 2007).

V raziskavah PISA 2000 in PISA 2003 je ekspertna skupina za pregled podatkov informacije o prevajanju oziroma prirejanju gradiva dobila zlasti iz: (i) poročil nacionalnih centrov o poteku raziskave (na primer, ali so najeli profesionalne prevajalce; ali so prevajalce usposobili skladno s smernicami raziskave PISA; ali so bila gradiva prevedena dvakrat in nato usklajena pri pripravi končne različice; ali so bili prevodi verificirani; ali je gradiva pregledala nacionalna ekspertna skupina), in iz (ii) vsebinskih poročil strokovnjakov, ki so izvedli postopek verifikacije prevodov (o pogostosti in vrsti ugotovljenih napak).

V raziskavi PISA 2000 smo preučili tudi vse nacionalne različice gradiva v enem od treh najpogostejših jezikov sodelujočih držav, tj. v angleščini, francoščini ali nemščini (Grisay, 2003b). S pomočjo indeksov težavnosti smo poiskali tiste razlike v težavnosti nalog med državami, ki so bile značilne za vse ali večino držav znotraj neke jezikovne skupine (za razliko od ostalih dveh skupin) in so nakazovale, da so razlog za to morda napake pri prevajanju gradiv.

Kljub temu pa na podlagi podatkov, pridobljenih v raziskavah PISA 2000 in PISA 2003, ni bilo mogoče pripraviti uradnih poročil o enakovrednosti prevodov, ki bi upoštevala tudi spodaj navedeni smernici za prirejanje gradiva, ki ju je poleg ostalih sprejela Mednarodna komisija za teste (*International Test Commission*) (Hambleton, 1994; Hambleton in Merenda, 2005):

1. Smernica D5: »Sestavljavci preizkusov bi morali na podlagi sistematičnih dokazov – tako jezikovnih kot psiholoških – izboljševati zanesljivost postopkov prirejanja preizkusov in zbirati dokaze o enakovrednosti prevodov v vse jezike.«
2. Smernica D9: »Sestavljavci preizkusov bi morali zbirati statistične dokaze o enakovrednosti nalog za vse ciljne populacije.«

V predraziskavi PISA 2006 je bil pregled vprašanj o enakovrednosti prevodov veliko bolj sistematičen. Skladno z Zumbovim predlogom (2003) je bilo opravljenih dvoje analiz: analiza na ravni nalog in analiza na ravni lestvic dosežkov.

Na ravni nalog je Le (2006) natančno raziskal razlike v delovanju nalog (*Differential Item Functioning – DIF*) po državah, jezikih in spolu glede na tiste značilnosti nalog iz naravoslovja, ki so opredeljene v izhodiščih preverjanja – tip in kontekst naloge ter naravoslovne kompetence, ki jih preverja.

Na ravni lestvic dosežkov pa bomo v nadaljevanju predstavili nekaj možnih kazalnikov, s pomočjo katerih bi lažje izvedli interne aktivnosti prevajanja, verifikacije prevodov ter priprave uradnih poročil o enakovrednosti prevodov.

Pri organizaciji internih postopkov bi kazalnike lahko na primer uporabili za preverjanje: (i) ali nespremenljivost konstrukta, ki ga merimo (*construct invariance*) (Van de Vijver, 1998), kaže, da sta izvirni različici gradiva karseda enakovredni in brez vprašljivih ali pristranskih nalog; (ii) ali so nacionalne različice posameznih gradiv nekvalitetne in je zato potrebno sprejeti določene ukrepe (na primer ugotoviti, ali je treba zamenjati kakšnega prevajalca ali strokovnjaka, ki je opravil mednarodno verifikacijo prevoda); in (iii) ali se kakovost prevodov v neki državi iz cikla v cikel izboljšuje (ali slabša).

Kazalniki bi morali nuditi informacije, ki bi jih v uradnih poročilih lahko uporabili pri presojanju o kakovosti podatkov za neko državo ter pri poročanju o morebitnih državah, kjer so bili prevodi gradiva tako nekvalitetni, da niso bili več primerljivi z izvirnimi različicami ali z večino prevodov drugih držav.

V tem prispevku bomo predstavili in razpravljali o analizah, ki so bile prvi korak k oblikovanju primernih kazalnikov.

Ugotavljanje geografskih in jezikovnih vzorcev razlik s pomočjo klastrske analize (*cluster analysis*)

Učinkovita metoda za ugotavljanje sistematičnih vzorcev v delovanju več različnih merskih instrumentov, uporabljenih v mednarodni raziskavi, je clusterska analiza razlik med indeksi težavnosti nalog v posameznih nacionalnih različicah preizkusov.

Takšno metodo so pri ponovni analizi podatkov mednarodne študije o pismenosti odraslih IALS (*International Adult Literacy Survey*) uporabili Blum, Goldstein in Guérin-Pace (2001), da bi dokazali, da se hierarhija nalog, določena na podlagi deleža pravičnih odgovorov v vsaki sodelujoči državi, od države do države precej razlikuje. Eno značilno skupino so tvorile angleško govoreče države, ki so sodelovale v IALS, v dve posebni skupini pa so se uvrstile tudi francosko in nemško govoreče države. Avtorji analize so zaključili, da *»je bila uspešnost reševanja nalog odvisna od geografskih in jezikovnih dejavnikov, kar je v nasprotju s predpostavko te analize, da so podatki med seboj primerljivi in da so dosežki neodvisni od jezika, v katerem poteka preverjanje.«*

V podobnih analizah raziskav PISA 2000, PISA 2003 in PISA 2006 so prišli do podobnih spoznanj, le da so avtorji analiz tokrat namesto surovih deležev pravičnih odgovorov uporabili parametre težavnosti nalog, ki so jih dobili z neodvisno Raschevo analizo podatkov za posamezno državo. Shema 1 prikazuje dendrogram, dobljen s clustersko analizo težavnosti nalog (vrednosti *delta*) iz naravoslovja v predraziskavi PISA 2006.²

Vzorci podobnosti in razlik, ki so razvidni iz sheme 1, podobno kot tisti iz študije IALS razkrivajo vpliv jezikovnih, geografskih in kulturnih dejavnikov.

Nacionalne različice merskih instrumentov, pripravljene v istem jeziku, se skoraj po pravilu znajdejo v isti skupini. To na primer velja za večino preizkusov v nemščini, angleščini, francoščini, nizozemščini, španščini, ruščini, kitajščini in arabščini. Groba struktura dendrograma se zdi odvisna tudi od tipologije jezikov, saj tri velike skupine tvorijo vsi germanski jeziki (tj. nemščina, angleščina, nizozemščina in skandinavski jeziki), romanski jeziki (tj. francoščina, italijanščina, portugalščina, španščina in različna španska narečja) in slovanski ter baltski jeziki (ruščina, poljščina, srbsščina, hrvaščina, črnogorščina, slovenščina, slovaščina, litvanščina in latvijščina).

² Analiza je bila izvedena na 70 podatkovnih nizih predraziskave (en niz podatkov predstavlja en jezik v neki državi) in na 201 nalogi iz naravoslovja. Na teh nalogah so temeljile analize predraziskave, že pred tem pa je bilo iz analiz izločenih 46 nalog, ki so imele hude pomanjkljivosti tako v izvirnih kakor v številnih nacionalnih različicah, kar je kazalo na to, da so bile bolj kakor prevodi slabe same naloge.

Zelo očiten je tudi vpliv geografske bližine. Tako so se tri različice prevodov v ugrofinske jezike razvrstile v tri zelo različne »geografske« skupine: finska različica med nemške in skandinavske, estonska v majhno baltsko skupino in madžarska v ohlapno vzhodnoevropsko skupino, sestavljeno večinoma iz slovanskih, pa tudi iz enega romanskega jezika, romunščine. Podobno sta se različici v latinskoameriški španščini in portugalščini znašli v isti podskupini, neodvisno od evropskih različic v portugalščini, španščini, katalonščini in galicijsščini. Tudi hebrejska in arabska različica, ki sta bili uporabljeni v Izraelu, sta v isti skupini ne glede na jezikovno razliknost.

Nasprotno je bila angleška različica za Katar precej oddaljena od arabske različice, pripravljene za isto državo, pa tudi od ostalih dveh različic v arabščini, uporabljenih v Jordaniji in Tuniziji. Tudi ruska različica, uporabljena v Kirgizistanu, se je uvrstila v rusko podskupino (skupaj s preostalimi ruskimi različicami za Rusijo, Latvijo in Estonijo), precej stran od različice v kirgiškem jeziku. Očitno lahko v nekaterih primerih uporaba jezika manjšine pomeni tudi posebni kurikulum in/ali precej drugačne populacijske podskupine.

Opazimo lahko tudi vpliv kulturnih in socialno-ekonomskih značilnosti nekaterih držav. Najbolj razvite azijske države (Japonska, Južna Koreja, Hong Kong, Kitajski Tajpej in Macao) tako tvorijo eno skupino, ki je v mnogočem bolj podobna veliki skupini zahodno- in indoevropskih držav kakor pa skupini najbolj »daljnih« sodelujočih držav, kamor se uvrščata dve manj industrializirani tihomorski državi (Tajska in Indonezija) in vrsta arabsko govorečih držav (arabski Katar, Jordanija, Tunizija in vse države turških/altajskih jezikov (Turčija, Kirgizistan – uzbeški in kirgiški jezik ter Azerbajdžan – azeri). V tej skupini je BDP na prebivalca precej nižji od povprečja držav članic OECD, za države te skupine pa je značilen tudi večji delež učencev, ki se uvrščajo na dno lestvic dosežkov raziskave PISA.

Struktura razvrstitve držav v skupine se je le neznatno spremenila, ko smo analizo ponovili samo na podatkih za tiste naloge, ki so bile dovolj kvalitetne, da jih sestavljavci preizkusov in člani ekspertne skupine za naravoslovje niso izločili iz glavne raziskave PISA 2006. Vzorci, ki jih je razkrila ta analiza, so v grobem pokazali, da naloge iz naravoslovja v predraziskavi PISA 2006 niso delovale povsem enako v vseh sodelujočih državah in da so razlike med državami najverjetneje odvisne od jezika, v katerem je potekalo preverjanje.

Kako velik je morebitni vpliv jezikovne pristranosti?

Iz dendrograma na shemi 1 pa ni razvidno, v *kolikšni meri* je njegova struktura odvisna od jezikovnih razlik. Mnogi vzorci, ki jih lahko razberemo, bi na primer lahko nastali zaradi drobnih razlik pri zgolj dveh ali treh nalogah v neki skupini držav, njihov vpliv na splošno mednarodno primerljivost lestvice dosežkov pri naravoslovju pa bi bil zanemarljiv.

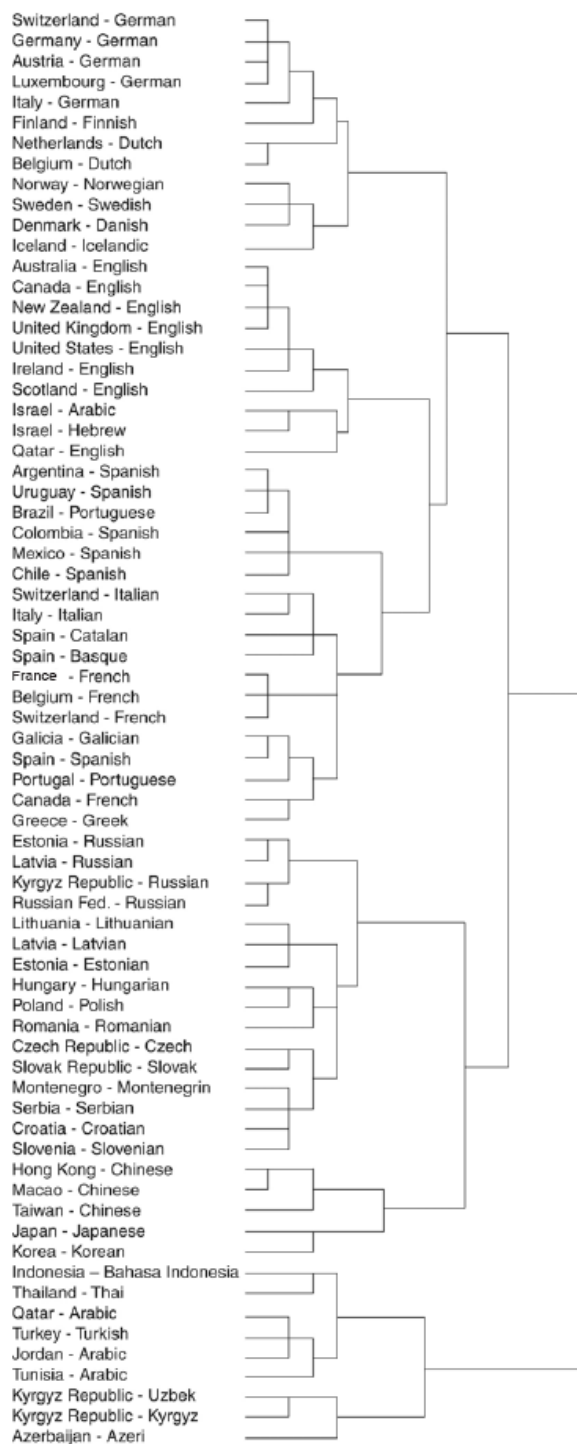
Za ugotavljanje deleža variance težavnosti nalog, ki je bila skupna vsem državam, v primerjavi z deležem variance, ki jo pojasnijo sekundarni faktorji oziroma faktorji, specifični za posamezne preizkuse ali skupine preizkusov, smo uporabili faktorsko analizo. Kovariančna matrika, ki smo jo uporabili pri tej analizi, je vsebovala 70 spremenljivk (število nacionalnih različic nalog) in 201 merjenje (končno število nalog v predraziskavi brez slabih nalog, ki so bile izločene).

Prvi faktor, ki ga je pokazala analiza, je pojasnil 79,2 odstotkov skupne variance; pri dveh tretjinah nacionalnih različic nalog je bilo moč z njim pojasniti 90 odstotkov variance, pri nobeni nalogi pa ta delež ni bil manjši od 74 odstotkov, kar je kazalo na visoko mednarodno primerljivost konstrukta.

Še dva faktorja sta imela lastno vrednost večjo od 1. Z drugim (nerotiranim) faktorjem je bilo mogoče pojasniti dodatnih 2,7 odstotkov variance in je zavzemal nizke pozitivne vrednosti za vrsto različic nalog, uporabljenih v islamskih državah (Azerbajdžan, Kirgizistan, Turčija, Indonezija, Jordanija, Tunizija, Katar), a je znašal 0 ali manj za vse ostale različice nalog. Tretjega faktorja (1,6 odstotka variance) ni bilo mogoče pojasniti.

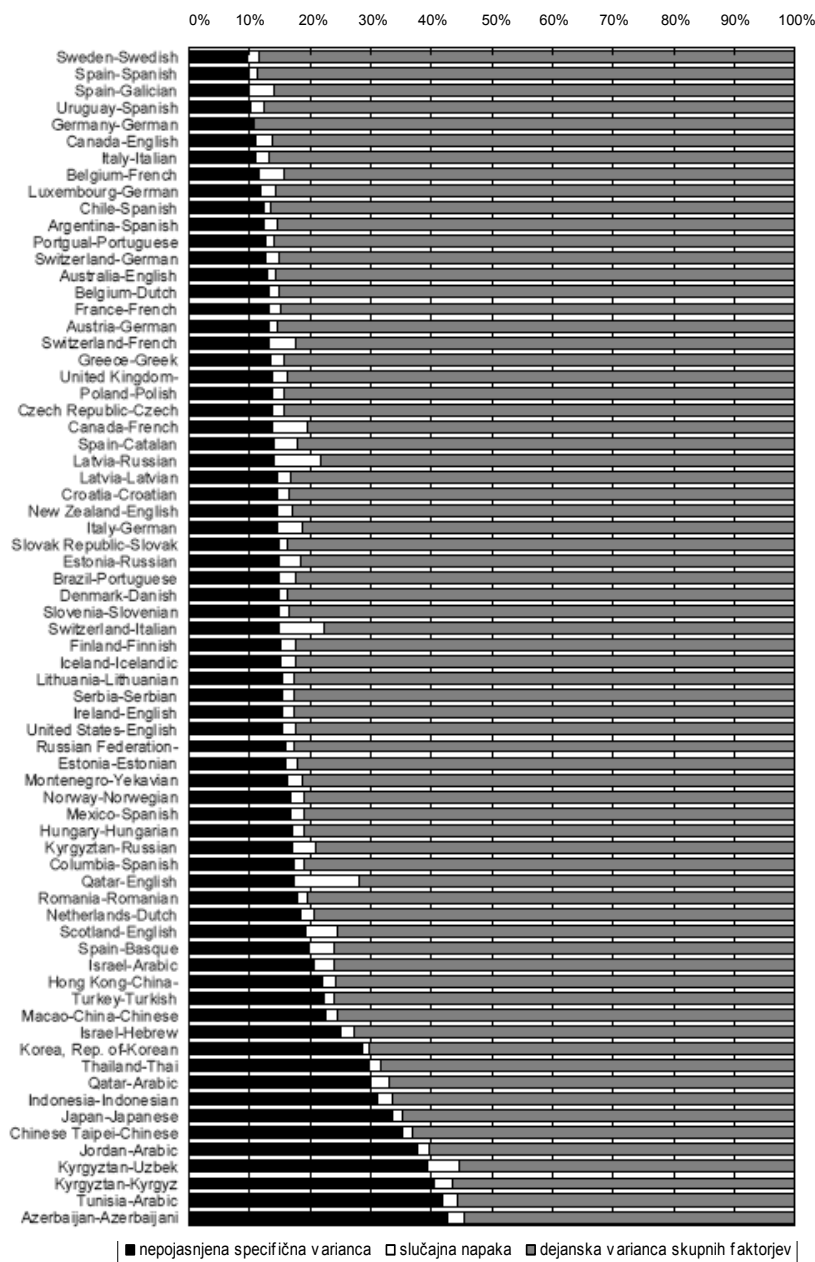
Treba je poudariti, da je podobna analiza, ki jo je izvedel Baye (2004) na podatkih glavne raziskave PISA 2000 s poudarkom na bralni pismenosti, prav tako izločila zelo močan prvi faktor (ki je pojasnjeval 82 odstotkov skupne variance) in dva manjša faktorja, ki ju ni bilo mogoče pojasniti. Tudi v ponovitvi analize na podatkih glavne raziskave PISA 2003 s poudarkom na matematični pismenosti je prvi faktor pojasnil 91 odstotkov skupne variance. V obeh primerih je dendrogram, ki smo ga dobili s clustersko analizo težavnosti nalog, pokazal podobno odvisnost od jezikovnih in kulturnih dejavnikov kakor tisti na shemi 1. Iz tega lahko sklepamo, da je specifična varianca, ki jo pojasnijo jezikovni in kulturni dejavniki, konsistentna in igra enako vlogo na vseh treh področjih pismenosti raziskave PISA, četudi je njen delež relativno majhen v primerjavi z varianco, ki jo lahko pripišemo skupni latentni dimenziji, ki jo s preizkusom merimo v vseh sodelujočih državah.

Shema 1: Clusterska analiza težavnosti nalog glede na nacionalne razlike preizkusa



Prvi kazalnik enakovrednosti nacionalnih različic preizkusa iz naravoslovja, ki je bil uporabljen v predraziskavi PISA 2006, smo določili na podlagi variance skupnih faktorjev s pomočjo enofaktorske analize glavnih komponent (*one-factor principal components analysis – PCA*). Prikazan je v shemi 2.

Shema 2: Varianca skupnih faktorjev ter specifična varianca in varianca napake parametrov za naloge iz naravoslovja po državah



V shemi 2 sivi del vsakega stolpca prikazuje delež variance skupnih faktorjev za dano nacionalno različico nalog, dobljene s pomočjo faktorske analize – *tj. varianco težavnosti nalog, ki je skupna vsem nacionalnim različicam*. Preostanek variance je razdeljen na dva dela: na *slučajno napako* (beli del vsakega stolpca) in na specifično komponento variance, ki jo morda lahko pojasnimo z *različnimi vrstami pristranosti* (črni del vsakega stolpca).

Beli del stolpcev ustreza *deležu slučajne napake*, ki je odvisna zlasti od velikosti vzorca učencev v predraziskavi in od drugih *modeliranih (modelled)* vzrokov za slučajno napako. Zlasti različice preizkusov, namenjene pripadnikom manjšin v številnih večjezikovnih državah, ponavadi rešuje manj učencev kakor preizkuse v prevladujočem jeziku, zato je delež slučajne napake tu večji. Ta delež smo določili s pomočjo simulacije, ki je vključevala sledeče korake:

1. V raziskavi PISA 2006 je vsak učenec reševal delovni zvezek s 50 nalogami oz. vprašanji, ki sledijo nalogi z uvodnim besedilom. Ker so posamezne naloge po delovnih zvezkih ciklično razporejene (*booklet rotation design*), je v normalni porazdelitvi 50 vrednosti za *dejansko* težavnost nalog.
2. Na podlagi *dejanskih* parametrov nalog, povprečnih dosežkov posameznih držav (iz predraziskave), variance dosežkov posameznih držav in velikosti vzorca (povprečnega števila učencev, ki so reševali nalogo) smo izločili odgovore učencev za vsako kombinacijo država-jezik.
3. Te odgovore smo analizirali in ocenili vrednost parametrov nalog. Ocenjene vrednosti parametrov za posamezne nacionalne različice nalog so nekoliko odstopale od dejanskih vrednosti zaradi (a) slučajne napake, (b) ciljanja določenih skupin učencev s posamezno nalogo, (c) velikosti vzorca in (d) variance dosežkov znotraj države.
4. Parametre za posamezne nacionalne različice nalog, ki smo jih dobili s simulacijo, smo uporabili v enofaktorski analizi glavnih komponent. V simulaciji so delež variance, ki je ne moremo pojasniti s skupnimi faktorji, skoraj v celoti *povzročili* slučajna napaka, ciljanje neke skupine učencev s posamezno nalogo in velikost vzorca, v manjši meri pa varianca dosežkov učencev znotraj države.
5. Korake od 2 do 5 smo 50-krat ponovili in napako izračunali kot povprečje 50 simulacij za vsako državo, da bi tako dobili čim natančnejše ocene. Komponento napake smo nato odšteli od *variance*, ki smo jo dobili iz realnih podatkov, ko smo izločili skupni faktor. S tem smo izničili vpliv slučajne napake, vpliv razlik med državami in znotraj njih ter vpliv velikosti vzorca. Povprečni delež variance napake je bil 2,4 odstotka, ocenjeni delež pa je nihaj od skoraj 0 odstotkov v Nemčiji (kjer je vsak delovni zvezek reševalo okrog 1400 učencev), pa do 11 odstotkov za angleško različico preizkusa v Katarju (kjer je en delovni zvezek reševalo samo 60–70 učencev).

Črni del stolpcev na shemi 2 je nepojasnjena komponenta, ki ostane po tem, ko od dejanske variance težavnosti nalog, ki je ne moremo pojasniti s skupnimi faktorji, v neki nacionalni različici preizkusa odštejemo komponento napake (kakor je opisano v koraku 5). Ta del stolpcev lahko razumemo kot možni pokazatelj tistega specifičnega dela variance podatkov neke države, ki ni niti skupen z mednarodno lestvico, niti ga ni mogoče pripisati zgolj naključnim dejavnikom. Črni del stolpcev je zato najverjetneje odvisen od *pristranosti, ki vpliva na primerljivost tega preizkusa z ostalimi različicami*.

Iz sheme 2 lahko sklepamo, da je bil delež specifične variance, ki jo lahko pojasnimo s pristranostjo, majhen in dokaj nespremenljiv v več kakor dveh tretjinah nacionalnih različic preizkusov. Še posebej ni bilo nobenega dokaza, da bi bilo v angleškem in francoskem preizkusu, ki sta nastala neposredno iz izvirnih različic, pristranosti bistveno manj kakor v ostalih nacionalnih različicah, ki jih je bilo treba iz angleščine in francoščine prevesti v druge zahodne ali evropske jezike in po potrebi prilagoditi.

Nasprotno pa so v skupini nacionalnih različic preizkusov, ki so jih reševali zlasti učenci v državah Bližnjega vzhoda in Azije in se uvrščajo na skrajno desno stran sheme 2, vrednosti kazalnikov »specifične variance« dokaj visoke (od 20 odstotkov variance za arabsko različico, uporabljeno v Izraelu, do 43 odstotkov za (neodvisno) arabsko različico, uporabljeno v Tuniziji, in za različico v jeziku azeri v Azerbajdžanu). Trije preizkusi iz te skupine (turški, japonski in korejski) so bili uporabljeni tudi v državah članicah OECD, ki vključujejo tako nekaj držav z najvišjimi rezultati s področja naravoslovja v predraziskavi PISA 2006 (Kitajski Tajpej, Japonska, Južna Koreja, Hong Kong), kot nekaj takih, kjer so bili naravoslovni dosežki učencev med najnižjimi (Azerbajdžan, Kirgizistan, Katar, Indonezija, Tajska, Tunizija, Turčija, Jordanija).

Razumeti nekatere možne izvore pristranosti

Da bi bolje razumeli, zakaj se v raziskavi PISA merski instrument za naravoslovje zdi v neki skupini sodelujočih držav manj »enakovreden« kakor v neki drugi, smo pripravili vrsto kazalnikov o morebiti zanimivih značilnostih teh držav in jih uporabili v multipli regresijski analizi, da bi tako lahko »napovedali« delež variance, ki je odvisna od teh značilnosti. Ti kazalniki so:

- **KAZALNIK JEZIKOVNIH IN KULTURNIH RAZLIK GLEDE NA JEZIK, V KATEREM POTEKA POUK.** Kljub nekaj poskusom, da bi našli natančnejšo klasifikacijo, smo ohranili dihotomno spremenljivko, ki loči indoevropske jezike na eni in neindoevropske jezike na drugi strani. Korelacija med kazalnikom specifičnih značilnosti in jezikom pouka v tej državi se ni bistveno izboljšala, ko smo uporabili natančnejše kategorije – na primer skupine germanskih, romanskih, slovanskih, altajskih ali ugrofinskih jezikov namesto bolj grobe delitve na indoevropske in neindoevropske jezike.

- **KAZALNIK MOŽNIH EKONOMSKIH RAZLIK.** Za predstavitev morebitnih razlik med sodelujočimi državami v relativni stopnji ekonomskega razvoja smo uporabili BDP na prebivalca, izražen v ameriških dolarjih glede na standard kupne moči. Če je bilo v neki državi pripravljenih več nacionalnih različic preizkusa, smo za vse upoštevali enak BDP.
- **KAZALNIK MOŽNIH RAZLIK V KAKOVOSTI PREVODOV,** dobljen iz poročil o verifikaciji prevodov. V raziskavi PISA 2006 so morali strokovnjaki, ki jih je za neodvisni pregled prevodov imenoval mednarodni center, na sistematičen način poročati o napakah, ki so jih odkrili in za katere so menili, da spodbujajo primerljivost z izvirnimi različicami merskih instrumentov. Napake so opisali v obliki Excelovih tabel (v angleškem jeziku) in jih nato poslali v pregled mednarodnemu centru raziskave PISA. Mednarodni center je v poročilih označil »ključne napake«, ki jih je bilo treba odpraviti pred avtorizacijo nacionalne različice gradiva za predraziskavo. Kot kazalnik smo uporabili število potrebnih »ključnih popravkov« na 100.000 znakov v besedilu vsake nacionalne različice preizkusa. Ta kazalnik je bil vse prej kot popoln, saj je bila večina odkritih napak praviloma odpravljena še pred predraziskavo, pa tudi zato, ker je bilo skupno število vseh »ključnih napak«, o katerih je bil obveščen mednarodni center, odvisno od zanesljivosti in osebne presoje več kakor štiridesetih strokovnjakov, ki so opravili neodvisni pregled prevodov. Kazalnik možnih razlik v kakovosti prevodov zato temelji zlasti na predpostavki, da je verjetnost ostalih napak, ki so bile v postopku verifikacije spregledane, veliko večja v preizkusih, kjer je bilo odkritih in odpravljenih tudi več hudih napak pri prevajanju.
- **KAZALNIK MOŽNIH RAZLIK V SNOVNI POKRITOSTI KURIKULA IZ NARAVOSLOVJA,** dobljen iz nacionalnega poročila o nalogah predraziskave PISA, ki ga je pripravil vsak nacionalni center. Nacionalni koordinatorji projekta so morali ustreznost vsake naravoslovne naloge glede na kurikulum svoje države oceniti na lestvici od 1 (ni vključeno v nacionalni kurikulum za 15-letne učence) do 5 (povsem ustrezna naloga glede na kurikulum za veliko večino 15-letnikov v državi). Kot kazalnik smo uporabili povprečno vrednost te spremenljivke za 247 nalog iz naravoslovja, ki so bile vključene v predraziskavo v vsaki državi. Če je bilo v neki državi pripravljenih več nacionalnih različic preizkusa, smo za vse upoštevali enako vrednost snovne pokritosti kurikula.

- **KAZALNIK MOŽNIH RAZLIK V KAKOVOSTI KODIRANJA ODPRTIH ODGOVOROV V NEKI DRŽAVI.** S pomočjo analize homogenosti smo ocenili konsistentnost koderjev odprtih odgovorov v neki državi. Vzorec odgovorov učencev za vsako nalogo odprtega tipa v predraziskavi PISA 2006 so ponovno kodirali štirje neodvisni ocenjevalci. Visoke vrednosti indeksa homogenosti pomenijo visoko stopnjo ujemanja različnih koderjev.³
- **KAZALNIK MOŽNIH UČINKOV CILJANJA NEKE SKUPINE UČENCEV S POSAMEZNO NALOGO (*targeting*),** ki se pojavijo poleg posledic slučajne napake, ki jih odpravimo v zgoraj opisani simulaciji. Učinek kopičenja rezultatov učencev na vrhu ali na dnu lestvice dosežkov (*ceiling or floor effects*) je lahko posledica slabše diskriminativnosti skupine nalog na skrajnem zgornjem in spodnjem delu distribucije nalog po težavnosti v državah z izrazito visokimi ali nizkimi povprečnimi dosežki. Kot kazalnik tega možnega izvora pristranosti smo uporabili povprečni indeks diskriminativnosti nalog iz naravoslovja, ki so jih učenci reševali v okviru predraziskave.

V tabeli 1 so prikazana razmerja med opisanimi spremenljivkami in glede na kazalnik specifične variance težavnosti nalog.

Tabela 1: Korelacije med kazalnikom specifične variance in kazalniki možnih izvorov pristranosti

spremenljivka	kazalnik specifične variance	povprečni dosežek pri naravoslovju	velikost vzorca v predraziskavi	neindoevropski jezik	BDP države	število ključnih popravkov prevoda na 100.000 znakov	snovna pokritost kurikula	indeks zanesljivosti kodiranja znotraj države	povprečni indeks diskriminativnosti nalog
kazalnik specifične variance	1,00								
povprečni dosežek pri naravoslovju	-0,23	1,00							
velikost vzorca v predraziskavi	-0,01	0,00	1,00						
neindoevropski jezik	0,75	-0,29	0,08	1,00					
BDP države	-0,36	0,61	-0,21	-0,32	1,00				
število ključnih popravkov prevoda na 100.000 znakov	0,20	-0,09	0,08	0,38	-0,27	1,00			
snovna pokritost kurikula	-0,16	0,09	0,04	-0,18	0,05	0,05	1,00		
indeks zanesljivosti kodiranja znotraj države	0,05	-0,15	0,11	0,06	-0,13	-0,15	-0,25	1,00	
povprečni indeks diskriminativnosti nalog	-0,51	0,40	0,02	-0,46	0,59	-0,44	0,03	-0,01	1,00

³ Opozoriti je treba, da ta indeks ne pokaže morebitne *pristranosti* kodiranja (npr. nekateri ocenjevalci so bolj blagi, nekateri pa strožji, in njihove ocene niso skladne z mednarodnimi priročniki za kodiranje odprtih odgovorov). V fazi glavne raziskave zato kakovost kodiranja preverijo še neodvisni ocenjevalci, ki jih imenuje mednarodni center raziskave PISA, na podlagi njihovih ocen pa se pripravijo dodatni kazalniki konsistentnosti kodiranja od države do države.

Pokazalo se je, da indeks zanesljivosti kodiranja odprtih odgovorov ni pomembno povezan s kazalnikom *specifične variance* ali s katerokoli drugo spremenljivko, zato tega indeksa nismo več uporabili v analizah.

Korelacija med kazalnikom *specifične variance* in kazalnikoma *snovne pokritosti kurikula* ter kakovosti prevoda (*ključni popravki*) je bila pričakovana: delež variance pristranosti je bil sorazmerno manjši v državah, kjer je kurikul za 15-letne učence vseboval vso snov, ki so jo v raziskavi PISA preverjale naloge iz naravoslovja, in kjer je bilo ob neodvisnem pregledu prevoda odkritih manj napak. Korelacije pa so bile zelo nizke, kar pomeni, da ti dve spremenljivki nista bili najboljši za ugotavljanje možnih virov pristranosti.

Postopna regresija (*stepwise regression*) spremenljivk *ključni popravki* in *snovna pokritost kurikula* s *specifično varianco* kot odvisno spremenljivko (glej tabelo 2) je pokazala, da se njun vpliv skoraj povsem prekriva z vplivom *BDP* in vključenosti v skupino *neindoevropskih jezikov*. Iz tega lahko sklepamo, da je stopnja neenakovrednosti prevoda odvisna od kombinacije dejavnikov: mnoge države z neindoevropskimi jeziki imajo nižji BDP od ostalih, prevodi v teh državah so bili morda nekoliko slabši in njihov kurikul ni pokrival vseh vsebin, ki so se preverjale v raziskavi PISA. Višja stopnja pristranosti je bila odvisna od kombinacije teh dejavnikov.

Tabela 2: Postopna regresija (stepwise regression) kazalnika specifične variance težavnosti nalog glede na različne druge kazalnike

upoštevani prediktorji	povečanje R^2	kumulativni R^2
Korak 1. Snovna pokritost kurikula	0,07	0,07
Korak 2. Kot zgoraj, plus ključni popravki	0,08	0,15
Korak 3. Kot zgoraj, plus BDP	0,12	0,27
Korak 4. Kot zgoraj, plus neindoevropski jezik	0,34	0,61
Korak 5. Kot zgoraj, plus povprečna diskriminativnost nalog	0,10	0,71

Regresijski koeficienti pri koraku 5

	koeficient	S.E.	t-vrednost	Pr> t
presečišče	0,670	0,076	8,86	<0,0001
snovna pokritost kurikula	-0,012	0,008	-1,40	0,166
ključni popravki	-0,000	0,000	-1,17	0,245
BDP	0,000	0,000	0,38	0,708
neindoevropski jezik	0,104	0,015	7,10	<0,0001
povprečna diskriminativnost nalog	-0,959	0,206	-4,65	<0,0001

Alternativni modeli

upoštevani prediktorji	pojasnjeni R^2
A. Samo neindoevropski jezik	0,56
B. Samo povprečna diskriminativnost nalog	0,43
C. Oboje zgoraj navedeno	0,69

Iz regresijskih koeficientov v tabeli 2 lahko razberemo, da v zadnjem koraku analize pomemben delež specifične variance pojasnujeta le spremenljivki *neindoevropski jezik* in *povprečna diskriminativnost nalog*. Če spremenimo vrstni red spremenljivk v posameznih korakih, lahko varianco kazalnika pristranosti razdelimo približno tako: 24 % specifične variance pojasni spremenljivka *neindoevropski jezik*; 10 % specifične variance pojasni *povprečna diskriminativnost nalog*; 2 % specifične variance lahko pripišemo *snovni pokritosti kurikula* in *ključnim popravkom*; 35 % specifične variance pojasni spremenljivki *neindoevropski jezik* in *povprečna diskriminativnost nalog* skupaj (delno se prekrivata tudi z *BDP*, *ključnimi popravki* in *snovno pokritostjo kurikula*); 29 % variance pa ostaja nepojasnjene.

Tak vzorec je nekoliko zaskrbljujoč, saj kaže, da lahko delež specifične variance v porazdelitvi težavnosti nalog v različnih nacionalnih različicah preizkusa v veliki meri (okrog 70 odstotkov variance tega kazalnika) pojasnimo z večjimi neskladji mednarodnih merskih instrumentov z nacionalnim kontekstom nekaterih držav – tj. z jezikovnimi in kulturnimi značilnostmi merskih instrumentov in ne povsem primernim ciljanjem nalog preizkusa glede na raven dosežkov učencev v teh državah.

Neskladnost s kulturnimi značilnostmi in nizka diskriminativnost nalog sta še posebej značilni za preizkuse bližnjevzhodnih in azijskih držav z nizkim BDP, ki niso članice OECD, kjer ima veliko učencev nizke dosežke. To pojasnjuje tudi visok delež variance skupnih komponent, ki jo je pokazala zgornja analiza.

Zanimivo je, da diskriminativnost nalog, kot kaže, ni imela posebne vloge v azijskih državah z visokimi dosežki, kot so Japonska, Južna Koreja, Kitajski Tajpej in Hong Kong. To pa ne velja za jezikovno neskladnost, kar posredno potrjuje veliko število popravkov v prevodih. Tudi nekatere nacionalne različice merskih instrumentov v neindoevropskih jezikih, ki so bile uporabljene v skupini zahodnih držav (finska, madžarska, estonska in baskovska), so imele ponavadi nekoliko višje vrednosti kazalnika specifične variance kot ostale evropske države, razloge za to pa bi tudi v tem primeru najbrž lahko našli v jezikovni neusklajenosti.

Ali bi se enakovrednost prevodov izboljšala, če bi za glavno raziskavo izbrali točno določene skupine nalog?

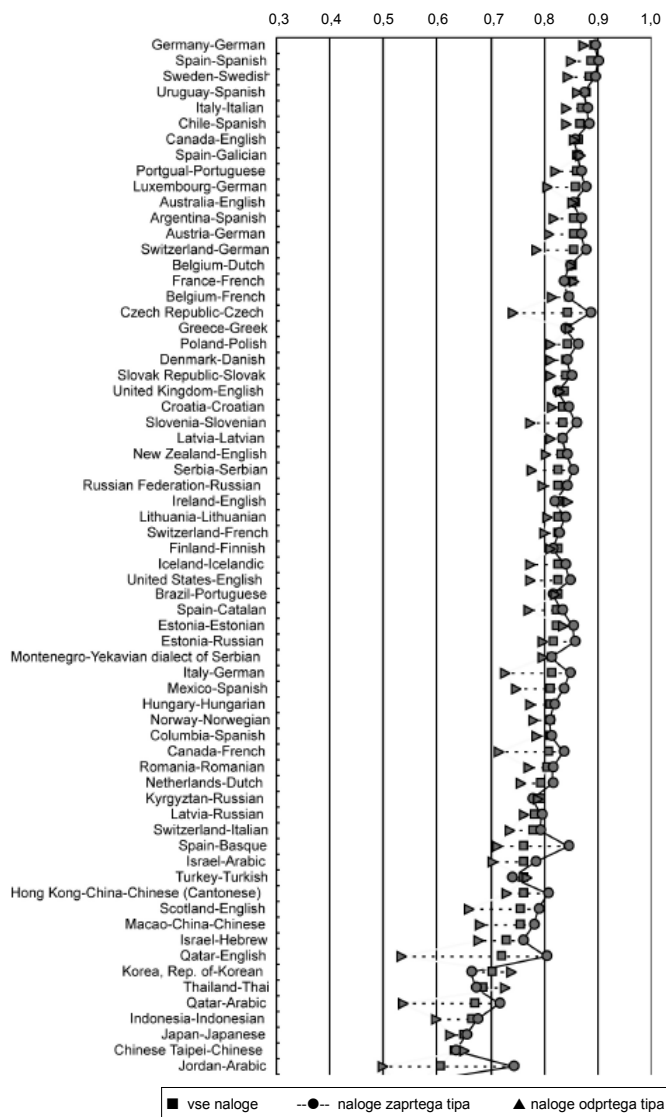
Z nekaj dodatnimi analizami smo želeli preveriti, ali bi drugače sestavljen niz nalog v preizkusu lahko bistveno vplival na delež variance skupnih faktorjev, še zlasti v državah z visokimi vrednostmi specifične variance.

Vpliv tipa nalog

Najprej smo z dvema neodvisnima faktorskima analizama izračunali dve vrednosti variance skupnih faktorjev, ki bi ju dobili, če bi bil preizkus iz naravoslovja v raziskavi PISA 2006 sestavljen izključno iz vprašanj zaprtega tipa

(tj. enostavnih ali kompleksnih nalog obkroževanja) ali izključno iz vprašanj odprtega tipa (tj. enostavnih ali strukturiranih nalog, kjer mora učenec sam oblikovati odgovor) (glej shemo 3). V analizo vprašanj zaprtega tipa sta bili vključeni dve tretjini od skupno 201 naloge iz naravoslovja (tj. 126 nalog), preostalo tretjino (75 nalog) pa smo uporabili pri analizi vprašanj odprtega tipa. S pomočjo sheme 4 lahko primerjamo deleže variance skupnih faktorjev, ki jih dobimo, če uporabimo naloge vseh tipov (kvadratki), če uporabimo le naloge obkroževanja (krožci) ali le naloge odprtega tipa (trikotniki).

Shema 3: Varianca skupnih faktorjev za potencialne preizkuse, sestavljene samo iz vprašanj zaprtega tipa ali samo iz vprašanj odprtega tipa



Delež variance skupnih faktorjev je bil zanimivo nekoliko večji za preizkus, sestavljen samo iz vprašanj zaprtega tipa (skupni faktorji so tu pojasnili 81 odstotkov celotne variance, v primerjavi s samo 79 odstotkov pojasnjene variance za preizkus, sestavljen iz vseh tipov nalog). Variance skupnih faktorjev za preizkuse, sestavljene iz vprašanj odprtega tipa, so bile navadno nižje (pojasnjevale so 75 odstotkov celotne variance). Iz sheme 3 je razvidno, da je bila razlika v vrednosti variance skupnih faktorjev glede na tip nalog, iz katerih je bil sestavljen preizkus, za nekatere nacionalne različice precejšnja: za češko, za baskovsko, uporabljeno v Španiji, za angleško in arabsko, ki sta bili uporabljeni v Katarju, za arabski različici v Jordaniji in Tuniziji, še zlasti pa za preizkus v jeziku azeri v Azerbajdžanu, kjer je bil delež variance skupnih faktorjev pri preizkusu z nizom vprašanj zaprtega tipa skoraj dvakrat višji kakor pri preizkusu, sestavljenem iz vprašanj odprtega tipa.

Takšna ugotovitev je nekoliko presenetljiva. Pričakovali bi namreč, da bodo kulturne razlike in nevajenost nalog obkroževanja bolj vplivale na odgovore učencev pri nalogah zaprtega tipa kot pri nalogah odprtega tipa. Vendar pa so slednje bolj kakor naloge zaprtega tipa odvisne od kakovosti kodiranja odgovorov učencev – to pomeni, da pristranost, ki jo ugotavljamo v nekaterih državah z visokim deležem kazalnika specifične variance, lahko vsaj delno izvira tudi iz napak pri kodiranju, ki jih indeks zanesljivosti koderjev ne pokaže.

Vpliv težavnosti nalog

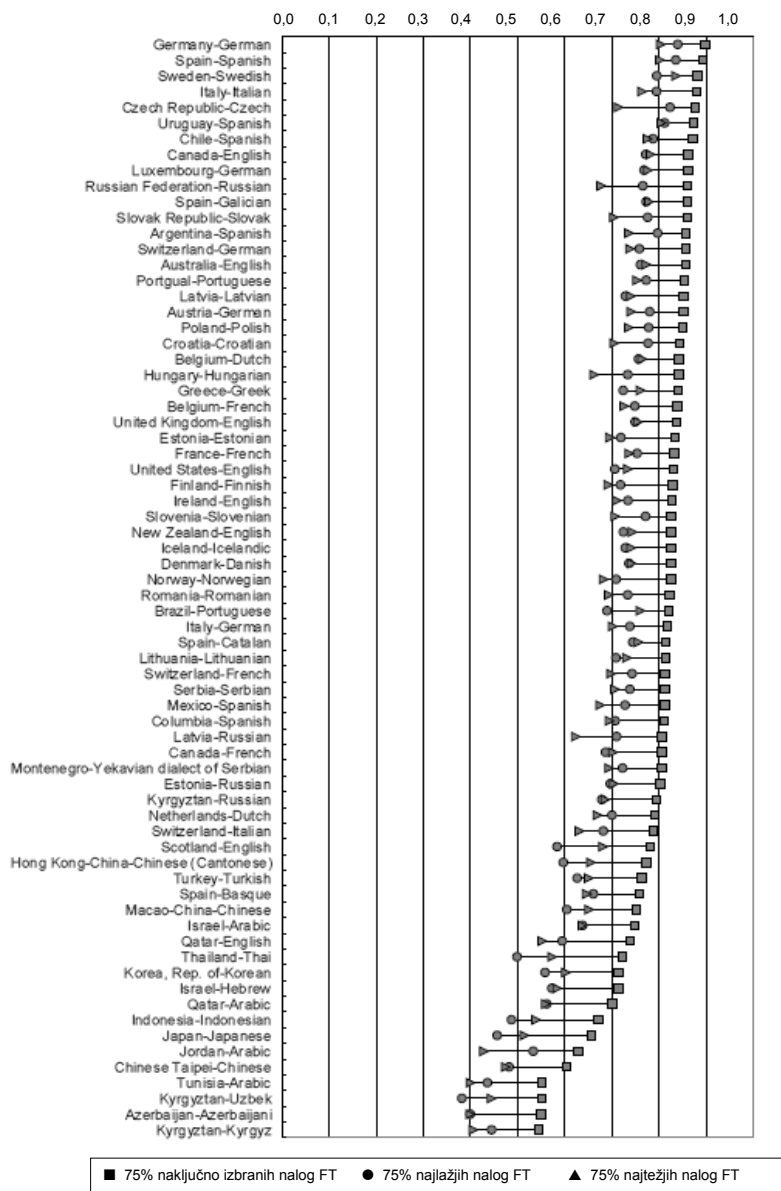
Analizo, prikazano na shemi 4, smo nadaljevali s tremi neodvisnimi faktorskimi analizami, v katere smo vključili: (i) 75 odstotkov najlažjih nalog iz predraziskave; (ii) 75 odstotkov najtežjih nalog; in (iii) 75 naključno izbranih nalog.

Iz sheme 4 je razvidno, da smo »najboljše« vrednosti specifične variance v vseh državah dobili takrat, ko smo uporabili naključni izbor lahkih, težkih in srednje zahtevnih nalog. Skupni faktor je v tej analizi pojasnil 79 odstotkov celotne variance, v primerjavi z 68-odstotnim deležem pojasnjene variance v analizi z večino lahkih in srednje zahtevnih nalog ali v analizi z večino težkih in srednje zahtevnih nalog. To najbrž pomeni, da se z znižanjem ali zvišanjem težavnosti preizkusa znanja ne bo izboljšala njihova enakovrednost.

Graf na shemi 4 pa kaže, da se »lahek« in »težek« preizkus znanja kljub približno enakemu deležu variance skupnih faktorjev v različnih državah ne bosta obnašala povsem enako. V nekaterih državah (kot so Japonska, Južna Koreja, Indonezija, Tajska, Škotska ali Brazilija) je bil delež variance skupnih faktorjev manjši za »lahek« kakor za »težek« preizkus. Za vrsto drugih nacionalnih različic preizkusa (Češka, Rusija, Slovaška, Hrvaška, Madžarska, ruska različica v Latviji, Argentina in Jordanija) je bilo ravno obratno.

Da bi raziskali še druge možne povezave med težavnostjo nalog in povprečnimi dosežki skupin učencev, ki so reševali različne verzije preizkusa iz naravoslovja v raziskavi PISA 2006, smo izračunali povprečne vrednosti pozitivnih ali negativnih razlik v delovanju nalog (*positive or negative DIF*), ki smo jih za vsako nacionalno različico preizkusa dobili iz 25-odstotnega deleža »najlažjih« in »najtežjih« nalog.

Shema 4: Varianca skupnih faktorjev za potencialne preizkuse, sestavljene iz lažjih ali težjih nalog



V splošnem je bila skupina najlažjih nalog manj stabilna od skupine najtežjih nalog. Zdelo se je, da so bile v predraziskavi PISA 2006 najlažje naloge za učence v večini držav z nizkimi povprečnimi rezultati (razen Mehike) in za učence iz Japonske in Južne Koreje težje, kot je bilo pričakovano. Nasprotno pa so te naloge nekoliko bolje od pričakovanj reševali na Finskem in Češkem.

Podoben »zrcalni« trend se je v določeni meri pokazal tudi pri najtežjih nalogah, ki so se v vrsti držav z nizkimi dosežki izkazale za lažje, kot je bilo pričakovati. Ne glede na to pa je bil delež tako pozitivnih kot negativnih razlik v delovanju nalog na splošno nižji za skupino težkih kakor za skupino lahkih nalog.

Primerjava kazalnikov enakovrednosti preizkusov iz matematike v raziskavah PISA 2003 in TIMSS 2003

Nazadnje smo opravili še analizo podatkov, ki smo jih dobili pri preverjanju znanja iz matematike v raziskavah PISA in TIMSS leta 2003, da bi tako ugotovili, ali je zgoraj opisani vzorec rezultatov značilen samo za naravoslovje ali velja tudi za matematiko. Raziskati smo želeli tudi, ali se manjša varianca skupnih faktorjev v ne-zahodnih državah, ki sodelujejo v raziskavi PISA, na enak način pojavlja tudi v raziskavi TIMSS, oziroma ali je v obeh raziskavah značilna za določeno skupino sodelujočih držav.

Za raziskavo TIMSS so bili podatki o težavnosti nalog v obliki povprečnih deležev pravih odgovorov po nalogah in po državah dostopni preko spleta, zato smo enake podatke izračunali tudi za naloge PISA in jih namesto parametrov delta uporabili v faktorskih analizah.

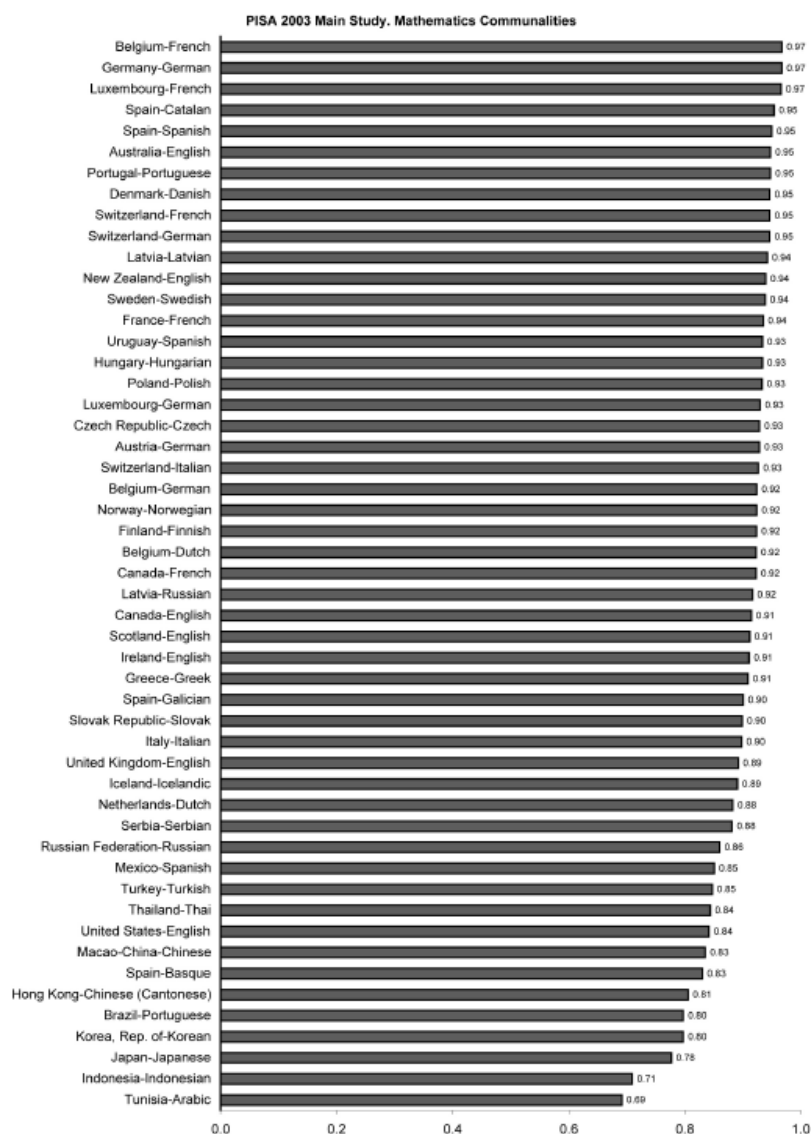
Rezultati, prikazani na shemi 5, so zanimivi z mnogih plati. Prvič, deleži variance skupnih faktorjev za matematiko v raziskavi PISA 2003 se v povprečju zdijo precej višji od tistih, ki so za naravoslovje v raziskavi PISA 2006 razvidni iz sheme 2 (varianca skupnih faktorjev predstavlja 91,3 odstotka celotne variance za matematiko in 79,2 odstotka celotne variance za naravoslovje). Enak trend je značilen tudi za TIMSS 2003, kjer je delež variance skupnih faktorjev skoraj v vseh državah višji za matematiko kot za naravoslovje.⁴ Sklepamo lahko, da je kurikulum za matematiko najbrž bolj univerzalen od kurikula za naravoslovje, kjer se učna snov od države do države bolj razlikuje.

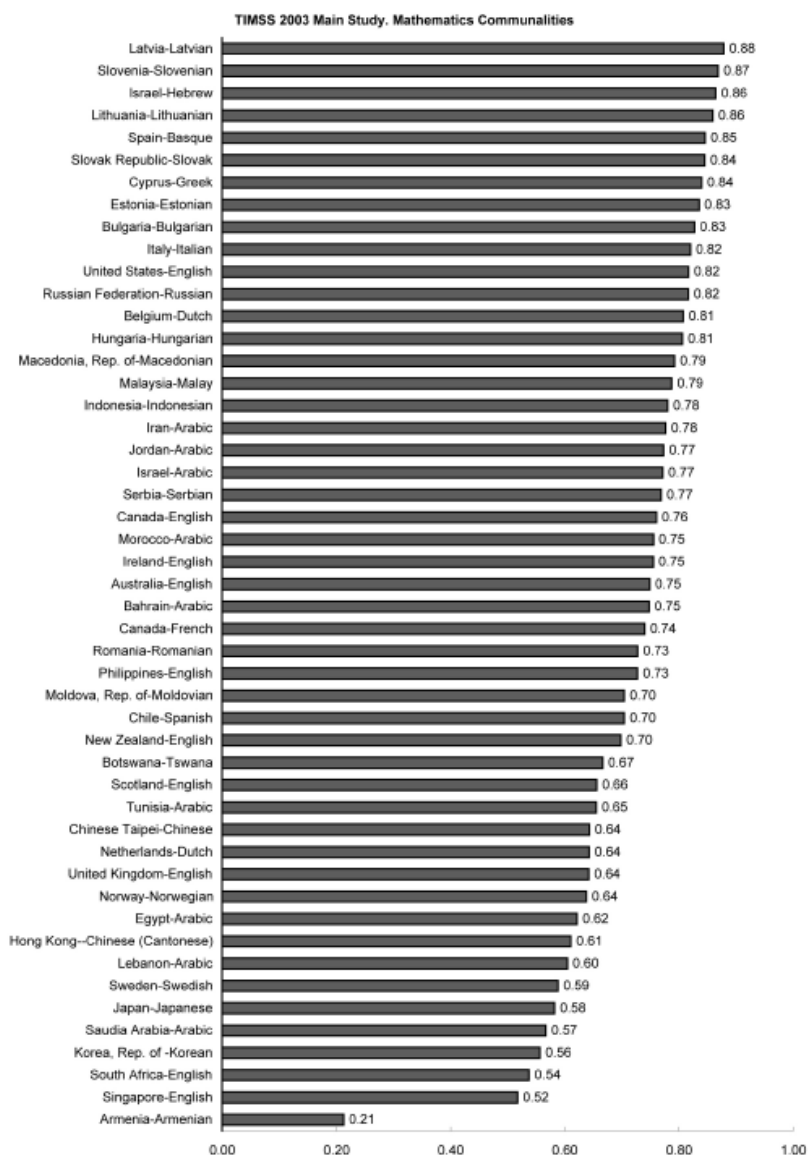
Drugič, deleži variance skupnih faktorjev za matematiko so v splošnem, presenetljivo, precej nižji za TIMSS kakor za raziskavo PISA (v povprečju predstavljajo 71,9 odstotka celotne variance za TIMSS in 91,3 odstotka celotne variance za raziskavo PISA). Tako veliko razliko lahko morda pojasnimo z (i) večjo kulturno homogenostjo držav, ki sodelujejo v raziskavi PISA, in (ii) z dejstvom, da so merski instrumenti raziskave TIMSS bolj neposredno povezani s kurikulumom. Iz

⁴ V tem prispevku uporabljamo samo podatke iz predraziskave PISA, zato raziskav PISA in TIMSS ne bomo neposredno primerjali.

sheme 6 je razvidno, da se je povprečna razlika v deležu variance skupnih faktorjev med obema raziskavama prepolovila (ne pa izginila: še vedno je predstavljala 77,5 odstotkov celotne variance v raziskavi TIMSS in 87,5 odstotkov v raziskavi PISA), ko smo analizo ponovili le na skupini 22 držav, ki so sodelovale v obeh raziskavah, in le s tistimi nalogami TIMSS, ki so vsebovale enak delež vprašanj s področja števil, algebre, geometrije, merjenja in uporabe podatkov kot naloge iz raziskave PISA.

Shema 5: Preverjanje znanja iz matematike v raziskavah PISA in TIMSS 2003. Varianca skupnih faktorjev pri težavnosti nalog.





Tretjič, neodvisno od teh razlik med raziskavama TIMSS in PISA smo odkrili tudi nekatere podobnosti v distribucijskih vzorcih variance skupnih faktorjev glede na evropske in neevropske države oziroma glede na industrializirane države in države v razvoju. Pri obeh raziskavah imajo najnižje deleže variance skupnih faktorjev zlasti azijske države (Hong Kong, Japonska, Kitajski Tajpej, Južna Koreja, Singapur) in nekatere arabsko govoreče države (Tunizija, Egipt, Libanon, Savdska Arabija). Za 22 držav, ki so sodelovale v obeh raziskavah, je korelacija rangov v ranžirni vrsti variance skupnih faktorjev za raziskavo TIMSS in raziskavo PISA znašala 0,63 ($p < 0,001$).

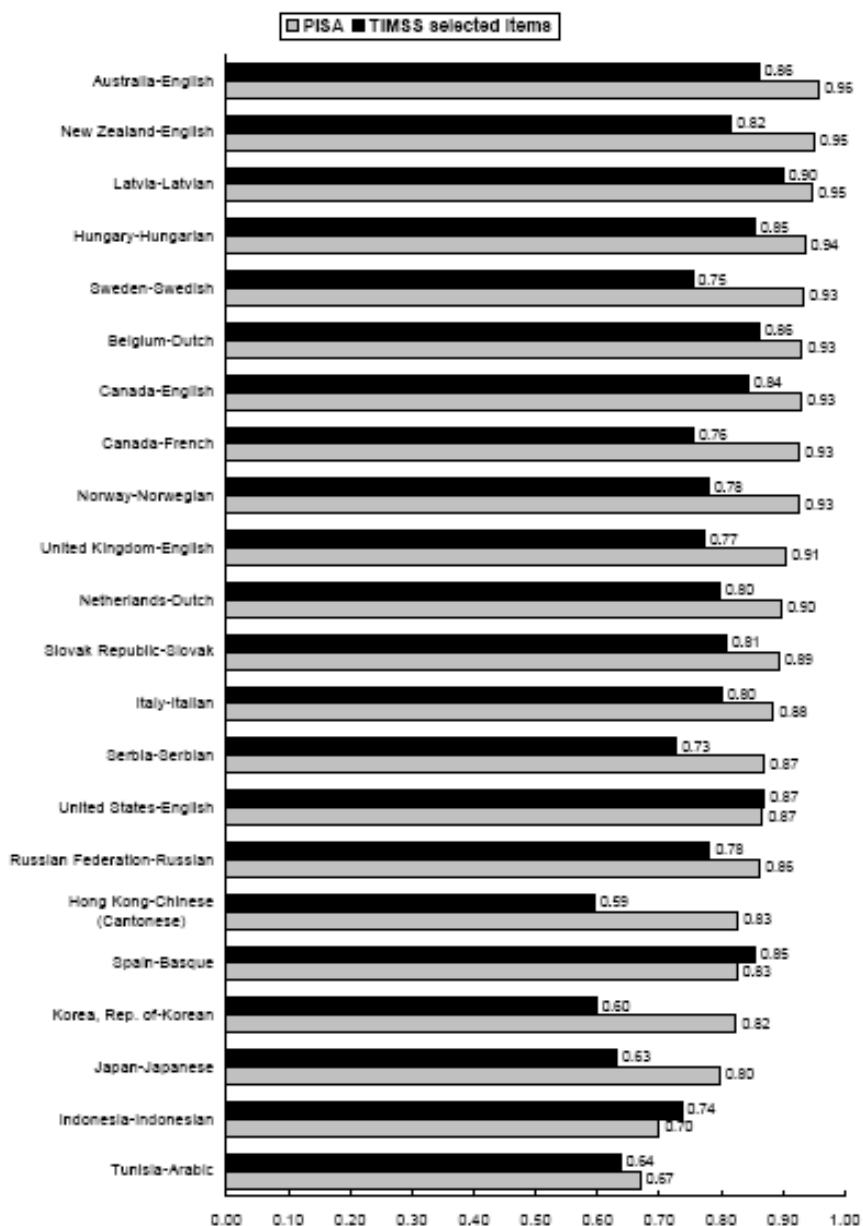
Razprava

Enakovrednost mednarodnega preizkusa lahko opredelimo kot enako verjetnost, da bodo pri neki nalogi pravilno odgovorili vsi učenci z določeno stopnjo znanja, ne glede na to, katero nacionalno različico preizkusa so reševali. Mnogo let pred uvedbo matematičnih modelov pojasnjevanja odgovorov je bila običajna praksa Mednarodnega združenja za evalvacijo učinkov izobraževanja (*International Association for the Evaluation of Educational Achievement – IEA*) primerjava vrstnega reda sodelujočih držav po težavnosti nalog s povprečnim deležem pravih odgovorov na mednarodni ravni. Zbiranje tovrstnih podatkov pa žal ni bilo redno, zato jih v mednarodnih raziskavah ne moremo uporabiti za primerjanje merskih instrumentov po kakovosti.

Študija bralne pismenosti združenja IEA je bila izjema: v dodatku mednarodnega poročila je Elley (1992) navedel, da je povprečna korelacija med rangom v nacionalni in rangom v mednarodni ranžirni vrsti po težavnosti nalog okrog 0,92 za preizkus bralnega razumevanja, ki so ga reševali 9-letni učenci, in 0,91 za preizkus, ki so ga reševali 14-letniki. V svojem komentarju teh podatkov Elley domneva, da sicer »lahko najdemo nekaj manjših značilnosti, ki bi jih redki opazovalci lahko imeli za učinek kulturne pristranosti«, vendar pa njihov vpliv verjetno ne bo večji od variance pristranosti, ki jo v vsakem nacionalnem preizkusu lahko pojasnimo s kulturnimi dejavniki. Elley zaključí, da se rezultati preizkusov IEA/RLS (*Reading Literacy Study*) »zdijo primerljivi tako med državami kakor znotraj posameznih držav.«

Takšen optimizem pa je v očitnem nasprotju z dognanji nedavne raziskave, ki jo sta jo izvedla Ercikan in Koh (2005); v njej sta preučila primerljivost angleške in francoske različice preizkusa TIMSS 1995, ki sta bili uporabljeni v Angliji, Franciji in ZDA. S pomočjo konfirmatorne faktorske analize sta testirala model enakovrednosti meritev (tj. model, ki upošteva invarianco številnih faktorjev, faktorskih uteži za posamezne naloge, napak v nalogah in korelacij med faktorji iz različnih podatkovnih nizov) za vsakega od osmih delovnih zvezkov iz matematike in naravoslovja, ki so jih učenci reševali v raziskavi TIMSS. Dobljene vrednosti RMSEA so pokazale dobro ujemanje celotnega modela s podatki za tri delovne zvezke iz matematike in enega iz naravoslovja v primerjavi Kanada (angl.) / Kanada (franc.) in primerjavi Anglija / Francija, a za nobenega od osmih matematičnih ali naravoslovnih delovnih zvezkov iz primerjave ZDA / Francija.

Shema 6: Varianca skupnih faktorjev težavnosti nalog za 22 držav, ki so sodelovale tako v raziskavi PISA 2003 kot v raziskavi TIMSS 2003



V zaključku svojega članka Ercikan in Koh navajata, da sta ugotovila precejšnje razlike med francoskimi in angleškimi preizkusi za preverjanje matematike in naravoslovja. Zapišeta, da »lahko razlike, nastale v postopku prilagajanja preizkusov, in jezikovne razlike, od katerih so lahko odvisni dosežki

sodelujočih učencev, vplivajo na enakovrednost konstruktov, ki jih preverjamo v različnih državah«, in zaključita, da »dognanja te študije kažejo na razlike v konstrukti, ki jih v različnih državah merimo v raziskavi TIMSS in poudarjajo pomen empiričnih dokazov, s katerimi je treba podpreti primerljivost konstrukta, preden lahko podatke raziskave TIMSS uporabimo v nadaljnjih raziskavah.«

Analize, ki smo jih izvedli v tem prispevku, niso bile niti približno tako natančne kot konfirmatorna faktorska analiza, ki sta jo opravila Ercikan in Koh, pa vendar so potrdile potrebo po bolj sistematičnem poročanju o enakovrednosti nacionalnih različic merskih instrumentov v mednarodnih raziskavah.

Korelacije, ki jih navaja Elley (1992), so pravzaprav zelo podobne primerljivim koeficientom, izračunanim za preizkus bralnega razumevanja v raziskavi PISA 2000, kar v obeh študijah kaže, da so v veliki večini sodelujočih držav nacionalne različice merskih instrumentov delovale na zelo konsistenten način.

Vendar pa v obeh študijah nekaj najnižjih korelacij opazamo v azijski skupini sodelujočih držav (med 0,75 in 0,85 za Singapur, Hong Kong in Tajsko v študiji IEA/RLS ter za Hong Kong, Južno Korejo, Japonsko, Tajsko in Indonezijo v raziskavi PISA 2000). Poleg tega se je nekaj zgoraj opisanih vzorcev preizkusa iz naravoslovja v raziskavi PISA 2006 pokazalo tudi v rezultatih študij IEA/RLS in TIMSS – korelacije so bile na primer relativno nižje v državah v razvoju kakor v industrializiranih državah. V državah v razvoju z neindoevropskimi jeziki, kot sta Bocvana ali Nigerija, so bile korelacije študije IEA/RLS še posebej nizke.

Nič presenetljivega ni, da so bili merski instrumenti študije IEA/RLS in prvih treh raziskav PISA (morda pa tudi ostalih mednarodnih raziskav) jezikovno in kulturno primernejši za skupino zahodnih držav, v katero se uvršča večina sodelujočih držav.

Analize v tem prispevku nakazujejo, da je bila enakovrednost merskih instrumentov v nekaterih drugih skupinah držav slabša, kar je vplivalo na nacionalne podatke teh držav. To je bilo še posebej značilno za vrsto bližnjevzhodnih in azijskih držav in bi lahko onemogočilo primerjave podatkov posamezne države z ostalimi državami, ki imajo znotraj te skupine iste jezikovne ali geografske značilnosti. Denimo, da bi želeli primerjati Tunizijo z ostalimi tremi arabsko govorečimi državami, ki so sodelovale v raziskavi PISA, glede na podatke za naravoslovje, ki smo jih opisali v tem prispevku. Povprečni delež pristranosti, ki vpliva na rezultate katerekoli izmed teh držav, ne bi bil nič manj kot 33 odstotkov skupne variance težavnosti nalog. Če bi v primerjavo vključili še Japonsko, Južno Korejo, Hong Kong in Tajpej, bi bil povprečni delež pristranosti 30-odstoten. Nasprotno pa bi delež pristranosti v isti primerjavi med Nemčijo in tremi ostalimi nemško govorečimi državami, ki so sodelovale v raziskavi PISA 2006, znašal le 11 odstotkov.

Ti rezultati jasno kažejo, da so še zlasti za arabske in kitajske različice merskih instrumentov potrebne dodatne, bolj poglobljene analize, da bi ugotovili, v kolikšni meri je problem odvisen od prevajanja in prilagajanja gradiva. Nekateri verifikatorji prevodov so predlagali, da bi za pomoč pri prevajanju v neindoevropske jezike pripravili posebne smernice za prevajalce. Mednarodna skupina strokovnjakov za prevajanje bi lahko s pomočjo skupine prevajalcev, ki so naravni govorniki ciljnega jezika, opravila analizo tistih nalog, ki so se obnašale v nasprotju s pričakovanji, podobno kot so to v svoji študiji angleških in francoskih različic za kanadsko raziskavo SAIP (*School Achievement Indicators Program*) storili Ercikan in ostali (2004).

Skupina svetovalcev za tehnično izvedbo raziskave PISA pa bo morala razmisliti, ali je smiselno določiti standard (s pomočjo kazalnikov, ki smo jih predlagali v tem prispevku, ali kakšnih drugih), na podlagi katerega bi identificirali tiste države, v katerih nacionalne različice instrumentov latentnega konstrukta ne merijo na način, ki je dovolj primerljiv z ostalimi državami, da bi omogočal zanesljive mednarodne primerjave.

Viri

- Baye, A. (2004). La gestion des spécificités linguistiques et culturelles dans les évaluations internationales de la lecture, *Politiques d'éducation et de formation*, 11, 55-70.
- Blum, A., Goldstein, H. in Guerin-Pace, F. (2001). An analysis of international comparisons of adult literacy. *Assessment in Education*, 8(2), 225-246.
- Elley, W. B. (1992). *How in the world do students read? IEA study of reading literacy*. The Hague: IEA.
- Ercikan, K., Gierl, M. J., McCreith, K., Puhan, G., Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education*, 17(3), 301-321.
- Ercikan, K. in Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS, *International Journal of Testing*, 5(1), 23-35.
- Grisay, A. (2003a). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing*, 20(2), 225-240.
- Grisay, A. (2003b). *PISA 2000: Differences in item difficulty between English, French and German countries*, Unpublished PISA TAG paper.
- Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests, *European Journal of Psychological Assessment*, 10, 229-240.

- Hambleton, R. K., Merenda, P. F. in Spielberger, C. D. (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum.
- Le, L. (2006, April). *Analysis of differential item functioning in PISA 2006*. Paper prepared for the annual meeting of the American Educational Research Association, San Francisco.
- McQueen, J. in Mendelovits, J. (2003). Cultural equivalence in a cross-cultural study. *Language Testing*, 20(2), 208-225.
- OECD (2007). *Technical standards for PISA 2009*. OECD document [GB(2007)4/REV1]. Retrieved from OECD, Directorate for Education web site: <http://www.pisa.oecd.org>
- Van de Vijver, F. J. R. (1998). Towards a theory of bias and equivalence, *ZUMA-Nachrichten Spezial*, 3, 41-52.
- Zumbo, B. D. (2003). Does item level DIF manifest itself in scale level analyses? Implications for translating language tests. *Language testing*, 20(2), 136-147.

KAKO SE PRI MERJENJU STALIŠČ IZOGNITI KULTURNO POGOJENI NARAVNANOSTI K IZBIRANJU SKRAJNIH ODGOVOROV?

Maurice Walker

Avstralski inštitut za raziskovanje v izobraževanju, Melbourne

S pomočjo Raschevih modelov teorije odgovora na postavko želimo na podatkih predraziskave PISA 2006 pokazati, da je pri odgovarjanju na vprašanja Likertovega tipa, ki ugotavljajo stališča učencev, naravnost k izbiranju skrajnih odgovorov podobno izražena v državah s podobnim kulturnim ozadjem. V tem prispevku si bomo eno lestvico stališč podrobneje ogledali v osmih kulturno različnih državah. Raziskali bomo dve poti, kako razrešiti problem pogostega izbiranja skrajnih odgovorov: prvič, z uporabo dihotomnih variant postavk, in drugič, z upoštevanjem naravnosti, značilne za posamezno državo, v Raschevem modelu pojasnjevanja odgovorov. Analiza različnih tipov postavk pokaže, da so rezultati in korelacija z dosežki učencev pri vseh podobni, vendar s postavkami Likertovega tipa dobimo več informacij o preizkusu. Posebna oblika hierarhične analize razkrije, da se podatki bistveno bolje ujemajo z modelom, ki ne zanemari medsebojnega vpliva države in parametrov delta posameznih postavk. Nekoliko orišemo tudi, kaj bi bilo treba upoštevati pri merjenju stališč s pomočjo postavk Likertovega tipa v različnih kulturnih okoljih.

Uvod

V mednarodnih primerjalnih raziskavah, kot so študije s področja izobraževanja Mednarodne zveze za evalvacijo dosežkov v izobraževanju (*International Association for the Evaluation of Educational Achievement – IEA*) ali Organizacije za ekonomsko sodelovanje in razvoj (*Organization for Economic Co-operation and Development – OECD*), se za merjenje stališč, prepričanj, mnenj in lastne aktivnosti učencev pogosto uporablja vprašalnike. Eden od glavnih razlogov za zbiranje tovrstnih podatkov – odslej jih bomo imenovali s skupnim terminom *spremljajoči podatki* – je pridobiti dodatne informacije, s katerimi bi morda lahko pojasnili nekatere značilnosti dosežkov teh učencev. Povezave med spremljajočimi podatki in dosežki učencev so pogosto od države do države različne (Kirsch et al., 2002; Martin, Mullis, Gonzales in Chrostowski, 2004; Mullis, Martin, Gonzales in Chrostowski, 2004; OECD, 2001, 2004), obstaja pa vrsta domnev o razlogih za takšno raznolikost. Na primer, Martin in ostali (2004) menijo, da so razlike med občutkom za lastno učinkovitost in dosežki na področju naravoslovja pogojene s kulturnim ozadjem učencev. Kirsch in ostali (2002) ugotavljajo, da so korelacije med pogostostjo branja in bralno pismenostjo najnižje v državah z nizkim povprečjem dosežkov na področju bralne pismenosti. Lie in Turmo (2005) opažata, da so učenci iz držav z nizkim povprečjem dosežkov s področja matematične pismenosti pogostejše bolj naklonjeni matematiki kakor njihovi vrstniki iz držav z visokim povprečjem matematične pismenosti.

Po mnenju nekaterih bi te ugotovljene razlike lahko delno pripisali pristranosti odgovorov. Paulhus (1991, str. 17) pristranost odgovora opredeljuje kot »sistematično naravnost, da na vrsto vprašanj v vprašalniku odgovarjamo na podlagi nekih drugih vsebin in ne zgolj na podlagi vsebine vprašalnika«. Po Paulhusu obstajajo trije glavni tipi pristranosti odgovorov:

- **socialna zaželenost** ali nagnjenost k podajanju odgovorov, za katere menimo, da nas »prikažejo v dobri luči«;
- **popustljivost**; z neko trditvijo se navadno raje strinjamo kakor ne strinjamo;
- **naravnost k izbiranju skrajnih odgovorov**; pri lestvicah stališč raje izbiramo odgovore na skrajnih koncih kakor tiste na sredini lestvice.

V tem prispevku bomo obravnavali le naravnost k izbiranju skrajnih odgovorov. Mnoge raziskave ugotavljajo medkulturne razlike v slogu odgovarjanja (Choi, Mericle in Harachi, 2006; Heine, Lehman, Peng in Greenholz, 2002), nekatere pa poleg tega zaključujejo, da je naravnost k izbiranju skrajnih odgovorov izrazitejša v točno določenih skupinah (Lee, Jones, Mineyama in Zhang, 2002; van Herk, Poortinga in Verhallen, 2004). Za pričujoči prispevek je ključnega

pomena trditev, da v mednarodnih primerjalnih študijah na področju izobraževanja kulturni dejavniki lahko vplivajo na slog odgovarjanja.

V mednarodnih primerjalnih študijah se spremljajoče dejavnike navadno meri z nizom postavk Likertovega tipa, ki tvorijo lestvico. Zaradi medkulturnih razlik pri odgovarjanju na tak tip vprašanj pa Heine trdi, da z njimi »bolj zanesljivo ugotavljamo razlike znotraj ene skupine kakor pa med različnimi skupinami« (Heine et al., 2002, str. 914). Medkulturna psihologija se pogosto ukvarja z vprašanjem, ali je medkulturne razlike pri odgovarjanju na vprašanja Likertovega tipa mogoče pojasniti glede na to, ali je neka kultura bolj nagnjena k individualizmu ali h kolektivizmu. Individualistična družba je tista, ki bolj poudarja cilje, vrednote in pravice posameznika, medtem ko kolektivistična družba daje prednost skupnim družbenim ciljem (za zgoščen pregled literature gl. Heine et al., 2002). Drugi dejavnik, ki naj bi vplival na pogostejše izbiranje skrajnih odgovorov, je pismenost anketiranca (npr. Flakerud, 1988): manj pismene osebe naj bi slabše razločevale pomenske odtenke besednih zvez, kot so *delno se strinjam*, *strinjam se* ali *zelo se strinjam*, zato naj bi ponavadi izbrale najpreprostejši odgovor, ki izraža njihovo mnenje (v tem primeru *strinjam se*). Še ena točka razprave je jezikovna enakovrednost prevodov danih odgovorov; v nekaterih jezikih lahko namesto *zelo se strinjam* pravzaprav iščemo ustrezen prevod odgovora *popolnoma se strinjam*.

Kakorkoli že, namen tega prispevka ni raziskati možnih vzrokov takšnih razlik. Zgolj dejstvo, da razlike obstajajo, ni sporno, le razlage, ki jih zanje najdemo v literaturi, so zelo različne. Študija, ki jo bomo izvedli, izhaja iz preprostega pragmatičnega vprašanja: če je naravnost k izbiranju skrajnih odgovorov pri odgovarjanju na postavke Likertovega tipa povezana s kulturnim ozadjem anketirancev, kako se ji v mednarodnih raziskavah izogniti?

Ena izmed možnosti je, da ugotovimo optimalno število možnih odgovorov za lestvico Likertovega tipa, ki bi izravnali kulturno pogojeno pristranost pri odgovarjanju na vprašanja. Lee in ostali (2002) so na primer ugotovili, da je veljavnost lestvice za merjenje *občutka za koherentnost (sense of coherence scale)* pri Kitajcih in Američanih večja, ko so možni odgovori štirje, pri Japoncih pa takrat, ko lahko izbirajo med sedmimi odgovori.

Težavo pogostega izbiranju skrajnih odgovorov pa bi še lažje rešili, če bi se preprosto izognili lestvicam stališč Likertovega tipa in bi uporabili le dihotomne postavke. T. i. popustljivosti pri odgovarjanju s tem sicer ne bi preprečili ali zmanjšali, naravnost k izbiranju skrajnih odgovorov pa bi postala nepomembna. Takšno razmišljanje je spodbudilo pričujočo raziskavo. Da bi lahko primerjali Likertovo in dihotomno različico vprašalnikov, sta bili v okviru predraziskave preverjanja naravoslovne pismenosti PISA 2006 pripravljeni dve verziji merskih instrumentov.

Ključna vprašanja raziskave so bila:

- Ali so razlike v naravnosti k izbiranju skrajnih odgovorov pri postavkah Likertovega tipa kulturno pogojene?
- Ali oblika danih odgovorov vpliva na ključne ugotovitve, do katerih pridemo s pomočjo vprašalnikov v primerjalnih študijah?
- Ali se je mogoče naravnosti k izbiranju skrajnih odgovorov izogniti že v fazi načrtovanja raziskave?

Metoda

V predraziskavi PISA 2006 sta bili naključno uporabljeni dve različici merskih instrumentov, ki sta merili *veselje do naravoslovja* (*enjoyment of science – ENJSCI*). Prva polovica učencev je reševala različico s postavkami Likertovega tipa, druga polovica učencev pa je dobila enake postavke v dihotomni obliki. V tem prispevku bomo najprej neposredno primerjali dobljene rezultate prvega in drugega niza postavk. Nato bomo rezultate postavk Likertovega tipa priredili v dihotomne tako, da bomo združili kategoriji *sploh se ne strinjam* in *ne strinjam se* ter *zelo se strinjam* in *strinjam se*, ter tako dobljene podatke primerjali z rezultati že prvotno dihotomnih postavk.

Vprašalnika v dveh izvirnih jezikih, angleščini in francoščini, sta najprej prevedla dva neodvisna prevajalca, sledilo je usklajevanje obeh prevodov in neodvisni jezikovni pregled. Kljub temu je mogoče, da so pomenski odtenki postavk vplivali na rezultate posameznih držav. Postopke prevajanja v raziskavi PISA natančno opisuje Grisay (2002).

Izvirna različica postavk vprašalnika o *veselju do naravoslovja* v angleščini je prikazana v tabeli 1. Pri postavkah Likertovega tipa v angleščini so bile kategorije možnih odgovorov *zelo se strinjam* (*strongly agree*), *strinjam se* (*agree*), *ne strinjam se* (*disagree*) in *sploh se ne strinjam* (*strongly disagree*); pri dihotomnih postavkah pa sta bili uporabljeni kategoriji *strinjam se* (*agree*) in *ne strinjam se* (*disagree*).

Tabela 1: Postavke za merjenje veselja do naravoslovja v predraziskavi PISA 2006

Postavka 1:	Ponavadi se ob učenju naravoslovnih tem zabavam. (<i>I generally have fun when I am learning science topics.</i>)
Postavka 2:	Rad berem o naravoslovju. (<i>I enjoy reading about science.</i>)
Postavka 3:	Uživam pri reševanju naravoslovnih problemov. (<i>I am happy doing science problems.</i>)
Postavka 4:	Rad pridobivam nova znanja iz naravoslovja. (<i>I enjoy acquiring new knowledge in science.</i>)
Postavka 5:	Zanima me učenje o naravoslovju. (<i>I am interested in learning about science.</i>)

V predraziskavi PISA 2006 so štiri vprašalnike naključno razdelili učencem vsake sodelujoče šole, da bi postavke preizkusili. Dva vprašalnika sta vsebovala dihotomne različice postavk za merjenje *veselja do naravoslovja*, dva pa Likertove različice. Vsi vprašalniki so vsebovali vrsto demografskih in drugih postavk, vključno s postavkami za merjenje odnosa do naravoslovja in motivacije zanj. Učenci so vprašalnike reševali po dveurnem preizkusu naravoslovne pismenosti. Rešili naj bi jih v 30 minutah, predvideni čas pa so po potrebi podaljšali.

Sodelujoče države

Primerjali bomo rezultate osmih držav, ki so sodelovale v predraziskavi PISA 2006. Izbrane so bile na podlagi jezikovne raznolikosti (vsaka država je uporabila preizkus v drugačnem jeziku), razlik v povprečnih dosežkih na področju naravoslovne pismenosti in geografske lege.¹ Izbira teh držav je bila namerno zasnovana tako, da je omogočala medkulturne primerjave. Četudi učenci znotraj posamezne države niso bili kulturno homogeni, je pouk v šolah te države potekal v enem jeziku, ki je bil drugačen od jezika poučevanja v preostalih sedmih državah. Iz tega lahko sklepamo, da so učenci iz posamezne države pripadali kulturno značilni družbi.

Namen predraziskave je bil preveriti učinkovitost merskih instrumentov, praktično izvedbo in postopke za obdelavo podatkov ter se s tem pripraviti na izvedbo glavne raziskave PISA. Poročanje o izsledkih predraziskave ni bilo predvideno, zato tudi vzorčenje ni bilo izpeljano tako, da bi bil vzorec povsem reprezentativen. Za namene predraziskave je bilo vzorčenje zasnovano tako, da je bil izbran priložnostni vzorec šol, ki je bil znotraj posamezne države dovolj reprezentativen glede vrste šole in učnega programa.

Tabela 2: Velikosti vzorcev za dve različici vprašalnika o veselju do naravoslovja

	dihotomni	Likertov
Država A	2745	2733
Država B	583	606
Država C	696	682
Država D	637	633
Država E	857	859
Država F	1104	1109
Država G	622	622
Država H	882	893
Skupaj (prečni vzorec/ <i>pooled sample</i>)	8126	8137

Ponavadi je bilo naključno izbranih 30–40 šol, znotraj vsake šole pa 30–40 učencev, starih 15 let. Nekatere države so se odločile za večji vzorec, da so lahko primerjale različne skupine učencev znotraj države (npr. glede na geografske regije). Zajem podatkov je potekal v letu 2005. Velikosti vzorcev so navedene v tabeli 2.

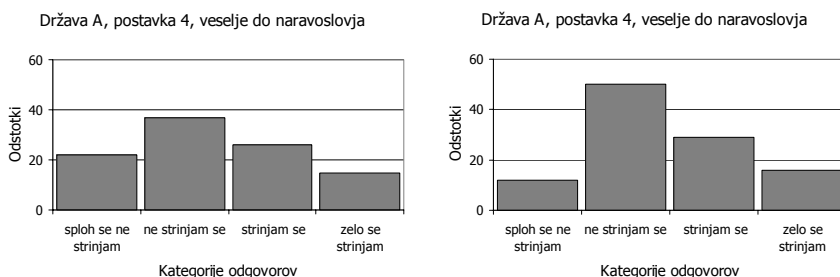
¹ Uporabili bomo podatke iz predraziskave PISA, ki še niso bili objavljeni, zato imena držav niso razkrita.

Rezultati

Razlike v naravnosti k izbiranju skrajnih odgovorov

Že ob prvem pregledu frekvenc posameznih odgovorov so se pokazali različni vzorci odgovarjanja po državah. Shema 1 prikazuje deleže posameznih odgovorov na postavko 4 (*Rad pridobivam nova znanja iz naravoslovja*) za dve državi – imenovali ju bomo država A in država B. Iz teh preprostih grafov je jasno razvidno, da so učenci iz države A pogosteje izbirali skrajne kategorije odgovorov, tako pritrdilne kot nikalne, kakor učenci iz države B. Tak vzorec je bil bolj ali manj konsistenten pri vseh petih postavkah, ki so merile veselje učencev do naravoslovja, vendar je bil najbolj izražen pri postavki 4, ki nam bo zato služila kot primer. Enako konsistentni vzorci odgovarjanja so se pokazali tudi v ostalih šestih državah, ker pa so bili najbolj izraziti prav v državah A in B, bomo za prikaz v prispevku uporabili ti dve državi.

Shema 1: Deleži posameznih kategorij odgovorov na postavko 4 (veselje do naravoslovja) v državah A in B

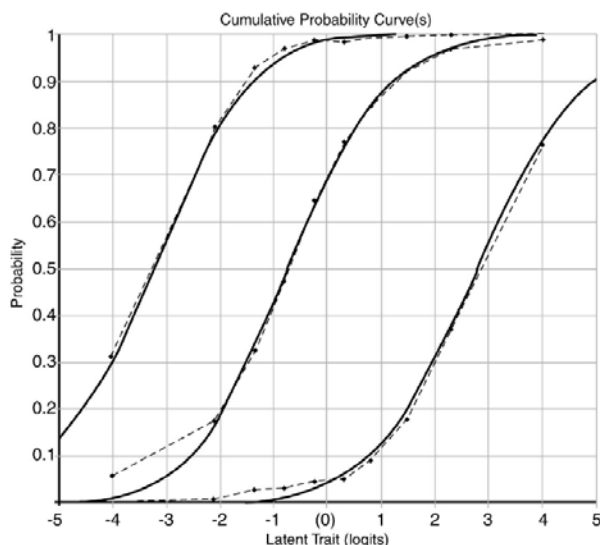


Toda čeprav že preproste frekvence odgovorov kažejo različne vzorce odgovarjanja, ne povedo ničesar o prikriti latentni potezi teh učencev.

Oba merska instrumenta smo zato pretvorili v lestvico s pomočjo modela za analizo dosežkov pri nalogah z delno pravilnimi odgovori (*Partial Credit Model*) (Masters, 1982) in aplikacije ConQuest (Wu, Adams in Wilson, 1997), ki jo je razvil Avstralski inštitut za raziskovanje v izobraževanju.

Prileganje postavk modelu smo ocenili s posebno ConQuestovo statistiko infit MS, ki temelji na razliki med opazovanimi in z modelom ocenjenimi vrednostmi (Wu, 1997). Ocenjene vrednosti so se gibale med 0,94 in 1,02 za parametre postavk in med 0,90 in 1,00 za parametre delta (Masters, 1982), kar kaže na dobro prileganje modelu. Vse analize ujemanja z modelom so bile opravljene na podatkih zbirnega vzorca.

Shema 2: Dejanske kumulativne krivulje verjetnosti za postavko 4 konstrukta veselje do naravoslovja (iz analize, kjer smo uporabili model dosežkov za naloge z delno pravilnimi odgovori /partial credit model) – za skupni vzorec (pooled sample)



Stopnjo prileganja modelu za postavko 4 lahko prikažemo s prekrivanjem predpostavljenih kumulativnih krivulj verjetnosti in podatkov zbirnega vzorca, kot je razvidno iz sheme 2. Dobljeni podatki (črtkana krivulja) na shemi natančno sledijo modelu (polna krivulja), kar dokazuje dobro prileganje. Shema prikazuje kumulativno verjetnost, da bo učenec izbral določeno kategorijo odgovora (ali višjo) glede na to, v kolikšni meri je pri njem izražena neka latentna poteza (v tem primeru *veselje do naravoslovja*). Če kategorije odgovorov točkujemo z 0 za *sploh se ne strinjam*, z 1 za *ne strinjam se*, z 2 za *strinjam se* in s 3 za *zelo se strinjam*, krivulji na levi strani prikazujeta verjetnost, da bo učenec z določeno stopnjo latentne poteze (prikazane na x-osi) dosegel vsaj 1 točko, krivulji na sredini verjetnost, da bo dosegel vsaj 2 točki, in krivulji na desni verjetnost, da bo dosegel vsaj 3 točke.

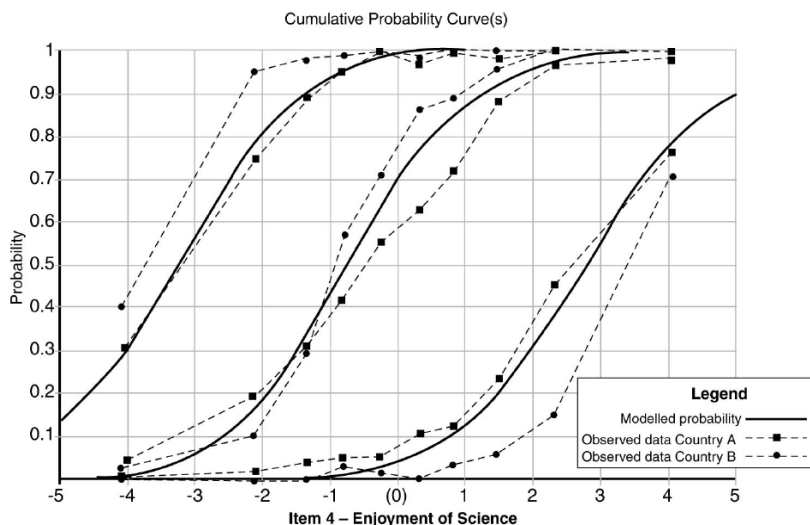
Ko pa smo podatke grupirali po državah in izmerjene vrednosti za vsako državo prikazali v istem modelu, so se razkrile precejšnje razlike v odgovorih med državami. Natančneje, podatki za posamezno državo se niso vedno prilegali modelu. Shema 3 prikazuje, kako se modelirane kumulativne krivulje verjetnosti za postavko 4 ujemajo z zbranimi podatki za državi A in B.

Zanimata nas zlasti levi in desni niz podatkov, ki se nanašata na verjetnost, da bo učenec izbral skrajno kategorijo odgovora. Krivulja na levi strani kaže, da je bila verjetnost, da bo učenec v državi B raje izbral *ne strinjam se* (ali višjo kategorijo odgovora) kot *sploh se ne strinjam*, veliko večja kakor za učenca iz

države A, ki ga naravoslovje enako veseli. Za učence iz države B, ki se nahajajo pri –2 točki *logit* na lestvici *veselja do naravoslovja*, je bila na primer verjetnost, da bodo izbrali neko drugo kategorijo odgovora in ne *sploh se ne strinjam*, okrog 0,96; ista verjetnost za enako uvrščene učence iz države A pa je bila le okrog 0,77. Z drugimi besedami – bolj verjetno je bilo, da bo učenec iz države A izbral skrajni odgovor *nikakor se ne strinjam*.

Krivulji na desni strani sheme 3 kažeta, da je bila verjetnost, da bo učenec iz države B izbral odgovor *zelo se strinjam*, veliko manjša kot za učenca A, ki ima enako izraženo *veselje do naravoslovja*. Verjetnost, da bo učenec iz države B pri vrednosti 2 točki *logit* na lestvici *veselja do naravoslovja* izbral odgovor *zelo se strinjam*, je bila okrog 0,11; verjetnost izbire takega odgovora za učenca z enakim *veseljem do naravoslovja* iz države A je bila 0,38. Povedano drugače – bolj verjetno je bilo, da bo učenec iz države A izbral skrajni odgovor *zelo se strinjam*. Podatki torej kažejo, da so učenci iz države B raje izbirali manj skrajne odgovore kakor učenci iz države A z enako izraženim *veseljem do naravoslovja*.

Shema 3: Dejanske kumulativne krivulje verjetnosti za postavko 4 konstrukta veselje do naravoslovja (iz analize, kjer smo uporabili model dosežkov za naloge z delno pravičnimi odgovori /partial credit model) – za državi A in B



Shema 3 izmed osmih držav in petih postavk res prikazuje najočitnejši primer neujemanja z modelom, vendar pa so bila takšna neujemanja, čeprav v manjši meri, tudi sicer precej pogosta.

Primerjava z različico v dihotomni obliki

Z ugotovljeno stopnjo neujemanja z modelom za posamezne države in nekaj zbranimi dokazi o razlikah v naravnosti k izbiranju skrajnih odgovorov smo

lahko dihotomne različice postavk primerjali s: 1) postavkami Likertovega tipa in 2) s postavkami Likertovega tipa, prirejenimi v dihotomno obliko.

V tabeli 3 so navedene povprečne ocene verjetnosti (*weighted likelihood estimate*, WLE, Warm, 1989) za vsako državo, in sicer po modelu dosežkov za naloge z delno pravilnimi odgovori (*partial credit model*) za Likertove različice postavk in po Raschevem modelu za dihotomne različice postavk. Vpisan je tudi rang glede na povprečje države in atenuirana korelacija med ocenjeno vrednostjo odnosa učencev do naravoslovja in njihovimi naravoslovnimi dosežki (WLE) za vsako državo.

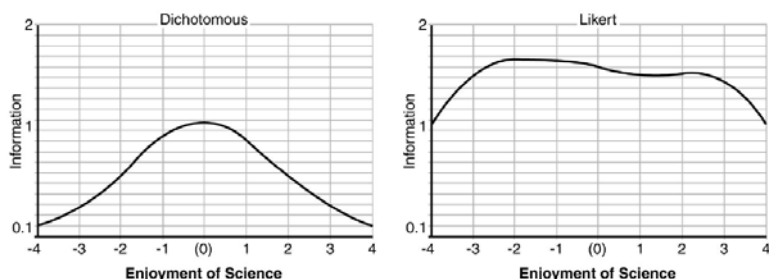
Tabela 3 kaže, da z uporabo preprostih Raschevih modelov različne metode zajema in obdelave podatkov privedejo le do neznatnih razlik v relativni razvrstitvi povprečnih vrednosti *veselja do naravoslovja*, korelacije z naravoslovnimi dosežki pa ostanejo dokaj konsistentne. Po teh kriterijih zato ne moremo nobene različice postavk izbrati kot najprimernejše.

Vendar pa z Likertovo lestvico dobimo več informacij o učencih z bolj ali manj izraženo latentno potezo. To je razvidno tudi iz grafov na shemi 4. Absolutne vrednosti sicer neposredno niso primerljive, krivulji (*test information curves*) iz obeh grafov pa se po obliki očitno razlikujeta. Z uporabo dihotomnih različic postavk smo dobili manj informacij o učencih, ki se na grafu umeščajo več kot eno standardno deviacijo od povprečja lestvice – torej o tistih, ki na x-osi ležijo onkraj vrednosti -1 in $+1$. Krivulja (*test information curve*) za Likertove različice postavk kaže, da le-te dajejo bolj konsistentne informacije za celotno populacijo – krivulja pade šele pri odklonu treh standardnih deviacij od srednje vrednosti. Ena izmed prednosti te krivulje je v tem, da stališča, kot je *veselje do naravoslovja*, obravnavamo kar kot izobraževalne dosežke. Raziskovalci zato pogosto preučujejo lastnosti učencev s trdnimi stališči in kako le-ta vplivajo na njihove dosežke.

Tabela 3: Povprečna vrednost veselja do naravoslovja (WLE), korelacija med veseljem do naravoslovja in naravoslovnimi dosežki ter rang v ranžirni vrsti povprečnih vrednosti veselja do naravoslovja za vsako državo. Rezultati, dobljeni s tremi metodami zajema in obdelave podatkov.

država	dihotomne postavke			Likertove postavke			Likertove postavke, prirejene v dihotomno obliko		
	korelacija z			korelacija z			korelacija z		
	povprečje naravoslovnimi	dosežki	rang	povprečje naravoslovnimi	dosežki	rang	povprečje naravoslovnimi	dosežki	rang
WLE				WLE			WLE		
A	-0,092	0,26	6	-0,010	0,21	6	-0,029	0,20	6
B	0,457	0,23	3	0,250	0,26	4	0,410	0,24	5
C	0,432	0,26	4	0,388	0,18	3	0,528	0,19	3
D	0,307	0,30	5	0,222	0,31	5	0,447	0,30	4
E	-0,509	0,31	8	-0,513	0,31	8	-0,541	0,28	8
F	-0,234	0,32	7	-0,376	0,32	7	-0,398	0,32	7
G	0,474	0,03	2	0,536	0,11	2	0,551	0,11	2
H	0,594	0,21	1	0,565	0,17	1	0,697	0,19	1

Shema 4: Krivulji (test information curves) za dihotojno in Likertovo različico konstrukta veselje do naravoslovja



Hierarhična analiza Raschevih modelov

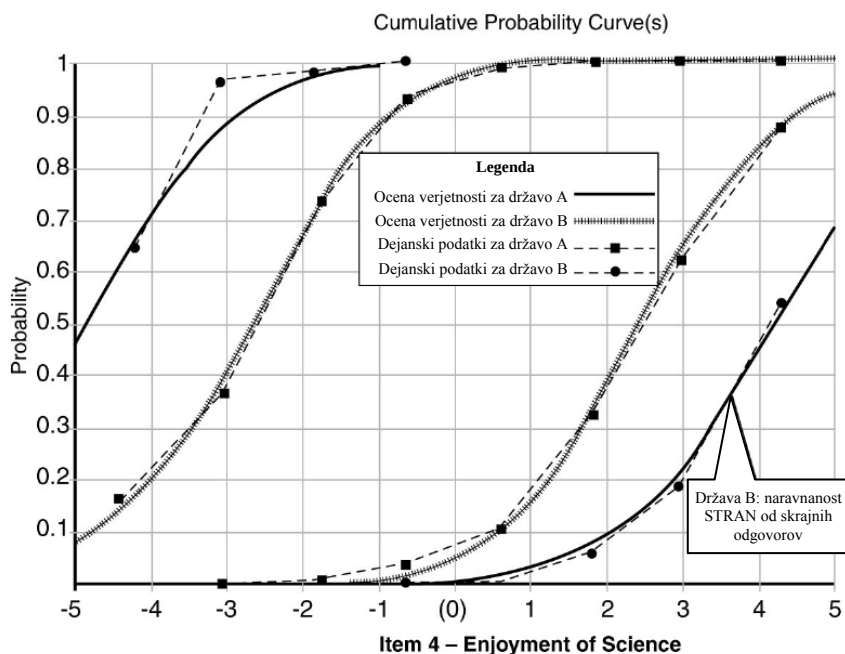
Konstrukte smo že raziskali s pomočjo osnovnega modela za analizo dosežkov pri nalogah z delno pravilnimi odgovori (*partial credit model*), sedaj pa bomo z uporabo multi-facetnega Raschevega modela (*multi-facet Rasch model*) (Linacre, 1994) preučili odstopanja nekaterih držav po težavnosti postavk in parametre posameznih kategorij odgovorov. V tabeli 4 so predstavljeni pogoji facetnega in ConQuestovega modela, ki ju bomo uporabili v analizi, in opisi učinkov, ki jih ima posamezni pogoj ali parameter.

Hierarhično smo testirali vrsto modelov, da bi našli najbolj ekonomičnega. Pri hierarhični analizi se zbirni vzorec najprej umeri po kompleksnem modelu, nato pa se iz modela enega za drugim izloča posamezne elemente, tako da postaja vse bolj enostaven. Nato primerjamo podatke o odklonih (*deviance statistics*) vsakega koraka iz postopka umerjanja. Če je razlika v odklonu statistično pomembna, se podatki bolje prilegajo kompleksnejšemu izmed dveh modelov, ki ju primerjamo. V tabeli 5 je predstavljenih sedem različnih modelov, ki smo jih primerjali med seboj, rezultati teh primerjav pa so navedeni v tabeli 6. Iz prvega stolpca tabele 6 je razvidno, katera dva modela primerjamo, zadnji pa kaže, da so bile vse primerjave statistično pomembne. Podatki se modelu 1 kot najkompleksnejšemu prilegajo precej boljše kakor ostalim.

Tabela 4: Opis pogojev modela v aplikaciji ConQuest

Pogoj modela ConQuest	Parameter	Opis predpostavljenega učinka
ITEM	d_i	Splošni učinek postavke
STEP	t_k	Splošni učinek kategorije odgovora na postavko
CNT	a_c	Splošni učinek države
ITEM*STEP	t_{ik}	Učinek medsebojnega vpliva postavke in kategorije odgovora na postavko
ITEM*CNT	b_{ic}	Učinek medsebojnega vpliva postavke in države
ITEM*STEP*CNT	g_{ick}	Učinek medsebojnega vpliva postavke, države in kategorije odgovora na postavko

Shema 5: Dejanske kumulativne krivulje verjetnosti za postavko 4 konstrukta veselje do naravoslovja (iz analize, kjer smo izmed multi-facetnih Raschevih modelov /multi-faceted Rasch models/ uporabili model 1) – za državi A in B



Shema 5 kaže dobro prileganje modela 1, saj se izmerjeni podatki za državi A in B ter teoretične kumulativne krivulje verjetnosti za postavko 4 precej natančno ujemajo. V tem prispevku se ne bomo posvečali krivuljam verjetnosti za srednje kategorije odgovorov (tj. verjetnosti, da bodo učenci izbrali kategorijo *strinjam se* ali višjo), zato smo jih zaradi boljše preglednosti izločili iz sheme 5; ta zato prikazuje le krivulje verjetnosti za skrajne kategorije odgovorov. Podatki se kompleksnemu modelu na shemi 5 prilegajo precej bolje kot preprostejšemu modelu s sheme 3.

Zaključimo lahko, da se podatki na splošno bolje ujemajo z modelom, ki upošteva medsebojno učinkovanje države, postavke in kategorije odgovora na postavko (delta). Učinke teh dejavnikov bomo v nadaljevanju upoštevali tudi pri razlagi podatkov o naravnosti k izbiranju skrajnih odgovorov.

Pregled ocenjenih vrednosti parametrov delta

Določili smo že najustreznejši model pojasnjevanja odgovorov, zato bomo v zadnji analizi, ki jo bomo opisali v tem prispevku, raziskali ocene parametrov delta za postavke Likertovega tipa. Kot smo že omenili, je hierarhična analiza pokazala, da se podatki precej bolje prilegajo najkompleksnejšemu modelu kot pa kateremu od preprostejših modelov, ki smo jih preiskali. Model, ki se najboljše ujema s podatki, upošteva medsebojni vpliv parametrov delta ter učinke države in postavke. Z drugimi besedami – to pomeni, da je pri vsaki postavki razkorak med posameznimi kategorijami odgovorov (med *sploh se ne strinjam* in *ne strinjam se*, med *ne strinjam se* in *strinjam se*, ter med *strinjam se* in *zelo se strinjam*) odvisen od države.

Iz tega lahko sklepamo, da je mogoče države, kjer učenci pogosteje izbirajo skrajne odgovore, odkriti tako, da preučimo parametre delta. V tabeli 7 so navedene ocene parametrov delta za državi A in B pri postavki 4.

Tabela 5: Hierarhična primerjava modelov v aplikaciji ConQuest za Likertove različice konstruktov

	Facet model formulation	ConQuest model term	Deviance	Number of parameters
Model 1	$\ln\left(\frac{p_{mic,k-1}}{p_{mic,k}}\right) = \theta_n - \alpha_c - \delta_i - \beta_{ic} - \gamma_{ick}$	CNT + ITEM + ITEM*CNT + ITEM*CNT*STEP	79612.7	121
Model 2	$\ln\left(\frac{p_{mic,k-1}}{p_{mic,k}}\right) = \theta_n - \alpha_c - \delta_i - \beta_{ic} - \tau_{ik}$	CNT + ITEM + ITEM*CNT + ITEM*STEP	80927.1	51
Model 3	$\ln\left(\frac{p_{mic,k-1}}{p_{mic,k}}\right) = \theta_n - \alpha_c - \delta_i - \tau_{ik}$	CNT + ITEM + ITEM*STEP	82846.5	23
Model 4 [the Partial Credit Model]	$\ln\left(\frac{p_{mic,k-1}}{p_{mic,k}}\right) = \theta_n - \delta_i - \tau_{ik}$	ITEM + ITEM*STEP	83040.8	16
Model 5	$\ln\left(\frac{p_{mic,k-1}}{p_{mic,k}}\right) = \theta_n - \delta_i - \tau_k$	ITEM + STEP	83374.7	8
Model 6	$\ln\left(\frac{p_{mic,k-1}}{p_{mic,k}}\right) = \theta_n - \alpha_c - \delta_i - \tau_k$	CNT + ITEM + STEP	83183.5	15
Model 7	$\ln\left(\frac{p_{mic,k-1}}{p_{mic,k}}\right) = \theta_n - \alpha_c - \delta_i - \beta_{ic} - \tau_k$	CNT + ITEM + ITEM*CNT + STEP	81228.4	43

Tabela 6: Razlike med sedmimi multi-facetnimi Raschevimi modeli (multi-faceted Rasch models), ugotovljene s pomočjo hi-kvadrat testa

testirani modeli	hi-kvadrat	stopnje prostosti	pomembnost
2 – 1	1314,4	70	$p < 0,001$
3 – 2	1919,4	28	$p < 0,001$
4 – 3	194,3	7	$p < 0,001$
5 – 4	333,9	8	$p < 0,001$
5 – 6	191,2	7	$p < 0,001$
6 – 7	1955,1	28	$p < 0,001$
4 – 6	142,6	1	$p < 0,001$
2 – 7	301,3	8	$p < 0,001$

Tabela 7: Parametri delta za četrto Likertovo postavko konstrukta veselje do naravoslovja, za državi A in B

	država A	država B
Parameter delta 1 (kategorija odgovora <i>ne strinjam se</i> ali višja)	-2,538	-4,263
Parameter delta 2 (kategorija odgovora <i>strinjam se</i> ali višja)	-0,009	-0,341
Parameter delta 3 (kategorija odgovora <i>zelo se strinjam</i> ali višja)	2,546	4,603
Razlika med parametrom delta 3 in delta 1	5,084	8,866

Shema 6: Razlike med prvim in tretjim parametrom delta za vse postavke, ki merijo veselje do naravoslovja, po državah

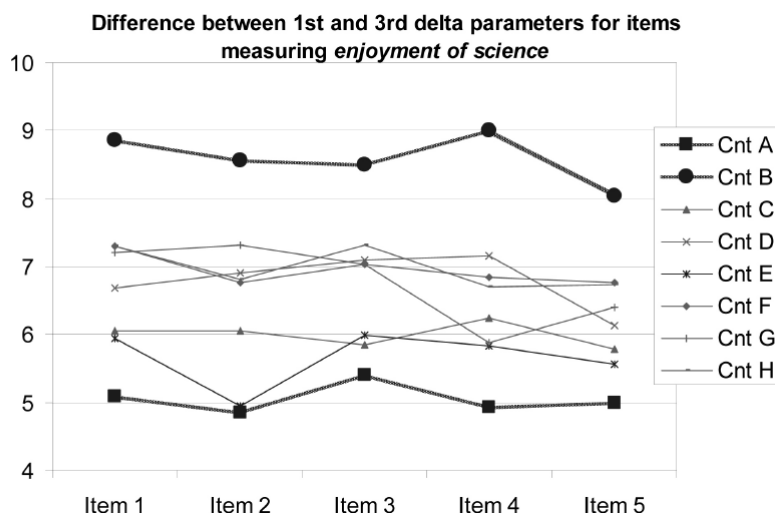


Tabela 7 prikazuje tudi razlike med deltama 1 in 3 – večja kot je razlika med njima, manj pogosto učenci v tej državi izbirajo skrajne odgovore. Tako lahko sklepamo, ker so v tem modelu že upoštevani morebitni splošni učinki države ali postavke ter učinki medsebojnega vpliva države in postavke. Ocene parametrov zato predstavljajo le medsebojni vpliv države in parametrov delta.

S tem v mislih lahko razliko med parametri delta 3 in 1 grafično prikažemo za vsako državo in za vsako postavko, ki meri nek konstrukt (shema 6).

Shema 6 razkriva konsistentne vzorce izbiranja skrajnih odgovorov pri vseh postavkah, ki merijo *veselje do naravoslovja*. Za državo A so razlike med prvim in tretjim parametrom delta dosledno nizke pri vseh postavkah (z velikostjo največ okrog pet), kar kaže, da učenci v tej državi pogosteje izberejo skrajne odgovore. Po drugi strani pa so razlike med prvim in tretjim parametrom delta za državo B pri vseh postavkah precejšnje (velikosti med osem in devet), kar kaže, da se učenci v tej državi izogibajo skrajnim kategorijam odgovorov. Takšna konsistentnost kaže, da naravnost k izbiranju skrajnih odgovorov ni odvisna od postavke – vsaj ne od niza postavk, ki merijo *veselje do naravoslovja*.

Zaključek

Vprašanje, ki je spodbudilo to analizo, je bilo: če je naravnost k izbiranju skrajnih odgovorov pri odgovarjanju na vprašanja Likertovega tipa res odvisna od kulturnih razlik, kako se ji lahko v mednarodnih raziskavah izognemo? Eden izmed predlogov je bil, da bi z uporabo dihotomnih postavk *ipso facto* lahko preprečili prepogosto izbiranje skrajnih odgovorov. Lestvice na podlagi preprostih Raschevih modelov so pokazale, da so vzorci korelacije med naravoslovnimi dosežki in *veseljem do naravoslovja* v vseh državah zelo podobni, in sicer tako pri postavkah Likertovega tipa, kot pri dihotomnih postavkah in pri postavkah Likertovega tipa, prirejenih v dihotomne z združevanjem več kategorij odgovorov. Kljub enakovrednosti na vseh ostalih področjih pa je Likertova različica postavk nekoliko v prednosti, saj nudi več informacij o tistem delu populacije, ki se po izraženosti latentne poteze uvršča nekoliko stran od povprečja.

Hierarhični test različnih facetnih modelov (*facet models*) pa je pokazal, da se podatki bolje prilegajo kompleksnejšemu Raschevemu modelu, ki upošteva medsebojni vpliv države in parametrov delta. Analiza postavk Likertovega tipa s pomočjo takšnega bolj kompleksnega modela je razkrila konsistentne vzorce, ki bi jih od države do države *morda* lahko interpretirali kot naravnost k izbiranju skrajnih odgovorov.

Pomembno vprašanje je, kdaj lahko neko značilnost odgovarjanja označimo za *ekstremno*. Nekdo bi lahko na primer trdil, da bi bilo o državah, ki smo jih v tem prispevku označili kot tiste z nizko stopnjo naravnosti k izbiranju skrajnih odgovorov, bolje govoriti kot o državah, kjer učenci najpogosteje izbirajo odgovore iz sredine lestvice.

Pregled, ki smo ga predstavili, je zelo omejen, saj se osredotoča le na en konstrukt (*veselje do naravoslovja*), ki smo ga v predraziskavi PISA 2006 merili z dvema vzporednima različicama postavk, Likertovih in dihotomnih. Z dodatnimi

analizami bi lahko ugotavljali skrite korelacije med tem konstruktom, izmerjenim in obravnavanim na dva različna načina, in ostalimi konstrukti iz vprašalnikov ali delovnih zvezkov. Morda bi prišli do drugačnih ugotovitev, če bi preizkusili več merskih instrumentov z dvema vzporednima različicama postavk, zlasti če le-ti ne bi bili tako močno povezani z naravoslovjem. Dokler ne bo raziskanih več konstruktov, kulturno pogojene naravnosti k izbiranju skrajnih odgovorov ni mogoče potrditi: učinki, o katerih smo poročali v tem prispevku, so morda značilni samo za konstrukt *veselje do naravoslovja*.

Zaključimo lahko, da je naravnost k izbiranju skrajnih kategorij odgovorov morda res značilnejša za nekatere države, vsaj pri nekaterih stališčih učencev. Če tehtamo različne možnosti, se zdi v mednarodnih primerjalnih raziskavah pri merjenju spremljajočih dejavnikov smiselno nadaljevati z uporabo postavk Likertovega tipa. Raziskovalci bi morali vse konstrukte, ki jih merimo s postavkami Likertovega tipa, natančno preiskati od države do države. Morebitno naravnost k izbiranju skrajnih odgovorov lahko izravnamo, če v analizah postavke Likertovega tipa pretvorimo v dihotomne. Enako pozornost bi morali posvetiti tudi izbiri analitičnega modela. Premami nas lahko, da bi v model vključili vidike, ki omogočajo (oz. ne onemogočajo) medsebojni vpliv države in parametrov delta posameznih postavk. Iz raziskovalnega vidika je ta možnost sicer privlačna, saj se podatki takemu analitičnemu modelu bolje prilagodijo, vendar po drugi strani takšen pristop zabriše kulturne učinke – pri raziskovanju dosežkov učencev pa nas včasih zanimajo prav ti. Povedano drugače, z upoštevanjem učinka države v teoretičnem modelu izravnamo nekatere razlike med državami; merjenje prilagodimo posamezni državi, a o rezultatih meritev poročamo na enotni lestvici. Nekomu, ki ni podrobno seznanjen s takšnim metodološkim pristopom, je težko pojasniti, zakaj se končni rezultat merjenja neke latentne poteze za učenca iz ene države, ki se je »strinjal« z vsemi trditvami, razlikuje od končnega rezultata učenca s povsem enakim vzorcem odgovarjanja iz neke druge države. Če se izkaže, da je naravnost k izbiranju skrajnih odgovorov statistično pomembna in konsistentna, bi bilo morda dobro o rezultatih poročati na dva načina: prvič prilagojeno pristranosti pri odgovarjanju in drugič ne.

Prihodnje raziskave bi lahko odkrile, da kulturno pogojena pristranost obstaja in je neodvisna od latentne poteze, ki jo želimo izmeriti – argument za upoštevanje te pristranosti pri oblikovanju teoretičnega modela bi tako imel veliko večjo težo. Povedali smo že, da v tem prispevku pristranosti nismo dokazali, le nakazali.

Viri

- Choi, Y., Mericle, A. in Harachi, T. W. (2006). Using Rasch analysis to test the cross-cultural item equivalence of the Harvard trauma questionnaire and the Hopkins symptom checklist across Vietnamese and Cambodian immigrant mothers. *Journal of Applied Measurement*, 7(1), 16-38.
- Flaskerud, J. H. (1988). Is the Likert scale format culturally biased? *Nursing Research*, 37(3), 185-186.
- Grisay, A. (2002). Translation and cultural appropriateness of the test and survey materials. V: R. J. Adams and M. Wu (ur.), *PISA 2000 technical report* (pp. 57-70). Pariz: OECD Publications.
- Heine, S. J., Lehman, D. R., Peng, K. in Greenholz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales?: The reference group effect. *Journal of Personality and Social Psychology*, 82(6), 903-918.
- Kirsch, I., de Jong, J., Lafontaine, D., McQueen, J., Mendelovits, J. in Monseur, C. (2002). *Reading for change: Performance and engagement across countries. Results from PISA 2000*. Pariz: OECD Publications.
- Lee, J. W., Jones, P. S., Mineyama, Y. in Zhang, X. E. (2002). Cultural differences in responses to a Likert scale. *Research in Nursing and Health*, 25, 295-306.
- Lie, S. in Turmo, A. (2005). *Cross-country comparability of student's self-reports: Evidence from PISA 2003*. Oslo, Norway: University of Oslo.
- Linacre, J. M. (1994). *Many-facet Rasch measurement*. Chicago: MESA Press.
- Martin, M. O., Mullis, I. V. S., Gonzalez, E. J. in Chrostowski, S. J. (2004). *TIMSS 2003 international science report*. Boston: TIMSS and PIRLS International Study Centre.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.
- Mullis, I. V. S., Martin, M. O., Gonzalez, E. J. in Chrostowski, S. J. (2004). *TIMSS 2003 international mathematics report*. Boston: TIMSS and PIRLS International Study Centre.
- OECD. (2001). *Knowledge and skills for life: First results from PISA 2000*. Pariz: OECD Publications.
- OECD. (2004). *Learning for tomorrow's world: First results from PISA 2003*. Pariz: OECD Publications.
- Paulhus, D. L. (1991). Measurement and control of response bias. V: *Measures of Personality and Social Psychological Attitudes* (Vol. 1, pp. 17-59). San Diego: Academic Press.

- van Herk, H., Poortinga, Y. H. in Verhallen, T. M. M. (2004). Response styles in rating scales: Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology*, 5(3), 346-360.
- Warm, W. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54, 427-445.
- Wu, M. L. (1997). *The development and application of a fit test for use with marginal maximum likelihood estimation and generalized item response models*. Unpublished masters thesis, University of Melbourne, Melbourne, Australia.
- Wu, M. L., Adams, R. J. in Wilson, M. R. (1997). ConQuest: Multi-aspect test software [Computer program]. Camberwell, Avstralija: Australian Council for Educational Research

.

VPLIV TRUDA UČENCEV NA IZIDE MEDNARODNIH RAZISKAV

Jayne Butler
Univerza v Melbournu

Raymond J. Adams
Avstralski inštitut za raziskovanje v izobraževanju
Univerza v Melbournu

Mednarodno preverjanje znanja učencev v raziskavah, kot sta Mednarodna raziskava trendov v znanju matematike in naravoslovja (TIMSS) in Program mednarodne primerjave dosežkov učencev (PISA), ima za razvoj šolske politike, ki temelji na podatkih, vse večji pomen. Takšne raziskave imajo lahko daljnosežni vpliv, zato je toliko bolj pomembno, da so veljavne in zanesljive. V izobraževanju se vse bolj poudarja motivacijske dejavnike, ki vplivajo na učenje. V tem prispevku bomo preučili, kaj bi lahko ogrozilo veljavnost tovrstnih raziskav tako, da bomo raziskali, kako trud učencev vpliva na njihove dosežke. K sreči se je že izkazalo, da je trud učencev po državah približno enak, kar spodbija trditev, da sistemske kulturne razlike na tem področju onemogočajo mednarodne primerjave. V resnici je trud učencev z bralnimi dosežki povezan približno toliko kakor ostali dejavniki, kot so enostarševska družina, spol in socialno-ekonomsko ozadje. Ko govorimo o trendih, bi torej morali upoštevati tudi trud, ki ga učenci namenijo preizkusu, saj ta lahko pomaga pri interpretaciji bralnih dosežkov glede na državo in spol.

**The Impact of Differential Investment of Student Effort on the Outcomes of
International Studies**

JOURNAL OF APPLIED MEASUREMENT, 8(3), 279–304

Copyright © 2007

Uvod

V popularni literaturi pogosto naletimo na pomislek, da preverjanje znanja, ki nima neposrednih posledic za učence, učitelje ali šolo, podcenjuje zmožnosti učencev, kar se še stopnjuje, ko učenci pobližje spoznajo preizkuse (Holliday in Holliday, 2003). To vprašanje je še posebej pomembno pri mednarodnih primerjalnih raziskavah, kot sta TIMSS in PISA. Številni menijo, da nepoznane regionalne in kulturne razlike v odnosu učencev do preizkusa znanja in njihovi motiviranosti resno ogrožajo veljavnost preverjanja, še zlasti takrat, ko se rezultate primerja po regijah, kulturah ali med leti (Bracey, 1999; Holliday in Holliday, 2003). Vprašanje razlik v odnosu učencev do preizkusa med državami, ki sodelujejo v mednarodnih raziskavah, se ponavlja načne takrat, ko neka država ne doseže pričakovanih rezultatov.

V strokovni literaturi najdemo nekaj študij o vplivu motivacije učencev na dosežke pri preverjanju znanja, a njihovo število je omejeno, izsledki pa nekoliko dvomljivi. Večinoma se osredotočajo na vprašanje, ali učenci k reševanju preizkusov, ki zanje niso odločilni (kot so na primer mednarodne primerjalne raziskave ali pilotni programi preverjanja znanja), pristopajo manj motivirani (Wise in DeMars, 2005). Ker tovrstni preizkusi za nekatere učence niso toliko pomembni, lahko pri reševanju pokažejo manj znanja, kakor ga imajo v resnici. Iz tega razloga mnogi poudarjajo, da skupni rezultati ne pokažejo pravega znanja in zmožnosti učencev, s čimer je ogrožena veljavnost rezultatov in njihove interpretacije (Kiplinger in Linn, 1996; Mislevy, 1995; O'Neil, Sugrue in Baker, 1996; Wainer, 1993; Wolf in Smith, 1995).

Preučevanje motivacije za reševanje preizkusov znanja se je pojavilo kot raziskovalno področje v okviru atribucijske teorije motivacije. Atribucijska teorija je socialno-kognitivna teorija motivacije, ki jo je v 80. letih prejšnjega stoletja razvil Bernard Weiner. Po tej teoriji naj bi naša spoznanja (pripisovanje vzroka oz. atribucija) odločilno vplivala na čustva. Ključna predpostavka Weinerjevega modela je, da so naša čustva odvisna od kognitivnega procesa iskanja vzročnih razlag – od tega, kako poskušamo ugotoviti vzrok za neko dejstvo ali dogodek.

Weiner ugotavlja, da svoje dosežke najpogosteje pripisujemo lastni sposobnosti, trudu, težavnosti nalog ali sreči. Atribucije (pripisovanje) delimo glede na tri dimenzije: kontrola, stabilnost in možnost nadzora. Dimenzija kontrole ima dva pola: zunanega in notranjega. Dimenzija stabilnosti pove, ali se razlogi, katerim pripisujemo svoj uspeh ali neuspeh, spreminjajo, ali ostajajo konstantni. Možnost nadzora zajema vrsto vzrokov – od tistih, ki jih posameznik lahko nadzoruje, na primer veščine ali učinkovitost, do vzrokov, na katere nima vpliva, na primer sposobnost, razpoloženje, dejanja drugih ali sreča (Weiner, 1986).

Količina truda, ki ga bo posameznik vložil v neko aktivnost, je odvisna od njegove percepcije oziroma od razlogov, katerim pripisuje svoj uspeh ali neuspeh.

Trud je notranji in nestabilen faktor, na katerega pa posameznik v veliki meri lahko vpliva. Temeljni princip atribucijske teorije, kolikor se navezuje na motivacijo, je prav to, da posameznikova lastna percepcija in razlogi, za katere meni, da vodijo v uspeh ali neuspeh, določajo, koliko truda bo vložil v neko aktivnost (Weiner, 1986).

Motivacija za reševanje preizkusa znanja je posebna oblika motivacije. Odvisna je od konteksta in jo lahko opredelimo kot posameznikovo dojetje lastne motiviranosti za to, da bo pri reševanju preizkusa dal vse od sebe. (Pintrich in Schunk, 2002). V raziskavah s tega področja najdemo različne kazalnike motivacije, kot so izbira nalog, trud, vztrajnost in dosežki. Kljub številnim vprašanjem o vlogi učenčeve motiviranosti pa je pri obsežnih preverjanjih znanja le malo študij raziskalo, kako učenci pristopijo k preizkusu znanja, in kako različni načini reševanja vplivajo na njihove dosežke (Baumert in Demmrich, 2001).

Pri raziskovanju tega vprašanja so se nekateri strokovnjaki odločili preučiti odnos med motivacijo za reševanje preizkusa znanja in različnimi vrstami spodbud. O'Neil, Sugrue, Abedi, Baker in Golan (1997) so v eksperimentalnih študijah o motiviranosti za reševanje preizkusov pokazali, da denarna nagrada, ki jo učenec prejme za vsak pravilen odgovor, poveča trud in izboljša dosežke učencev osmega razreda. Pri učencih dvanajstega razreda denarna spodbuda za pravilne odgovore ni imela vpliva niti na trud niti na dosežke učencev. V neki drugi študiji, izvedeni na objavljenih nalogah raziskave TIMSS, prav tako niso ugotovili nobenega praktičnega učinka na dosežke učencev dvanajstega razreda, ki naj bi za vsak pravilen odgovor dobili denarno nagrado (O'Neil, Abedi, Lee, Myoshi in Mastergeorge, 2004). Kljub temu pa so ti učenci v reševanje preizkusa vložili več truda kakor učenci iz kontrolnih skupin.

Objava rezultatov PISA 2000 je v Nemčiji sprožila široko razpravo v šolskih krogih. Keitel, Kilpatrick in Haenisch so v svojih člankih iz leta 1998 (gl. Baumert in Demmrich, 2001, str. 442) trdili, da učenci v Nemčiji ne jemljejo resno preizkusov znanja, ki se ne ocenjujejo. Baumert in Demmrich sta temu nasprotovala in trdila, da zaključki avtorjev člankov ne temeljijo na empiričnih dokazih. V eksperimentalni raziskavi, ki sta jo izvedla na 467 učencih devetega razreda, sta Baumert in Demmrich ugotovila, da postopek in pogoji preverjanja znanja, kot so povratne informacije, ocene ali finančne spodbude, nimajo nobenega vpliva na načrtovani ali dejanski trud učencev za reševanje preizkusa.

Splošne ugotovitve tako praktičnih kot teoretičnih raziskav o motivaciji učencev za reševanje preizkusov znanja in o njihovih dosežkih so kontradiktorne. Raziskave o tem, kako spodbude vplivajo na trud učencev pri reševanju preizkusov znanja, kažejo različne rezultate. Tudi pri preverjanjih znanja, ki za učence niso odločilnega pomena, je odnos med motivacijo učencev in njihovimi dosežki nejasen. V nekaterih študijah ugotavljajo, da so učenci dokaj motivirani in se pri reševanju potrudijo tudi takrat, ko preizkus zanje nima odločilnih posledic

(Baumert in Demmrich, 2001). Spet druge trdijo, da so pomembnost preizkusa ali posledice, ki sledijo iz preizkusa, za učence res povezana z motivacijo in dosežki (Wolf in Smith, 1995).

Izsledki poskusov o motivaciji za reševanje preizkusov znanja, vzorcih glede na spol in dosežkih učencev so ravno tako kontradiktorni. Študije o vplivu spola niso prinesle enotnih zaključkov. Raziskava, ki sta jo izvedla Karmos in Karmos (1984), je pokazala, da je povezava med motivacijo za reševanje preizkusa in dosežki močnejša pri dečkih kakor pri deklicah. Nasprotno pa Brown in Walberg (1993) nista ugotovila medsebojnega vpliva med motivacijo za reševanje preizkusa znanja in spolom.

Upoštevati je treba mnoge vidike, povezane s tem, kakšen pomen učenci pripisujejo preverjanju znanja. Naše raziskovanje je utemeljeno na prepričanju, da učenci sooblikujejo proces preverjanja znanja in vanj prinašajo svoje znanje in pretekle izkušnje. Cilj tega prispevka je raziskati vpliv, ki ga ima ugotovljeni trud učencev na izide preverjanj. Osredotočili pa se bomo tudi na vzorce razlik, ki se pokažejo pri trudu učencev glede na spol, pri preverjanjih znanja brez odločilnih posledic za učence.

Pregledali bomo podatke iz prvih dveh ciklov raziskave PISA v letih 2000 in 2003, da bi ugotovili povezave med trdom učencev in njihovimi dosežki pri preverjanju znanja. Ugotoviti bomo poskušali, v kolikšni meri na ključne rezultate, na primer na razlike v bralnih dosežkih med državami in med različnimi podskupinami znotraj posameznih držav, vpliva trud, ki ga učenci iz različnih okolij vložijo v reševanje preizkusov znanja.

Merjenje truda učencev

Pregled raziskave PISA, vključno z metodologijo, publikacijami in razlago namena, je na voljo v članku Turnerja in Adamsa (2007). Kljub temu bomo v tem prispevku dokaj natančno opisali instrument, ki je bil v raziskavi uporabljen za kvantifikacijo motivacije za reševanje preizkusa – t.i. termometer truda (*Effort Thermometer*), ki ga je razvila skupina raziskovalcev na Inštitutu Maxa Plancka v Berlinu (Kunter et al., 2002).

Tri sodelujoče države (Nemčija, Avstralija in Norveška), ki niso verjele, da se učenci posebej potrudijo za obsežne študije, kot je PISA, so leta 2000 v preverjanje vključile tudi vprašalnik o tem. Leta 2003 so podobni vprašalniki postali sestavni del PISE in so jih reševali vsi udeleženci.

Navodila za uporabo termometra truda

Termometer truda se uporabi ob koncu dveurnega preizkusa znanja. Učenci dobijo kratka navodila:

*Prosimo, končaj z reševanjem nalog.
Pojdi na zadnjih nekaj strani delovnega zvezka, kjer boš našel vprašanji o uporabi računalna in o trudu, ki si ga vložil v preizkus. Prosimo te, da odgovoriš na vprašanji, nato pa zapreš svoj delovni zvezek.*

Termometer truda obsega tri desetstopenjske lestvice, kakor je prikazano na shemi 1. Te tri lestvice smo poimenovali velik osebni trud, trud za PISO in trud za šolske ocene. Prva lestvica označuje največji trud posameznika v situaciji, ki je zanj izjemno pomembna. Na drugi lestvici lahko izmerimo, koliko se učenci potrudijo za raziskavo PISA. Tretja lestvica kaže pričakovani trud učencev, če bi bilo preverjanje znanja zanje v kontekstu šole zelo pomembno.

Iz sheme 1 lahko razberemo, da s križcem označeno število 10 na lestvici označuje optimalen trud posameznika. Tri lestvice so prikazane s stolpci oštevilčenih kvadratkov. Učenec, ki označi kvadrateg ob številu 10, torej zase meni, da se je pri nečem potrudil, kot je bilo najbolj mogoče.

Uvodnih besedil, ki so del lestvic za ocenjevanje lastnega truda, in razlage termometra truda učitelj učencem ne prebere na glas, zato je mogoče sklepati, da termometer truda učencem, ki ne berejo najbolje, predstavlja precejšen izziv. Uvodno besedilo in navodila vsebujejo štiri kompleksne povedi; skica termometra učencem, ki te besede ne poznajo, ne prinaša natančnih informacij za dekodiranje njegovega pomena; prav tako lestvice niso prikazane vodoravno, kot je običajno pri tovrstnih vprašalnikih, temveč navpično, da s tem sledijo metafori termometra. Učence s težavami pri branju lahko to odvrne od odgovarjanja na vprašanja ali pa je veljavnost njihovih odgovorov vprašljiva.

Shema 1: Termometer truda

Koliko si se potrudil?

Predstavljaj si neko resnično situacijo (na primer v šoli ali v kakšnem drugem kontekstu), ki je zate osebno zelo pomembna, in bi zato dal vse od sebe in se zelo potrudil, da bi bil uspešen.

V takšni situaciji bi na »termometru truda« označil najvišjo vrednost, kot je prikazano spodaj.

Koliko si se potrudil pri reševanju tega preizkusa v primerjavi s situacijo, ki si si jo zamislil?

Koliko bi se potrudil pri reševanju tega preizkusa, če bi se ocene vpisale v redovalnico?

	☒ 10	☐ 10	☐ 10
	☐ 9	☐ 9	☐ 9
	☐ 8	☐ 8	☐ 8
	☐ 7	☐ 7	☐ 7
	☐ 6	☐ 6	☐ 6
	☐ 5	☐ 5	☐ 5
	☐ 4	☐ 4	☐ 4
	☐ 3	☐ 3	☐ 3
	☐ 2	☐ 2	☐ 2
	☐ 1	☐ 1	☐ 1

Uvodna navodila za učence

Preden nadaljujemo, bo koristno pregledati navodila za učence in ugotoviti, ali učence motivirajo in jih spodbujajo k premišljenemu reševanju nalog. Uvodna navodila za učence so kratka in jih nikakor ne moremo imeti za motivacijski govor, kljub temu pa vsebujejo nekaj ključnih informacij, ki se nanašajo na področje in namen preverjanja.

Spodaj navedena navodila so vsi sodelavci za izvedbo zajema podatkov na šolah pred začetkom dveurnega reševanja preizkusa znanja prebrali učencem, ki so sodelovali v raziskavi PISA 2003. Namen navodil je bil učencem podati bistvene splošne informacije o preverjanju znanja v raziskavi PISA.

Izbrani ste bili, da sodelujete v pomembni mednarodni raziskavi, ki se imenuje Mednarodna primerjava dosežkov učencev ali na kratko PISA. Cilj raziskave je ugotoviti, kako dobro učenci vaše starosti po svetu berejo in kaj vedo o matematiki ter naravoslovju. <Državo> predstavlja približno <število> učencev. Po vsem svetu v raziskavi sodeluje okrog 200.000 učencev, ki prihajajo iz več kot 700 šol v 40 državah.

Rezultati raziskave bodo državam pomagali ugotoviti, kaj se učenci učijo. Raziskava lahko v prihodnosti vpliva na učence po vsem svetu, zato vas prosimo, da se potrudite, kolikor je mogoče.

Rezultati

Ta prispevek je v nadaljevanju razdeljen na štiri sklope. V prvem sklopu je opisano, kako so v različnih državah določili primerljivo spremenljivko za trud učencev. Drugi sklop obravnava razmerje med *relativnim trdom* učencev in njihovimi bralnimi dosežki, tretji pa razmerje med *relativnim trdom* in spolom na mednarodni in nacionalni ravni. Četrty sklop se osredotoča na empirično analizo in interpretacijo bralnih dosežkov iz dveh ciklov raziskave PISA pri učencih iz Nemčije in Avstralije.

Spremenljivka truda učencev

Predhodna analiza o trdu učencev za PISO in za šolske ocene je temeljila na opisni statistiki. Sledilo je preučevanje spremenljivk truda učencev in njihovih bralnih dosežkov. Predhodna analiza je bila izvedena na podatkih raziskave PISA 2003.

Določitev spremenljivke truda učencev

Za potrebe kasnejših analiz smo določili novo spremenljivko, in sicer t.i. *razliko v trdu*:

$$\text{razlika v trdu} = \text{trud za PISO} - \text{trud za šolske ocene}$$

Uporaba *razlike v trdu* kot ključne neodvisne spremenljivke namesto zgolj podatkov o trdu učencev za PISO nam pomaga dognati, kako resno učenci jemljejo preizkuse znanja PISA v primerjavi z ostalimi področji šolskega dela, ki imajo zanje pomembne posledice. Z določitvijo spremenljivke *razlika v trdu* lahko tudi bolje izravnamo morebitne kulturne razlike, ki bi se pokazale pri zbranih podatkih o trdu učencev.

Obstajata dva razloga, zakaj je trud učenca bolje prikazati kot diferenčno kakor pa kot absolutno vrednost. Prvič, četudi si s termometrom truda prizadevamo zagotoviti kulturno primerljivost in zato uporabimo lestvico *velik osebni trud* kot sidro, bodo neposredne medkulturne primerjave še vedno vprašljive.

Drugič, povsem mogoče je, da je učinkovitost izobraževalnih sistemov, kjer učenci niso tako motivirani in prizadevni, precej nižja kakor v izobraževalnih sistemih, kjer sta njihova motivacija in prizadevanje za uspeh večja. Če torej v neki državi ugotovimo nižje vrednosti spremenljivke *trud za PISO*, to še ne pomeni, da so razlike v dosežkih učencev med državami *posledica* razlik v trdu učencev – verjetneje gre za vpliv različnih izobraževalnih sistemov ali kultur. Z uporabo spremenljivke, ki jo izrazimo kot razliko med *trdom za PISO* in *trdom za šolske ocene*, torej nevtraliziramo ta vidik razlik v trdu učencev glede na šolski kontekst.

Vrednosti *razlike v trudu* lahko segajo od -9 do +9. Tabela 1 za vsako od teh vrednosti prikazuje (utežene) deleže učencev, ki so sodelovali v raziskavi PISA 2003, skupaj s povprečnimi bralnimi dosežki posameznih skupin učencev.¹ Delež učencev, ki niso odgovarjali na termometer truda, je bil 17,5-odstoten. Mogoče je, da je odstotek takih učencev odvisen od njihovih bralnih zmožnosti in razumevanja koncepta termometra truda, na kar smo že opozorili v prejšnjem poglavju.

Negativna vrednost spremenljivke *razlika v trudu* pomeni, da bi se učenec bolj potrudil pri testu za oceno kakor pri preizkusu raziskave PISA. Takšni učenci realno ocenjujejo trud, ki je potreben za tovrstna preverjanja znanja brez odločilnih posledic.

Iz tabele 1 je razvidno, da je večina učencev (92,8 %), ki so rešili termometer truda, to storila na predvidljiv in realen način. Odstotek učencev s pozitivnimi vrednostmi *razlike v trudu*, ki očitno kaže na nerealno oceno, predstavlja 5,9 % uteženega vzorca oz. 7,2 % veljavnih odgovorov. Rezultat 0 vključuje učence, ki menijo, da so se za PISO potrudili enako, kakor bi se za šolski test za oceno, in predstavlja 22,9 % uteženega vzorca oz. 27,7 % veljavnih odgovorov.

Tabela prav tako kaže, da je povprečni bralni dosežek učencev s pozitivnimi vrednostmi spremenljivke *razlika v trudu* nižji v primerjavi z učenci, ki so bili pri odgovarjanju bolj razumni.

Tabela 1: Deleži učencev po kategorijah razlike v trudu, prikazani z uteženimi podatki raziskave PISA 2003

vrednost razlike v trudu	% učencev	povprečni dosežek	vrednost razlike v trudu	% učencev	povprečni dosežek
-9	0,5	437,34	1	3,2	399,68
-8	0,3	456,14	2	1,4	404,67
-7	0,6	463,38	3	0,6	396,47
-6	1,0	463,14	4	0,3	399,62
-5	2,7	470,19	5	0,2	395,35
-4	3,8	480,60	6	0,1	399,59
-3	8,5	488,71	7	0,0	417,37
-2	16,5	490,84	8	0,0	364,68
-1	19,7	489,83	9	0,1	423,15
0	22,9	464,23	niso odgovarjali	17,53	389,66

Spremenljivka *razlika v trudu* ima dve nezaželeni značilnosti, ki otežujeta analizo in interpretacijo podatkov. Prvič, vrednosti nad 0 imajo zelo nizke frekvence, in drugič, večina učencev ima negativne rezultate. Iz tega razloga je bilo izvedeno dvojno rekodiranje. Pri prvem smo združili vse kategorije s pozitivnimi vrednostmi v eno samo kategorijo z vrednostjo -10. Temeljilo je na

¹ Vsi podatki, ki jih navajamo v tem prispevku, so bili izračunani z uporabo ustreznih uteži (*weights*) in dosežkov v obliki pričakovanih vrednosti za zmožnosti učencev (*plausible values*) skladno s tehničnimi navodili raziskave PISA (OECD, 2005).

opažanju, da učencev, ki so podajali nerealne odgovore, ni bilo veliko in da so imeli nizke bralne dosežke. Pri drugem rekodiranju smo vsaki vrednosti spremenljivke *razlika v trudu* prišteli 10, da bi poenostavili opisovanje in interpretacijo rezultatov analize. Spremenljivko *razlika v trudu* v rekodirani obliki smo poimenovali *relativni trud*.

Tabela 2 za vsako kategorijo *relativnega truda* prikazuje delež učencev in njihov povprečni bralni dosežek, združena kategorija vseh predhodno pozitivnih vrednosti pa je označena kot vrednost 0. Vrednost 10 sedaj postane najpogostejša kategorija in vključuje učence, ki naj bi se tako za preverjanja, ki imajo zanje pomembne posledice, kot za tista, ki niso tako odločilna, enako potrudili.

Za lažje opisovanje in razpravo o trudu učencev je koristno določiti opisnike za pet stopenj *relativnega truda*. Učence v nulti kategoriji označujemo kot *nerealne ocenjevalce*; tiste v prvi kategoriji kot *PISA ciniki*; v osmi kot *PISA realiste*; v deveti kot *prizadevne realiste*; in učence v deseti kategoriji kot *podpornike PISE*.

Podatki v tabeli 2 kažejo, da imajo *nerealni ocenjevalci* (tj. učenci, ki pravijo, da se za PISO potrudijo bolj kot za šolski test za oceno), najnižje povprečje bralnih dosežkov. To sicer ni predmet tega prispevka, toda enako nizki dosežki so za to skupino učencev značilni tudi pri matematični in naravoslovni pismenosti.

Podporniki PISE (ki pri *relativnem trudu* dosegajo vrednost 10) so največja skupina, kamor se uvrščajo učenci, ki trdijo, da so se za raziskavo PISA potrudili enako kakor bi se za test, ki šteje za *oceno* (gl. tabelo 2).

Prizadevni realisti (ki pri *relativnem trudu* dosegajo vrednost 9) so skupina učencev z drugim najvišjim povprečjem bralnih dosežkov v raziskavi PISA.

PISA realisti (za katere ima spremenljivka *relativni trud* vrednost 8) dosegajo najvišje povprečne bralne dosežke v raziskavi PISA.

PISA ciniki (z vrednostjo 1 pri *relativnem trudu*) trdijo, da v raziskavo PISA vložijo zelo malo truda, njihovi povprečni bralni dosežki pa so drugi najslabši.

Tabela 2: Deleži učencev pri rekodirani spremenljivki relativni trud v raziskavi PISA 2003

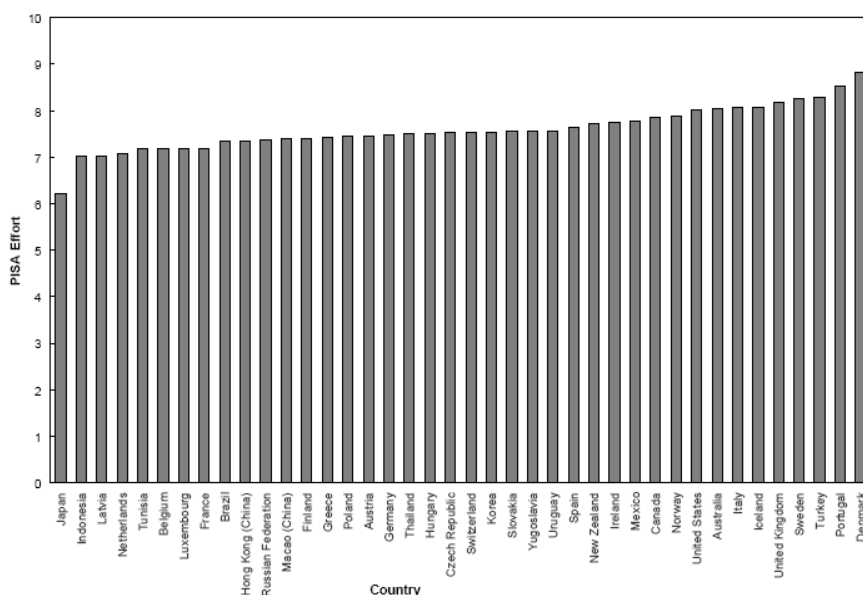
vrednost relativnega truda	% učencev	povprečni dosežek	oznaka glede na <i>relativni trud</i>
0	5,9	401,54	nerealni ocenjevalci PISA ciniki
1	0,5	437,34	
2	0,3	456,14	
3	0,6	463,38	
4	1,0	463,14	
5	2,7	470,19	PISA realisti prizadevni realisti podporniki PISE
6	3,8	480,60	
7	8,5	488,71	
8	16,5	490,84	
9	19,7	489,83	
10	22,9	464,23	
niso odgovarjali	17,5	389,66	
skupaj	100	459,58	

Če strnemo, smo pri preučevanju spremenljivke truda učencev izvedli tri postopke. Prvič, nerealne ocene smo združili v eno samo skupino; drugič, racionalizirali in definirali smo spremenljivko *relativni trud*, in tretjič, določili smo oznake za nekatere skupine učencev na lestvici *relativnega truda*, da bi s tem poenostavili opisovanje in razlago podatkov.

Trud učencev in bralni dosežki

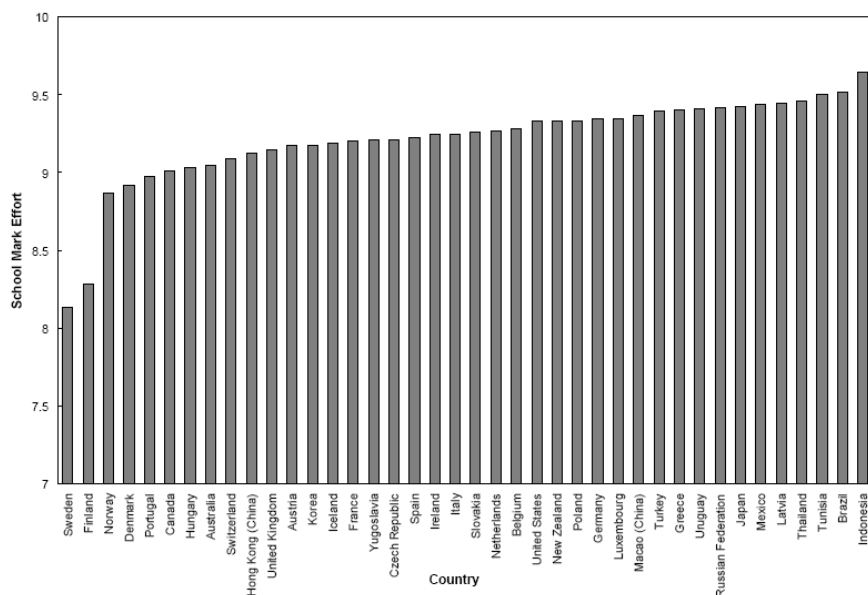
Prvi korak raziskovanja truda učencev na svetovni ravni je preučiti *relativni trud* po državah in vsako spremenljivko truda povezati z bralnimi dosežki posamezne države. Povprečne vrednosti *truda za PISO*, *truda za šolske ocene* in *relativnega truda* za države, ki so sodelovale v raziskavi PISA 2003, so prikazane na shemah 2, 3 in 4.

Shema 2: Trud za PISO po državah



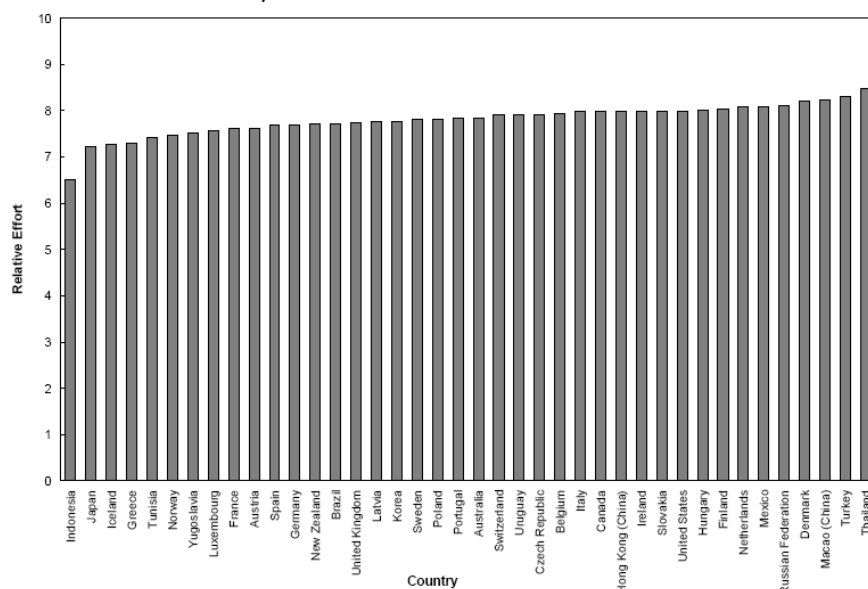
Iz sheme 2 lahko razberemo, da po *trudu za PISO* večina držav dosega vrednosti med 7 in 8. Najnižji rezultat ima Japonska, najvišjega pa Danska. Pet držav, ki se po *trudu za PISO* uvrščajo nad 8, je manj ekonomsko razvitih. Takšen rezultat se ujema z izsledki prejšnjih raziskav in potrjuje, da anketiranci iz ekonomsko slabše razvitih držav pri tovrstnih lestvicah pogosteje dajejo socialno zaželene odgovore (King, Murray, Salomon in Tandon, 2004). Razlika med državo z najvišjo in državo z najnižjo vrednostjo *truda za PISO* je 2,61. Razlike med državami glede na to, kako učenci ocenjujejo svoj trud za PISO, so dokaj majhne, kar ne govori v prid anekdotičnim dokazom, da v nekaterih državah učenci za tovrstne raziskave niso tako motivirani kakor v drugih (Wainer, 1993).

Shema 3: Trud za šolske ocene po državah



Shema 3 kaže, da ima z 8,13 najnižji rezultat glede na trud učencev za šolske ocene Švedska. Najvišji rezultat ima Indonezija z vrednostjo 9,64. Najvišja razlika med dvema državama pri *trudu za šolske ocene* je torej 1,51. Ta razlika je pri *trudu za PISO* višja, kar je morda odraz statusa raziskave PISA in posledica dejstva, da raziskava za učence nima pomembnih posledic.

Shema 4: Relativni trud po državah

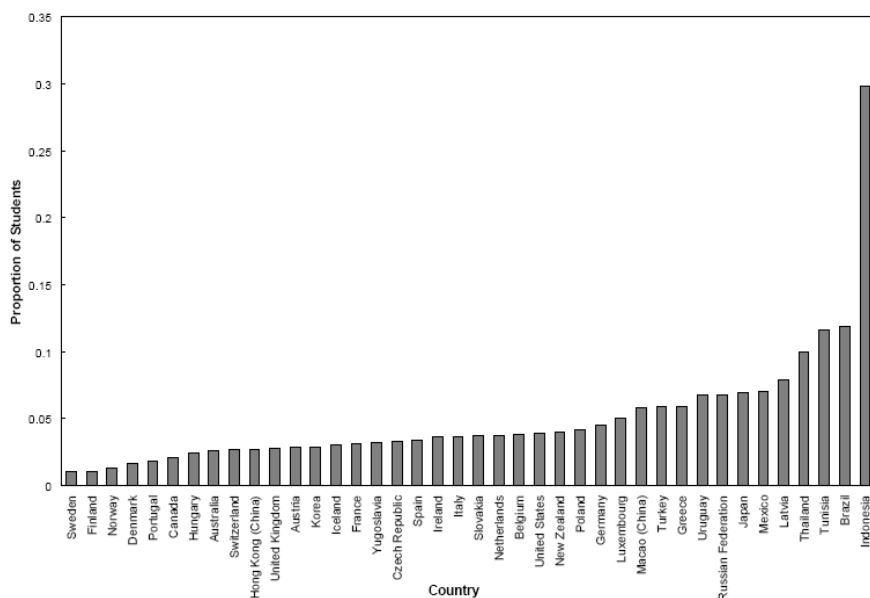


Na shemi 4 ima Indonezija najnižje povprečje *relativnega truda* (6,49) in Tajska najvišjega (8,49). Pri večini držav je povprečna vrednost *relativnega truda* dokaj podobna in znaša okrog 8. To se sklada s ocenami učencev, ki v povprečju svoj trud za PISO na termometru truda označijo za okrog dve »stopinji« nižje od truda za šolske ocene.

Japonska v skupni razvrstitvi povprečnih bralnih dosežkov zaseda 14. mesto (OECD, 2006b), po povprečnem *relativnem trudu* učencev pa se uvršča na 40. mesto. Vendar pa na vprašanje o tem, koliko so se potrudili, japonski učenci odgovarjajo konzervativno tako pri lestvici *truda za PISO* kot pri lestvici *truda za šolske ocene*. Njihovo povprečje znaša 6,22 pri *trudu za PISO* in 8,13 pri *trudu za šolske ocene*. Japonski učenci torej v primerjavi s svojimi vrstniki iz drugih držav dajejo nizke ocene na obeh lestvicah. Tak vzorec pa ne velja za turške učence, ki na obeh lestvicah dosegajo visoke vrednosti. Kot kaže, lahko govorimo o nacionalnih vzorcih konsistentno visokih ali nizkih ocen *relativnega truda* učencev.

Podatki, ki smo jih navajali zgoraj, vključujejo vse učence. Povedali pa smo že, da nekaj učencev svoj trud ocenjuje nerealno – še posebej 7,2 % učencev, ki trdijo, da so se za PISO potrudili bolj, kot bi se za šolski test za oceno.

Shema 5: Delež učencev, ki nerealno ocenjujejo svoj trud, po državah



Na shemi 5 so za vsako državo prikazani deleži učencev, ki so podajali nerealne ocene. Še posebej izstopa Indonezija s kar 29,8 % *nerealnih ocenjevalcev*, kar je veliko več kot v ostalih državah. Pet držav z največjim

deležem *nerealnih ocenjevalcev* – Tajska, Tunizija, Brazilija in Indonezija – pa ima hkrati tudi nizke povprečne bralne dosežke.

Prav tako je na shemi 5 očiten vzorec maloštevilnih *nerealnih ocenjevalcev* v skandinavskih državah. Švedska, Finska, Norveška in Danska zasedajo prvih pet mest med državami, kjer je najmanj učencev, ki bi nerealno ocenjevali svoj trud. V splošnem je za države z nadpovprečnimi bralnimi dosežki glede na povprečje OECD značilno nizko število nerealnih ocen truda učencev; največ takšnih ocen pa je v manj razvitih državah z bralnimi dosežki pod povprečjem OECD. Tak vzorec je morda posledica učenčevih zmožnosti branja in razumevanja različnih uvodnih besedil pri termometru truda.

Navaden primer je Indonezija, kjer je 29,8 % učencev nerealno odgovarjalo pri termometru truda. Obstajata dve hipotezi, ki naj bi pojasnili ta vzorec. Po prvi naj bi bile nerealne ocene posledica socialne zaželenosti odgovorov – učenci podajajo odgovore, ki jih prikažejo v dobri luči. Po drugi razlagi gre za učence, ki slabo berejo in zato niso razumeli lestvic *truda za PISO* in *truda za šolske ocene* glede na lestvico *velikega osebnega truda*, ki služi kot izhodišče.

Korelacije med trdom učencev in bralnimi dosežki

Tabela 3 prikazuje korelacije med povprečnimi dosežki države in povprečnimi vrednostmi truda učencev. Vključene so vse države razen Indonezije. V primeru, ko je neka država z netipičnimi rezultati izključena, je potrebno natančneje preučiti zbrane podatke.

Tabela 3: Korelacije med trdom učencev in njihovimi dosežki za vse države in za vse države razen Indonezije

	vse države	vse države razen Indonezije
<i>relativni trud</i>	0,245	0,020
trud za PISO	-0,506	-0,464
trud za šolske ocene	0,247	0,060
nerealen trud	-0,697	-0,749

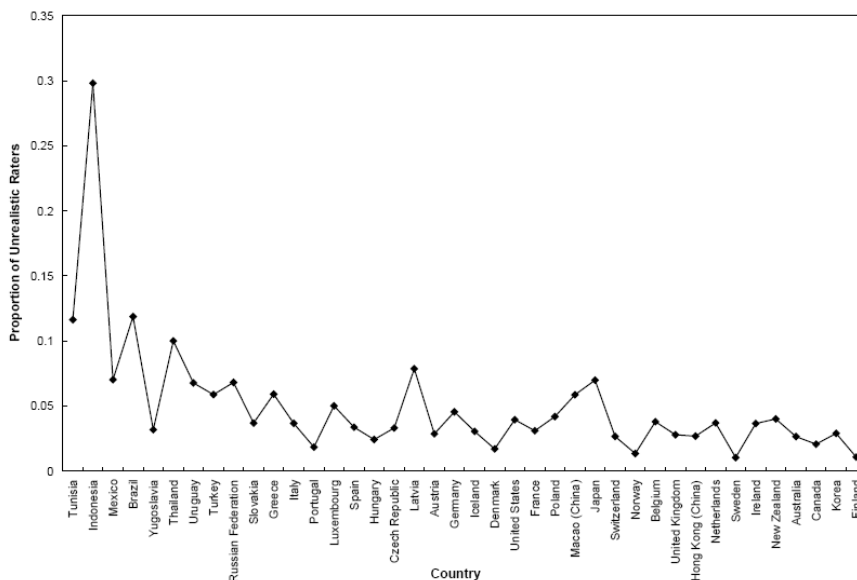
Iz tabele 3 je razvidno, da na ravni države *relativni trud* pozitivno korelira z bralnimi dosežki. Če izključimo Indonezijo, se ta korelacija približa vrednosti 0. Pozitivna korelacija med *relativnim trdom* oziroma *trdom za šolske ocene* in dosežki učencev izgine, če izključimo Indonezijo, medtem ko negativna korelacija med *trdom za PISO* oziroma ocenami *nerealnih ocenjevalcev* in dosežki učencev ostane, tudi če ne upoštevamo Indonezije.

Učenci, ki nerealno ocenjujejo svoj trud, imajo ponavadi nižje bralne zmožnosti, njihov delež pa je navadno večji v državah z nižjimi povprečnimi dosežki.

To priča o tem, da so takšne ocene truda prej posledica slabših bralnih zmognosti učencev kot pa pristranosti pri odgovarjanju, zaradi katere učenci pogostejše dajejo socialno zaželene odgovore. Razmerje med nerealnimi ocenami truda in stopnjo dosežkov po državah prikazuje tudi shema 6.

Na njej so prikazani deleži *nerealnih ocenjevalcev* truda po državah. Glede na bralne dosežke se države razvrščajo od Tunizije z najnižjimi do Finske z najvišjimi dosežki. Vrhovi grafa prikazujejo anomalije v deležu *nerealnih ocenjevalcev*. Tudi tukaj najbolj izstopa Indonezija. Deset najvišje uvrščenih držav glede na dosežke ima majhen delež učencev, ki podajajo nerealne ocene (takih učencev je manj kot 0,05 odstotka). Podobno očiten je tudi skandinavski vzorec majhnega števila *nerealnih ocenjevalcev*. Za Švedsko, Finsko, Norveško in Dansko so značilni visoki dosežki in majhni deleži učencev, ki bi svoj trud nerealno ocenili.

Shema 6: Delež učencev, ki dajejo nerealne ocene, po državah, razvrščenih glede na povprečne bralne dosežke



Vpliv nerealnih ocenjevalcev

Vedenje *nerealnih ocenjevalcev* poskušajo nekateri raziskovalci pojasniti s hipotezama, da učenci dajejo socialno zaželene odgovore in da so nerealni odgovori značilni za učence, ki imajo težave z branjem. Obe hipotezi porajata dvom o zanesljivosti termometra truda kot merskega instrumenta. Prav iz tega razloga bomo v nadaljevanju opravili dvojno analizo podatkov: v prvo bomo vključili *vse učence*, v drugo pa le *realne ocenjevalce*. S takšnim pristopom bomo dobili jasnejšo sliko in lažje izluščili pomembnejša vprašanja.

V skupini *vseh učencev* so vsi tisti, ki so v raziskavi PISA 2003 reševali termometer truda. Skupino *realnih ocenjevalcev* predstavljajo učenci, ki so v raziskavi PISA 2003 termometer truda rešili na realen in razumen način.

Pregled nekaterih prilagoditev modela

Sedaj si bomo približje ogledali, kako lahko *relativni trud* učencev vpliva na najvidnejše rezultate raziskave PISA – tj. v kolikšni meri lahko razlike v trudu učencev vplivajo na (relativno) uvrstitev neke države glede na dosežke.

Tabela 4 prikazuje razmerje med državo in bralnimi dosežki, in sicer posebej za *vse učence* iz vzorca in za *realne ocenjevalce*. To razmerje je izraženo v obliki multiple korelacije in kot R^2 . Pri izračunih smo uporabili podatke na ravni posameznih učencev, opravili pa smo jih s pomočjo regresije bralnih dosežkov glede na to, ali učenec izhaja iz neke države ali ne.

Iz tabele 4 je razvidno, da lahko skoraj 20 odstotkov variance bralnih dosežkov učencev pojasnimo s tem, iz katere države so. Če analizo omejimo na *realne ocenjevalce*, se delež pojasnjene variance zmanjša na 16,6 odstotkov.

Tabela 4: Multipli R in pri regresiji nacionalnega povprečnega dosežka glede na državo, če upoštevamo vse učence ali le realne ocenjevalce

	vsi učenci	realni ocenjevalci
multipli R	0,444	0,407
R^2	0,197	0,166

Tabela 5: Multipli R in R^2 pri regresiji nacionalnega povprečnega dosežka glede na državo in relativni trud, če upoštevamo vse učence ali le realne ocenjevalce

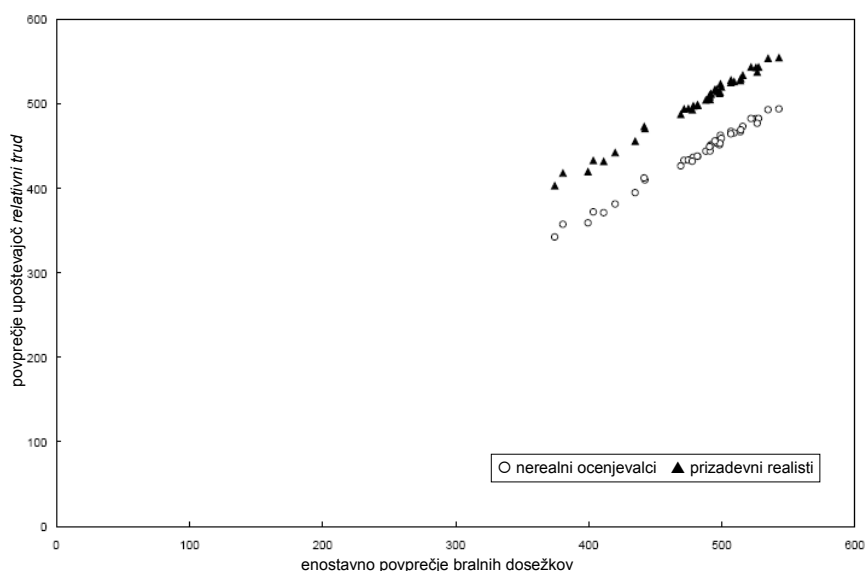
	vsi učenci	realni ocenjevalci
multipli R	0,449	0,420
R^2	0,202	0,177

V naslednjem koraku v model vključimo trud učencev in ugotovimo, kako *relativni trud* vpliva na razlike med državami.

Tabela 5 prikazuje razmerje med državo, *relativnim trdom* in bralnimi dosežki, in sicer posebej za *vse učence* in za skupino *realnih ocenjevalcev*. Izraženo je kot multipla korelacija in kot R^2 . Izračuni so pripravljeni na podlagi podatkov na ravni posameznih učencev s pomočjo regresije bralnih dosežkov glede na to, ali učenec izhaja iz neke države ali ne, in glede na njegov *relativni trud*. Primerjava dobljenih rezultatov s tistimi iz tabele 4 pokaže, da se delež variance dosežkov učencev, ki jo lahko pojasnimo z modelom, poveča za 0,5 odstotka (z 19,7 na 20,2 odstotka).

S pomočjo modelov, ki sta povzeta v tabelah 4 in 5, lahko za posamezno državo ocenimo enostavno nacionalno povprečje bralnih dosežkov in povprečje bralnih dosežkov glede na *relativni trud* učencev. Shema 7 prikazuje razpršenost enostavnih (ali neprilagojenih) povprečnih vrednosti v primerjavi s povprečji, pri katerih je upoštevan *relativni trud* učencev. Krožci predstavljajo predvidena povprečja posameznih držav, če bi se vsi učenci obnašali kot *nerealni ocenjevalci*, trikotniki pa državna povprečja, če bi bili vsi učenci *prizadevni realisti*. Povprečni dosežki se po tem, ko upoštevamo tudi *relativni trud* učencev, pri *nerealnih ocenjevalcih* v povprečju zmanjšajo za 46,76 točke, pri *prizadevnih realisti* pa povečajo za 13,96 točke.

Shema 7: Povprečni dosežki prizadevnih realistov in nerealnih ocenjevalcev



Enostavno (neprilagojeno) povprečje vseh držav znaša 480,93. Predvideno povprečje *prizadevnih realistov* iz vseh držav, če upoštevamo tudi *relativni trud* učencev, je 500,60. To pomeni, da bi bilo predvideno povprečje vseh sodelujočih držav za 19,67 točke višje, če bi se vsi učenci v vseh državah obnašali kot *prizadevni realisti*.

Če analizo omejimo samo na *realne ocenjevalce*, znaša enostavno povprečje 489,70, povprečje *prizadevnih realistov* pa 500,60. Sklepamo lahko, da bi bila predvidena povprečna vrednost bralnih dosežkov za 11,10 točke višja, če bi lahko v vseh državah vse učence, ki so dajali realne ocene, prišteli med *prizadevne realiste*.

Razlike v povprečnih dosežkih se gibljejo od 31,27 v Turčiji do 11,69 na Finskem. Razlika je najočitnejša v Indoneziji – tam znaša kar 37,51 točke. Seveda pa Indonezija še zdaleč ni tipični primer, na kar smo že opozorili.

Pomen in obseg takšnih razlik lahko, na primer, poskusimo osvetliti s kakšnim izmed treh dejavnikov, ki jih ponavadi preučimo v raziskavi PISA. Prvi dejavnik so ekonomske, socialne in kulturne značilnosti, s katerimi lahko pojasnimo petino razlik v dosežkih učencev iz držav OECD (OECD, 2003; Schulz, 2006). Ta delež za trud učencev znaša le 0,5 odstotka – koliko se učenci potrudijo za preverjanje, torej zelo malo vpliva na razlike v dosežkih in (kot je razvidno iz tabele 4) ne *pojasni* razlik med državami, na kar so nekateri strokovnjaki v svojih prispevkih o mednarodnih raziskavah že opozarjali. Drugi dejavnik je razlika med spoloma – na lestvici bralnih dosežkov deklice v povprečju za 26 točk prekašajo dečke. Ta razlika je le malo večja od razlike v dosežkih *vseh učencev* in *prizadevnih realistov*. Tretjič, v državah OECD imajo učenci iz enostarševskih družin v povprečju za 12 točk nižje bralne dosežke kot ostali učenci (OECD, 2003). Ta razlika je enaka kot razlika v dosežkih *realnih ocenjevalcev* in *prizadevnih realistov*.

Zdi se, da razlike v trudu učencev na rezultate ne vplivajo toliko, da bi onemogočale mednarodne primerjave. Po obsegu pa jih lahko enačimo z nekaterimi razlikami (npr. med spoloma), ki jih imamo navadno za dokaj pomembne.

Spol, relativni trud in bralni dosežki

V tem delu prispevka si bomo ogledali razlike med spoloma glede na *relativni trud* učencev in po državah. Raziskali bomo tudi vrzel v bralnih dosežkih deklic in dečkov ter razmerje med bralnimi dosežki in *relativnim trdom* glede na spol. Enako kot v prejšnjih poglavjih bomo *relativni trud* preučili v dveh vzporednih analizah: upoštevajoč *vse učence* in ločeno samo *realne ocenjevalce*.

Pomemben cilj šolskih sistemov je zagotoviti enake izobraževalne možnosti za vse učence, ne glede na spol. Analize izobraževalnih dosežkov so večkrat pokazale, da je ta cilj ogrožen. Na mednarodni ravni vse bolj zbuja skrb bralni dosežki dečkov; prednost deklic je na področju bralne pismenosti precejšnja. Vrzel med spoloma pa je v nekaterih državah izrazitejša kakor v drugih, kar dokazuje, da je razlika med deklicami in dečki z ustreznimi izobraževalnimi pristopi mogoče premagati (OECD, 2001).

Kontekstualna analiza PISA 2000 je pokazala, da deklice berejo pogosteje kot dečki. Dečki večinoma berejo le takrat, ko se to od njih zahteva, medtem ko se deklice branju večkrat posvečajo iz veselja. Razlike so tudi v tem, kakšno čtivo raje izberejo – med dečki so bolj priljubljeni časopisi, stripi, e-pošta in spletne strani, deklice pa raje posegajo po leposlovju (OECD, 2003). Takšni vzorci pogostosti branja in izbire čtiva kažejo, da bi bilo dobro raziskati, ali je motivacija

za reševanje preizkusa odvisna od spola. Če se izkaže, da to drži, je naslednji korak ugotoviti, kako razlike v motivaciji med spoloma vplivajo na dosežke pri preverjanju bralne pismenosti.

Karmos in Karmos (1984) sta motivacijo za reševanje preizkusa preučila s pomočjo lestvic, na katerih so učenci ocenjevali sami sebe. Ugotovila sta, da so deklice v svojih ocenah precej bolj pozitivne kakor dečki, vendar pa so bili rezultati dečkov pogostejše in močnejše povezani z njihovimi dosežki. Do povsem drugačnih ugotovitev pa sta prišla Brown in Walberg (1993), ki v svoji študiji trdita, da je vpliv motivacije pri dečkih in deklicah enak. Če sodimo po nasprotujočih si izsledkih študij, so potrebne dodatne raziskave razlik med spoloma in njihovega vpliva na motivacijo za reševanje preizkusov in bralne dosežke učencev.

Rezultati raziskave PISA 2000 kažejo nedvomno prednost deklic pri bralnih dosežkih. Boljši dosežki deklic so univerzalni, hkrati pa je razlika v primerjavi z dečki ogromna. V povprečju se učenke in učenci razlikujejo za 32 točk, kar je približno polovica točk območja, ki na lestvici PISA predstavlja eno izmed šestih ravni dosežkov. Ponavadi so razlike v dosežkih med spoloma večje od razlik v povprečnih dosežkih med državami (OECD, 2001).

Raziskava PISA 2003 kaže podobno sliko. Tudi tu so bili povprečni dosežki učenk pri bralni pismenosti precej boljši od povprečja učencev; prednost deklic ponovno znaša polovico točk ene ravni dosežkov na lestvici (OECD, 2004a). Če upoštevamo rezultate preverjanja bralne pismenosti za *vse učence*, je korelacija med spolom in bralnimi dosežki 0,136 ter ustreza razliki 29,41 točke, za kolikor se na lestvici bralne pismenosti PISA razlikujejo povprečni dosežki učenk in učencev.

Z vzporedno analizo smo poskušali ugotoviti, ali med deklicami in dečki prihaja do razlik pri spremenljivki *relativni trud* in (če take razlike res obstajajo) ali lahko z njimi pojasnimo razlike v dosežkih med spoloma. V tabeli 6 je prikazano povprečje *relativnega truda* za vse učence in posebej samo za *realne ocenjevalce*. Standardne napake so navedene v oklepajih.

Tabela 6: Povprečni relativni trud (s standardno napako) za deklice in dečke iz vseh držav raziskave PISA 2003

	<i>relativni trud</i> - deklice	<i>relativni trud</i> - dečki	povprečna razlika
vsi učenci	7,919 (0,021)	7,614 (0,027)	0,305 (0,029)
realni ocenjevalci	8,482 (0,014)	8,257 (0,016)	0,225 (0,018)

Iz tabele 6 lahko razberemo, da je povprečni *relativni trud* pri deklicah višji kakor pri dečkih, in sicer ne glede na to, ali upoštevamo samo *realne ocenjevalce* ali *vse učence*. Poleg tega je razlika v povprečnih vrednostih spremenljivke *relativni trud* večja v skupini *vseh učencev* kakor v skupini *realnih ocenjevalcev*. Slednji so k termometru truda pristopili bolj razumno, zato lahko rečemo, da vrednost 0,225 bolje odraža razliko v trudu učenk in učencev. Razlike med

spoloma za vse učence ($t = 10,517$, $p < 0,01$) in za realne ocenjevalce ($t = 12,5$, $p < 0,01$) so sicer statistično pomembne, niso pa velike – znašajo približno četrto točko na termometru truda.

Majhna multipla korelacija obstaja med spolom in *relativnim trdom* ter bralnimi dosežki ($r = 0,260$), vendar pa kljub dokaj nizki vrednosti znaša dvakrat toliko kot korelacija samo med spolom in bralnimi dosežki.

Poskušajmo sedaj ugotoviti, ali statistične prilagoditve, kjer upoštevamo tudi *relativni trud* učencev, privedejo do drugačnih rezultatov glede na povprečne bralne dosežke in spol na nacionalni ravni.

To storimo tako, da preverimo ujemanje podatkov z regresijskim modelom:

$$R_i = a_0^* + a_1^* G_i + b_1 E_i^1 + b_2 E_i^2 + \dots + b_{10} E_i^{10} + e_i, \quad (1)$$

pri čemer je R_i bralna zmožnost učenca i , G_i pa ima vrednost '1', če gre za dečka, in '0', če gre za deklico. Deset spremenljivk $E_i^1, E_i^2, \dots, E_i^{10}$ predstavlja *relativni trud* učenca i . Spremenljivka E_i^j ima vrednost '1', če je *relativni trud* učenca $i = j - 1$; v vseh ostalih primerih ima vrednost '0'. Če je na primer nek učenec *prizadevni realist*, je $E_i^9 = 1$, vsi ostali E_i^j pa imajo vrednost '0'.

V opisanem modelu je a_0^* predvideni povprečni dosežek za dečke, če upoštevamo tudi *relativni trud*, a_1^* pa je razlika med spoloma, upoštevajoč tudi trud učencev. Ocene parametrov tega modela so prikazane v tabeli 7.

Dečke izmed *podpornikov PISE* smo uporabili kot referenčno skupino in ugotovili, da ima skupina *nerealnih ocenjevalcev*, tako dečkov kot deklic, najslabše bralne dosežke. Dosežki dečkov v tej skupini so od dosežkov referenčne skupine v povprečju nižji za 62,39 točke, dosežki deklic pa za 45,99 točke.

Iz tabele 7 lahko razberemo, da ima najvišje dosežke skupina *PISA realistov*, ki imajo za *relativni trud* vrednost 8. Druga najuspešnejša skupina so učenci, ki imajo pri *relativnem trdu* rezultat 7, na tretjem mestu pa so *prizadevni realisti* z rezultatom 9 pri *relativnem trdu*. Zelo slabi so dosežki učencev, ki so nerealno ocenili svoj trud. To znova dokazuje, bodisi da učenci iz te skupine niso bili kos branju navodil za reševanje termometra truda, bodisi da s svojimi odgovori pogosteje želijo ugajati.

Tabela 7: Ocene parametrov modela

regresijski koeficient	ocena	standardna napaka
a_0^*	472,082	1,937
a_1^*	16,399	0,988
b_1	-61,759	2,545
b_2	-22,653	6,594
b_3	-5,284	6,539
b_4	2,009	4,976
b_5	1,558	4,935
b_6	7,704	3,265
b_7	17,444	2,902
b_8	25,441	2,085
b_9	27,091	1,912
b_{10}	25,408	1,967

Spol in *relativni trud*

Sedaj bomo poskušali ugotoviti, ali so vzorci razlik med spoloma glede na *relativni trud* učencev, ki se pojavljajo na mednarodni ravni, značilni tudi za posamezne države. Sodeč po mednarodnih podatkih je povprečni *relativni trud* pri učenkah višji kot pri učencih; to velja tako za skupino, kjer so zajeti *vsi učenci*, kakor za skupino *realnih ocenjevalcev*. Šibko korelacijo lahko opazimo tudi med spolom, *relativnim trdom* in bralnimi dosežki – znaša približno toliko kot korelacija med socialno-ekonomskim ozadjem učencev in njihovimi bralnimi dosežki (OECD, 2003).

Raziskava PISA 2003 je jasno pokazala, da učenke pri branju prekašajo učence. Razlike med državami pa so bile precejšnje – od 58 točk razlike med deklicami in dečki v Islandiji do 21 točk v Južni Koreji, Mehiki in na Nizozemskem ter 13 točk v kitajskem Macau.

Korelacije med spolom in bralnimi dosežki se gibljejo od zanemarljivih (Macao: $r = 0,090$) do občutnih (Islandija: $r = 0,295$, Norveška: $r = 0,229$, Avstrija: $r = 0,229$). Vrzal med dosežki deklic in dečkov pri raziskavi PISA 2003 je največja v Islandiji; nekaj možnih vzrokov za tak razkorak so preučili Olafsson, Halldorsson in Bjornsson (2006).

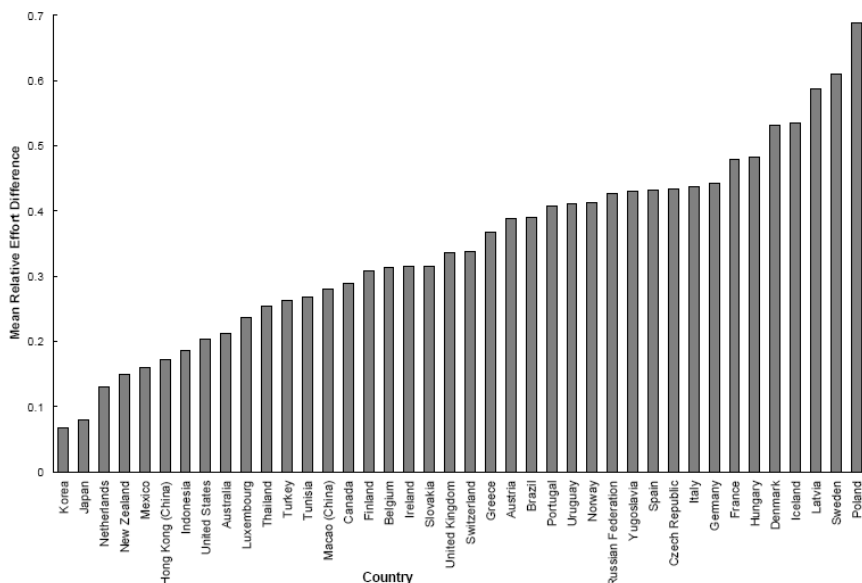
Korelacije med spolom in bralnimi dosežki so skladne z razlikami v bralnih dosežkih, ki segajo od 58,34 točke do 12,19 točke na PISA lestvici. Rečemo lahko, da je razlika v bralnih dosežkih med spoloma velika (več kot 47 točk) v Islandiji, Norveški in Avstraliji.

Če pogledamo razlike v *relativnem trdu* med deklicami in dečki po državah, lahko ugotovimo, ali so višje ocene deklic univerzalne. Shema 8 prikazuje razlike

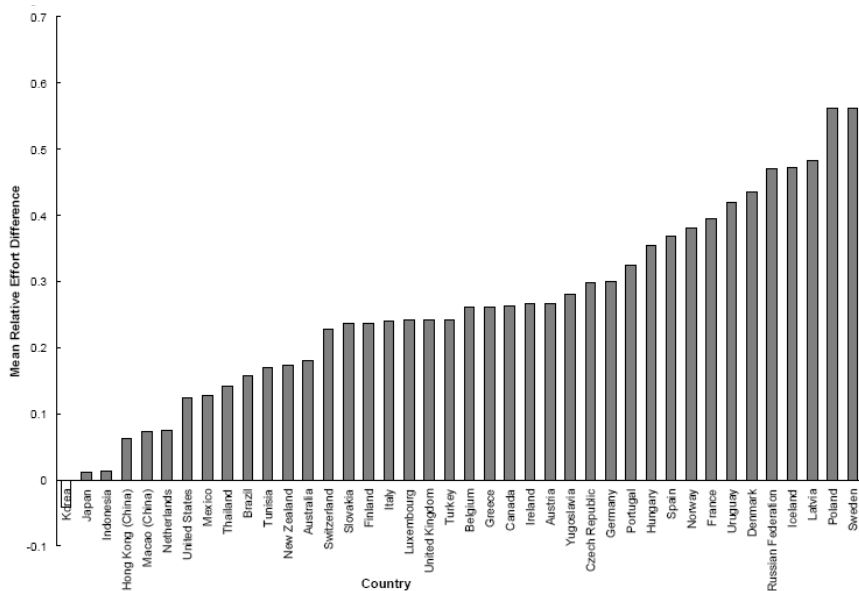
med povprečnim *relativnim trudom* deklic in dečkov za skupino *vseh učencev*, shema 9 pa za *realne ocenjevalce*.

Relativni trud učenk je na shemi 8 višji od *relativnega truda* učencev. Razlika med spoloma je najmanjša v Južni Koreji in na Japonskem, največja pa na Švedskem in Poljskem.

Shema 8: Razlike v relativnem trudu glede na spol za vse učence



Shema 9: Razlike v relativnem trudu glede na spol za realne ocenjevalce



Shema 9 kaže, da je v večini držav *relativni trud* pri deklicah višji. Izjema je Južna Koreja, kjer je *relativni trud* dečkov v primerjavi z deklicami višji. Vendar pa je tudi v tej državi kljub negativni vrednosti razlika med deklicami in dečki majhna. Med državami, kjer ima razlika v *relativnem trudu* deklic in dečkov pozitivno vrednost, je razlika med spoloma najmanjša na Japonskem in v Indoneziji, najizrazitejša pa na Poljskem in Švedskem. Očitno je torej vrstni red držav glede na razlike v *relativnem trudu* deklic in dečkov odvisen tudi od tega, ali vzorec omejimo le na učence, ki so realno ocenili svoj trud. Tako skrčen vzorec razkrije tudi nenavadne rezultate pri dečkih iz Južne Koreje. Po drugi strani pa z vzorcem, v katerega zajamemo le *realne ocenjevalce*, morda ne bomo mogli najbolje prikazati medsebojnega vpliva motivacije in dosežkov.

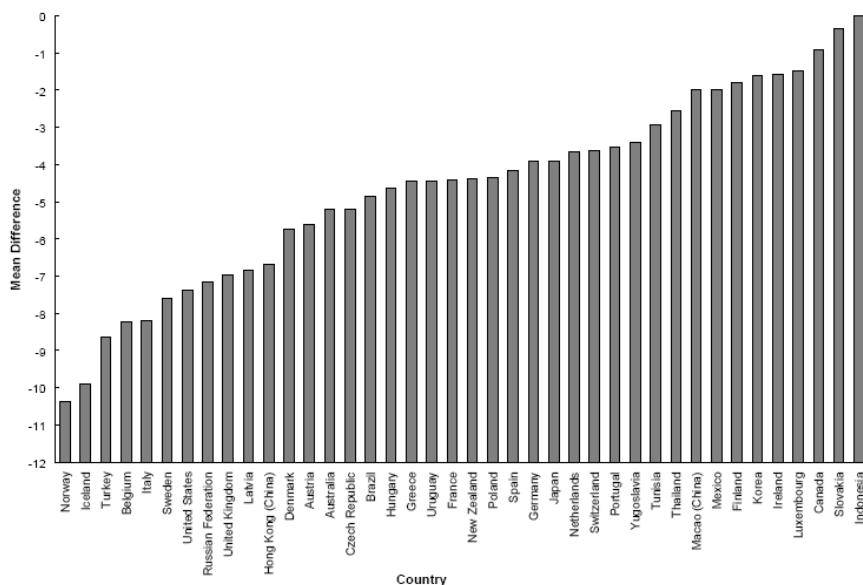
Spol, *relativni trud* in bralni dosežki

Izračunali smo multiple regresije za spol, *relativni trud* in bralne dosežke. V nekaterih državah so korelacije nizke: Japonska ($r = 0,199$), Južna Koreja ($r = 0,192$), Rusija ($r = 0,329$) in Tunizija ($r = 0,169$). Zmerna korelacija se je pokazala v Islandiji ($r = 0,385$), Latviji ($r = 0,329$), na Norveškem ($r = 0,364$) in Švedskem ($r = 0,327$). Če pri izračunih upoštevamo tudi spol in *relativni trud*, na ravni države dobimo jasnejšo sliko o tem, kako *relativni trud* vpliva na vrzel v bralnih dosežkih med deklicami in dečki.

Prilagajanje modela glede na spol in *relativni trud* učencev

V nadaljevanju bomo raziskali, ali je spol in *relativni trud* učencev treba upoštevati pri interpretaciji rezultatov. Pri tem bomo regresijski model (1) uporabili za vsako državo posebej.

Shema 10: Razlika med enostavnimi povprečji in povprečji, kjer je upoštevan spol in *relativni trud* učencev



Shema 10 kaže, da so razlike med obema povprečjema v nekaterih državah zanemarljive, v drugih pa predstavljajo tudi do 10 točk razlike na lestvici bralnih dosežkov. Države, kjer se vrzel v bralnih dosežkih med spoloma ob upoštevanju *relativnega truda* učencev ni bistveno spremenila, so Indonezija, Slovaška in Kanada. Nasprotno pa je na Norveškem, Islandiji, v Turčiji in Belgiji razkorak v bralnih dosežkih deklic in dečkov veliko večji, če pri izračunih upoštevamo tudi *relativni trud* učencev.

V vseh državah se ocena razlik med spoloma izkaže za preveliko, če upoštevamo tudi *relativni trud* učencev. Iz tega lahko sklepamo, da lahko pri raziskavi PISA vsaj del dejanskih razlik med spoloma pojasnimo z razlikami med deklicami in dečki glede na to, koliko se potrudijo pri reševanju nalog. Pri državah, ki se uvrščajo na sredino, lahko s trudom učencev pojasnimo približno 5 točk razlike.

Študija primera Avstralije in Nemčije

Na začetku tega prispevka smo že omenili, da so Avstralija, Nemčija in Norveška v raziskavo PISA 2000 že vključile termometer truda; v ciklu 2003 so ga že reševali učenci iz vseh sodelujočih držav. Kot primer smo izbrali Avstralijo in Nemčijo, saj lahko za ti dve državi primerjamo podatke o trudu učencev iz obeh ciklov raziskave in tako ugotovimo, ali bi bilo razlike v trudu učencev morda treba upoštevati pri interpretaciji trendov v raziskavi PISA.

Drugi razlog, zakaj smo izbrali prav Avstralijo in Nemčijo, je bil svojevrsten odziv na objavljene rezultate raziskave PISA 2000 v teh državah. V Avstraliji so bili rezultati raziskave objavljeni le v nekaterih slabo opaženih medijskih sporočilih in publikacijah (ACER, 2001). Nasprotno pa so v Nemčiji mediji obširno poročali o rezultatih, saj so nemško javnost šokirali podatki, da je njihovo državno povprečje bralnih dosežkov pod povprečjem držav OECD (Deutsche Presse Agentur, 2001; Fertig, 2003). Tudi klima med obema cikloma raziskave je bila v vsaki državi drugačna. Raziskava PISA v Avstraliji ni imela večjega vpliva in je ostala precej neznana, medtem ko se je v Nemčiji o rezultatih raziskave PISA 2000 na široko razpravljalo (Fertig, 2003).

Podatke, ki jih bomo analizirali, smo pridobili iz treh virov. Podatki raziskave PISA 2003 so dostopni na spletni strani PISA in so pripravljeni v skladu z mednarodnimi standardi. Prav tako smo na tej spletni strani dobili podatke o dosežkih učencev v raziskavi PISA 2000. Podatke o trudu avstralskih učencev za raziskavo PISA 2000, ki smo jih združili s podatki o njihovih dosežkih, nam je odstopil nacionalni center raziskave PISA pri Avstralskem inštitutu za raziskovanje v izobraževanju. Podatke o trudu nemških učencev za raziskavo PISA 2000 smo dobili na nacionalnem centru raziskave PISA, ki deluje na Pedagoškem inštitutu za naravoslovje (*Institut für die Pädagogik der Naturwissenschaften – IPN*); tudi te podatke smo nato združili s podatki o dosežkih učencev v Nemčiji.

Tabela 8 prikazuje velikosti neuteženih vzorcev za raziskavi PISA 2000 in PISA 2003 v Avstraliji in Nemčiji. V raziskavi PISA 2000 je sodelovalo 228 avstralskih in 213 nemških šol, v raziskavi PISA 2003 pa 301 avstralska in 211 nemških šol.

Sodeč po tehničnih poročilih raziskave PISA (Adams in Wu, 2002; OECD, 2005) tako avstralski kot nemški podatki izpolnjujejo zahtevne standarde raziskav PISA 2000 in PISA 2003.

Tabela 8: Podatki o vzorcih za raziskavi PISA 2000 in 2003 v Avstraliji in Nemčiji

	Avstralija		Nemčija	
	2000	2003	2000	2003
vsi učenci				
dečki	2806	6335	2438	2315
deklice	2629	6216	2574	2299
neveljaven podatek o spolu	42	0	61	46
skupaj	5477	12551	5073	4660
	Avstralija		Nemčija	
	2000	2003	2000	2003
realni ocenjevalci				
dečki	2517	5687	2112	2001
deklice	2387	5661	2245	2063
neveljaven podatek o spolu	16	0	42	39
skupaj	4920	11348	4399	4103

Trud učencev

Iz porazdelitve učencev glede na njihov *relativni trud* lahko razberemo različne vzorce za Avstralijo in Nemčijo, ki se s časom spreminjajo. Tabela 9 prikazuje deleže učencev za vsako kategorijo *relativnega truda* v Avstraliji in Nemčiji, in sicer ločeno za raziskavi PISA 2000 in PISA 2003.

Tabela 9: Porazdelitev učencev po kategorijah relativnega truda

kategorija relativnega truda	vrednost	2000		2003	
		Avstralija	Nemčija	Avstralija	Nemčija
podporniki PISE	10	19,3	13,8	16,6	18,4
prizadevni realisti	9	25,4	18,7	23,5	22,3
PISA realisti	8	23,0	22,0	24,7	21,6
	7	12,2	13,0	13,4	11,6
	6	5,2	6,6	5,6	5,7
	5	3,1	4,9	3,5	4,4
	4	1,2	1,7	1,6	1,5
	3	0,7	1,2	0,8	1,1
	2	0,3	0,7	0,6	0,4
PISA ciniki	1	0,5	0,6	0,5	0,5
nerealni ocenjevalci	0	3,0	1,9	2,5	4,2
neveljavni odgovori		6,2	14,8	6,8	8,4
razumen trud		79,9	67,5	78,2	73,9

Porazdelitev učencev po kategorijah *relativnega truda* za Avstralijo in Nemčijo v letu 2000 kaže, da se učenci v Nemčiji trudijo manj od svojih vrstnikov v Avstraliji. Primerjava podatkov za Avstralijo v letih 2000 in 2003 pa pokaže, da je trud učencev v treh letih upadel; v letu 2003 je manj *podpornikov PISE* in *prizadevnih realistov* kot v letu 2000. Podatki po letih za Nemčijo pokažejo ravno obraten vzorec – število *podpornikov PISE* in *prizadevnih realistov* je od leta 2000 do leta 2003 naraslo.

S pomočjo kombinirane utežene distribucije lahko ugotovimo, kolikšen delež učencev trdi, da se za preverjanje znanja potrudijo, kolikor je potrebno. Med takšne učence prištevamo *podpornike PISE*, *prizadevne realiste*, *PISA realiste* in prvo naslednjo skupino *realistov* (njihovi deleži so v tabeli 9 zapisani ležeče). Na splošno v obeh državah in obeh ciklih raziskave 74,9 odstotka učencev meni, da se preverjanja lotijo skrbno in se zanj dovolj potrudijo. Hkrati nam ta odstotek pove, koliko učencev se zelo potruzi tudi za preverjanje znanja, ki zanje nima tako pomembnih posledic.

Nasprotno pa je skupina *PISA cinikov* majhna in dokaj stabilna v obeh državah ter v obeh ciklih raziskave. Najbrž gre za skupino ključovalnih učencev, ki trdijo, da se za PISO zelo potrudijo, za šolski test pa ne, iz česar lahko sklepamo, da ti učenci raziskave PISA ne jemljejo resno. Prav tako zanimiva je skupina *nerealnih ocenjevalcev* v Nemčiji leta 2003, ki se je povečala glede na leto 2000, kar bi morda lahko pripisali dojemanju raziskave PISA kot pokazatelja nacionalnega uspeha.

Tabela 10 prikazuje povprečni *relativni trud* za vse učence in za *realne ocenjevalce* v Avstraliji in Nemčiji. Takoj lahko opazimo, da je povprečni *relativni trud* tako v letu 2000 kot v letu 2003 za Avstralijo večji kot za Nemčijo. Za leto 2000 razlika med državama znaša 0,310 ($t = 5,08$, $p < 0,01$), za leto 2003 pa 0,137 ($t = 2,69$, $p < 0,01$); razlike so statistično pomembne.

Tabela 10: Povprečni relativni trud (s standardno napako) po državah za vse učence in za realne ocenjevalce

	država	vsi učenci	realni ocenjevalci
2000	Avstralija	7,948 (0,048)	8,208 (0,037)
	Nemčija	7,638 (0,038)	7,812 (0,033)
2003	Avstralija	7,836 (0,025)	8,049 (0,025)
	Nemčija	7,699 (0,044)	8,066 (0,034)

Razlika med državama v letu 2003 je sicer še vedno statistično pomembna, vendar pa se je glede na leto 2000 zmanjšala. Razlog za to je statistično pomemben upad truda avstralskih učencev (0,112, $t = 2,08$, $p < 0,05$) in majhno, statistično nepomembno povečanje truda nemških učencev (0,061, $t = 1,04$, $p > 0,01$).

Naslednji korak primerjave je ugotoviti, ali se vzorci, značilni za skupino *vseh učencev*, na enak način pojavljajo tudi v skupini *realnih ocenjevalcev*. V raziskavi PISA 2000 je bil povprečni *relativni trud realnih ocenjevalcev* v Nemčiji nižji kakor v Avstraliji. Razlika 0,396 ($t = 7,987$, $p < 0,01$) je statistično pomembna.

V letu 2003 se je avstralsko povprečje nekoliko znižalo (0,159, $t = 4,29$, $p < 0,01$), torej lahko rečemo, da so se učenci v tem letu nekoliko manj potrudili, medtem ko je povprečje v Nemčiji naraslo (0,254, $t = 7,69$, $p < 0,01$). To pomeni, da so se učenci v letu 2003 bolj trudili v Nemčiji kakor v Avstraliji; vendar pa razlika med državama ni statistično pomembna (0,017, $t = 0,403$, $p > 0,01$).

Srednje vrednosti so pri *realnih ocenjevalcih* višje kakor pri skupini *vseh učencev*, *realni ocenjevalci* so se torej v Avstraliji in Nemčiji v letih 2000 in 2003 bolj potrudili za raziskavo PISA. Trud učencev iz Nemčije je od leta 2000 do leta 2003 narasel, podatki pa kažejo, da je bilo povečanje izrazitejše v skupini *realnih ocenjevalcev* kakor za *vse učence*. Sklepamo lahko, da so se učenci v Nemčiji zavedali polemik v zvezi z rezultati raziskave PISA 2000 in so se zato prizadevneje lotili preverjanja v letu 2003.

V tabeli 11 so prikazane srednje vrednosti *relativnega truda* učencev, ločeno po spolu in ciklih raziskave PISA. Podatki kažejo, da se tako v skupini *vseh učencev* kakor v tisti, ki vključuje le *realne ocenjevalce*, deklice trudijo bolj kakor dečki. Za skupino *vseh učencev* v Avstraliji razlika med spoloma znaša 0,204 za leto 2000 in 0,212 za leto 2003 ($t = 4,285$, $p < 0,01$), v Nemčiji pa 0,480 ($t = 7,741$, $p < 0,01$) za leto 2000 in 0,443 za leto 2003 ($t = 6,515$, $p < 0,01$). Razlike med spoloma so torej za oba cikla raziskave v obeh državah statistično pomembne, če upoštevamo *vse učence*.

Za skupino *realnih ocenjevalcev* znašajo razlike med spoloma v Avstraliji po posameznih ciklih raziskave 0,227 ($t = 3,771$, $p < 0,01$) in 0,181 ($t = 4,568$, $p < 0,01$), v Nemčiji pa 0,428 ($t = 7,824$, $p < 0,01$) in 0,299 ($t = 5,067$, $p < 0,01$). Tudi tukaj so vse razlike statistično pomembne. V primerjavi s skupino *vseh učencev* je razlika v povprečnem *relativnem trudu* deklic in dečkov v Nemčiji manjša v podskupini *realnih ocenjevalcev*. Od leta 2000 do leta 2003 se je ta razlika zmanjšala in se približala razliki med deklicami in dečki v Avstraliji. Zaključimo lahko, da je v skupini *realnih ocenjevalcev* v Nemčiji razlika med spoloma leta 2003 skorajda primerljiva z razliko, ki jo glede na trud deklic in dečkov opazimo v isti skupini avstralskih učencev.

Tabela 11: Povprečni relativni trud (s standardno napako) po spolu za vse učence in za realne ocenjevalce

	država	spol	vsi učenci	realni ocenjevalci
2000	Avstralija	deklice	8,059 (0,050)	8,332 (0,036)
		dečki	7,855 (0,066)	8,105 (0,055)
	Nemčija	deklice	7,881 (0,038)	8,025 (0,032)
		dečki	7,401 (0,059)	7,597 (0,051)
2003	Avstralija	deklice	7,943 (0,031)	8,140 (0,025)
		dečki	7,731 (0,038)	7,959 (0,038)
	Nemčija	deklice	7,929 (0,049)	8,220 (0,035)
		dečki	7,486 (0,063)	7,921 (0,054)

Bralni dosežki

V tem delu prispevka bomo raziskali vzorce, ki so se v dveh ciklih raziskave PISA na področju bralnih dosežkov pokazali v Nemčiji in Avstraliji pri deklicah in dečkih. Srednje vrednosti bralnih dosežkov so prikazane v tabeli 12; razvrščene so po ciklih raziskave, državah in spolu.

Iz tabele 12 je za Avstralijo razviden neznatni upad dosežkov (za 3,407 točke) od leta 2000 do leta 2003 ($t = 0,850$, $p < 0,01$), medtem ko so se dosežki učencev v istem obdobju v Nemčiji izboljšali za 7,367 točke ($t = 1,759$, $p < 0,01$) – vendar pa razlika tudi v Nemčiji ni tolikšna, da bi bila statistično pomembna. Podatki torej za obe državi kažejo neznatne spremembe v bralnih dosežkih učencev, do katerih je prišlo v triletnem časovnem obdobju.

Avstralski učenci so bili obakrat uspešnejši od svojih nemških vrstnikov, toda razlika med državama je leta 2003 precej manjša. Vrzel med dosežki obeh držav se je s 44,843 točke razlike v letu 2000 zmanjšala na 34,069 točke v letu 2003, pri čemer je razlika iz leta 2000 statistično pomembna ($t = 10,684$, $p < 0,01$). Razkorak v dosežkih obeh držav je torej precejšen.

V vseh štirih primerih (torej v obeh državah in v obeh ciklih raziskave) so bile deklice uspešnejše od dečkov. V raziskavi PISA 2000 je bila vrzel v dosežkih med spoloma v obeh državah podobna. V Avstraliji je znašala 31,574 ($t = 5,206$, $p < 0,01$), v Nemčiji pa 34,646 ($t = 6,929$, $p < 0,01$). Do leta 2003 pa se je razlika med dosežki deklic in dečkov v Nemčiji povečala in dosegla 42,123 točke ($t = 7,321$, $p < 0,01$) v primerjavi z 39,339 točke ($t = 10,312$, $p < 0,01$), kolikor je takrat znašala v Avstraliji. Za obe državi in v obeh ciklih raziskave je vrzel v dosežkih med spoloma statistično pomembna.

V Avstraliji so dosežki deklic v enem in drugem ciklu raziskave statistično primerljivi; razlika znaša 1,659 točke ($t = 0,318$, $p > 0,01$). Dosežki učenec so se v Nemčiji leta 2003 nekoliko povečali; razlika glede na leto 2000 znaša 10,729

točke, kar je tik pod mejo statistične pomembnosti, določene z ravno 0,05 ($t = 1,950$, $p > 0,01$). Pri dečkih niti v Avstraliji niti v Nemčiji ni opaziti statistično pomembnih sprememb. Dosežki dečkov so v Avstraliji leta 2003 za 7,132 točke nižji kakor tri leta pred tem ($t = 1,425$, $p > 0,01$), v Nemčiji pa za 3,252 točke višji glede na leto 2000 ($t = 0,616$, $p > 0,01$).

Če strnemo zgornje ugotovitve, lahko rečemo, da smo ugotovili neznatne spremembe v bralnih dosežkih obeh držav skozi čas in zanemarljive poraste in padce dosežkov pri deklicah in dečkih. Odkrili smo precejšnje vrzeli med dosežki obeh držav, pomembne razlike pa so se od prvega do drugega cikla raziskave PISA pokazale tudi med dosežki deklic in dečkov v posamezni državi. V nadaljevanju bomo skušali ugotoviti, ali so ugotovljeni vzorci enaki tako za skupino *vseh učencev* kakor za skupino, ki jo sestavljajo le *realni ocenjevalci*.

Iz tabele 13 lahko razberemo, da od prvega do drugega cikla raziskave dosežki v Avstraliji nekoliko upadejo (tj. za 3,427 točke; $t = 0,877$, $p > 0,01$). V Nemčiji v istem časovnem obdobju beležimo rahel porast dosežkov, vendar je razlika 0,273 točke tudi v tem primeru premajhna, da bi bila statistično pomembna ($t = 0,068$, $p > 0,1$).

Vrzel med dosežki se od leta 2000 do leta 2003 zoži s 35,785 točke na 32,085 točke. Vrzel med dosežki obeh držav v letu 2000 je statistično pomembna ($t = 8,602$, $p < 0,01$). Četudi se razlika med državama v letu 2003 zmanjša, ostaja nezanemarljiva ($t = 8,620$, $p < 0,01$). To pomeni, da so dosežki *realnih ocenjevalcev* v Avstraliji v obeh letih še vedno boljši kakor v Nemčiji. Razlika v dosežkih med državama je statistično pomembna in je manjša za skupino *realnih ocenjevalcev* kot pa za skupino, v katero prištevamo *vse učence*.

Deklice imajo v obeh državah in v obeh ciklih raziskave boljše dosežke od dečkov. Leta 2000 je razlika med spoloma v Avstraliji 31,574 točke ($t = 5,228$, $p < 0,01$), v Nemčiji pa 27,616 točke ($t = 6,502$, $p < 0,01$). Do leta 2003 razlika med dosežki deklic in dečkov v Nemčiji naraste na 37,558 točke ($t = 6,837$, $p < 0,01$), medtem ko v Avstraliji to leto znaša 36,607 točke ($t = 10,085$, $p < 0,01$). Vrzeli v dosežkih med spoloma v obeh državah so za oba cikla raziskave statistično pomembne.

Tabela 12: Povprečni bralni dosežki (s standardno napako) za vse učence

	vsi učenci	deklice	dečki	skupaj
2000	Avstralija	547,084 (4,549)	513,218 (4,012)	528,834 (3,397)
	Nemčija	502,198 (3,875)	467,552 (3,169)	483,991 (2,465)
2003	Avstralija	545,425 (2,553)	506,086 (2,835)	525,427 (2,126)
	Nemčija	512,927 (3,905)	470,804 (4,226)	491,358 (3,386)

Tabela 13: Povprečni bralni dosežki (s standardno napako) za učence, ki dajejo realne ocene

	realni ocenjevalci	deklice	dečki	skupaj
2000	Avstralija	553,337 (4,569)	521,763 (3,949)	536,714 (3,431)
	Nemčija	515,104 (2,915)	487,489 (3,089)	500,929 (2,353)
2003	Avstralija	551,781 (2,420)	515,174 (2,706)	533,287 (1,866)
	Nemčija	520,186 (3,767)	482,628 (3,998)	501,202 (3,221)

Dosežki avstralskih učenk so od enega do drugega cikla raziskave PISA statistično primerljivi; razlika znaša 1,556 točke ($t = 0,301$, $p > 0,01$). Pri učenkah v Nemčiji je opaziti manjše zvišanje dosežkov v višini 5,082 točke v primerjavi s skupino vseh učencev, ne glede na spol, vendar pa porast ni statistično pomembna ($t = 1,067$, $p > 0,01$). Pri dečkih niti rezultati v Avstraliji niti v Nemčiji niso statistično pomembni. Njihovi dosežki so se v Avstraliji od leta 2000 do leta 2003 znižali za 6,589 točke ($t = 1,376$, $p > 0,01$), v Nemčiji pa za 4,861 točke ($t = 0,962$, $p > 0,01$).

Razen pri rezultatih dečkov v Nemčiji smo v obeh ciklih raziskave ugotovili enake vzorce dosežkov pri skupini *realnih ocenjevalcev* in skupini *vseh učencev*. Na podlagi tega lahko sklepamo, da vzorci, ki smo jih razkrili pri skupini *vseh učencev*, veljajo tudi za *realne ocenjevalce*. Spremembe v bralnih dosežkih posamezne države, ki jih opazimo od enega do drugega cikla raziskave, so nepomembne. To pa ne drži za razlike, ki se od enega do drugega cikla raziskave pojavijo med obema državama in med spoloma – teh razlik ne moremo zanemariti. Nihanja dosežkov deklic ali dečkov iz skupine *realnih ocenjevalcev* iz leta v leto prav tako niso pomembna.

Prikazali smo torej razmerje med bralnimi dosežki in *relativnim trdom* učencev. Prav tako smo pokazali, da obstajajo razlike v *relativnem trdu* med skupino *vseh učencev* in skupino *realnih ocenjevalcev*, med dečki in deklicami ter med leti. Pokazale so se tudi razlike v bralnih dosežkih med skupino *vseh učencev* in skupino *realnih ocenjevalcev*, med Avstralijo in Nemčijo, med učenci in učenkami iz obeh držav ter med rezultati raziskav PISA 2000 in PISA 2003.

Prilagajanje modela glede na *relativni trud* učencev

V tem delu prispevka bomo raziskali, kaj se zgodi z dosežki nemških in avstralskih učencev, če v statističnem modelu upoštevamo tudi njihov *relativni trud*. Pri tem bomo preverili, kako se podatki prečnega vzorca Nemčije in Avstralije iz let 2000 in 2003 prilegajo dvema alternativnima regresijskima modeloma. Prvi regresijski model je:

$$\begin{aligned}
 R_i = & a_0 + a_1 C_i + a_2 Y_i + a_3 G_i \\
 & + a_4 (C_i \cdot Y_i) + a_5 (C_i \cdot G_i) \\
 & + a_6 (G_i \cdot Y_i) + a_7 (C_i \cdot Y_i \cdot G_i) \\
 & + e_i,
 \end{aligned}
 \tag{2}$$

pri čemer je R_i bralna zmožnost učenca i , C_i ima vrednost '1', če je učenec i iz Avstralije, v nasprotnem primeru pa ima vrednost '0'; G_i je '1', če gre za deklico, in '0', če gre za dečka; Y_i pa ima vrednost '1', če gre za učenca iz leta 2003, in vrednost '0' v ostalih primerih.

Parametre tega modela lahko razložimo takole:

- a_0 je ocenjeni povprečni dosežek za nemške dečke v letu 2000;
- a_1 je skupna razlika v dosežkih za državo;
- a_2 je skupna razlika v dosežkih med obema cikloma raziskave PISA;
- a_3 je skupna razlika v dosežkih med spoloma;
- a_4 je medsebojni vpliv države in cikla raziskave PISA;
- a_5 je medsebojni vpliv države in spola;
- a_6 je medsebojni vpliv spola in cikla raziskave PISA;
- a_7 je trosmerni medsebojni vpliv države, cikla raziskave PISA in spola.

Model (2) imenujmo *enostavni* regresijski model. Ocene parametrov tega modela so navedene v tabeli 14.

Tabela 14: Ocene parametrov enostavnega modela

regresijski koeficient	ocena	standardna napaka
a_0	466,437	3,145
a_1	46,177	5,408
a_2	4,076	5,535
a_3	35,761	5,151
a_4	-10,603	7,329
a_5	-1,290	7,879
a_6	6,653	7,400
a_7	-1,785	10,062

Drugi regresijski model je prikazan z enačbo (3). Pri tem modelu vključimo še 10 spremenljivk, $E_i^1, E_i^2, \dots, E_i^{10}$, ki predstavljajo *relativni trud* učenca i .

Spremenljivka E_i^j ima vrednost '1', če je *relativni trud* učenca $i = j - 1$; v vseh ostalih primerih ima vrednost '0'. Če je na primer nek učenec *prizadevni realist*, je $E_i^9 = 1$, vsi ostali E_i^j pa imajo vrednost '0':

$$\begin{aligned} R_i = & a_0^* + a_1^*C_i + a_2^*Y_i + a_3^*G_i \\ & + a_4(C_i \cdot Y_i) + a_5(C_i \cdot G_i) \\ & + a_6(G_i \cdot Y_i) + a_7(C_i \cdot Y_i \cdot G_i) \\ & + e_i \end{aligned} \quad (3)$$

V tem modelu ima vsak parameter alfa enak pomen kakor v modelu (2), le da so parametri tukaj prilagojeni glede na spremenljivke *relativnega truda* učencev. Ocenam na podlagi tega modela bi zato lahko rekli *prilagojene* ocene.

Kodiranje spremenljivke *relativni trud* glede na spremenljivke E_i^j poteka tako, da vzamemo *podpornike PISE* za referenčno kategorijo; učenec, ki sodi v to skupino, bo imel $E_i^j = 0$ za vse vrednosti j .

Ocene parametrov prilagojenega regresijskega modela so prikazane v tabeli 15.

Tabela 15: Ocene parametrov prilagojenega modela

regresijski koeficient	ocena	standardna napaka	regresijski koeficient	ocena	standardna napaka
a_0^*	492,017	3,712	b_1	-72,742	6,136
a_1^*	30,335	4,821	b_2	-70,341	11,832
a_2^*	-6,223	5,059	b_3	-55,802	11,074
a_3^*	24,160	3,850	b_4	-43,271	10,175
a_4^*	0,388	6,564	b_5	-42,814	7,868
a_5^*	4,469	6,666	b_6	-27,506	5,887
a_6^*	10,245	6,167	b_7	-17,930	4,294
a_7^*	-4,287	8,844	b_8	-7,908	3,579
			b_9	2,136	2,938
			b_{10}	13,846	3,608

Tabela 16: Cenilke za povprečja podskupin po enostavnem in prilagojenem regresijskem modelu

enostavni regresijski model				
	Avstralija		Nemčija	
	2000	2003	2000	2003
dečki	$a_0 + a_1$	$a_0 + a_1 + a_2 + a_4$	a_0	$a_0 + a_2$
deklice	$a_0 + a_1 + a_3 + a_5$	$a_0 + a_1 + a_2 + a_3 + a_4 + a_5 + a_6 + a_7$	$a_0 + a_3$	$a_0 + a_2 + a_3 + a_6$
prilagojeni regresijski model				
	Avstralija		Nemčija	
	2000	2003	2000	2003
dečki	$a_0^* + a_1^* + b_{10}$	$a_0^* + a_1^* + a_2^* + a_4^* + b_{10}$	$a_0^* + b_{10}$	$a_0^* + a_2^* + b_{10}$
deklice	$a_0^* + a_1^* + a_3^* + a_5^* + b_{10}$	$a_0^* + a_1^* + a_2^* + a_3^* + a_4^* + a_5^* + a_6^* + a_7^* + b_{10}$	$a_0^* + a_3^* + b_{10}$	$a_0^* + a_2^* + a_3^* + a_6^* + b_{10}$

S primerjavo rezultatov, ki smo jih dobili ob preizkušanju obeh opisanih regresijskih modelov, lahko poskusimo poiskati odgovor na ključno vprašanje: ali trud učencev šteje? Po tem, ko smo model prilagodili *relativnemu trudu* učencev, lahko raziščemo, kako trud učencev vpliva na njihove bralne dosežke.

V tabeli 16 vidimo, kako lahko ocene parametrov iz obeh regresijskih modelov uporabimo pri oblikovanju ocen povprečnih dosežkov različnih podskupin učencev, in sicer brez in z upoštevanjem njihovega *relativnega truda*. Cenilke, pri katerih so vrednosti prilagojene *relativnemu trudu* učencev, vključujejo tudi parameter b_{10} ; s tem vse nemške in avstralske učence obravnavamo, kakor bi bili *prizadevni realisti*.

Naše ugotovitve bomo najprej predstavili z opisom trendov dosežkov glede na spol, nato pa še z opisom trendov, značilnih za Avstralijo oziroma Nemčijo. Tudi tokrat bomo podatke predstavili ločeno za *vse učence* in za podskupino *realnih ocenjevalcev*.

Trendi dosežkov glede na spol

Najprej si bomo ogledali dosežke dečkov in ugotovili, ali se vzorci trendov pri skupini *vseh učencev* in pri *realnih ocenjevalcih* razlikujejo. Rezultati učencev iz Avstralije in Nemčije v letih 2000 in 2003 so prikazani v tabeli 17.

Tabela 17: Enostavna povprečja bralnih dosežkov in povprečja, prilagojena glede na relativni trud , za vse dečke in za dečke iz skupine realnih ocenjevalcev

vsi učenci		2000	2003	realni ocenjevalci		2000	2003
Avstralija	EP	512,614	506,087	Avstralija	EP	521,778	515,174
	PP	536,198	530,363		PP	536,956	531,138
Nemčija	EP	466,437	470,513	Nemčija	EP	486,600	482,375
	PP	505,863	499,640		PP	505,473	499,215

EP – enostavna povprečja bralnih dosežkov

PP – povprečja, prilagojena glede na *relativni trud*

Dosežki dečkov v Avstraliji so se od leta 2000 do leta 2003 poslabšali. Tak vzorec lahko opazimo tako pri enostavnih podatkih kakor pri podatkih, prilagojenih glede na *relativni trud* učencev, in je značilen tako za *vse učence* kakor za skupino *realnih ocenjevalcev*.

Enostavni povprečni dosežki dečkov v Nemčiji so v letu 2003 višji kakor leta 2000. Če pa upoštevamo tudi njihov *relativni trud*, se izkaže, da so se dosežki v treh letih pravzaprav nekoliko poslabšali. To pomeni, da lahko vsako očitno zvišanje dosežkov pri dečkih iz Nemčije pojasnimo z dejstvom, da so se pri preverjanju bolj potrudili.

Če pogledamo le skupino *realnih ocenjevalcev*, vidimo, da so se njihovi bralni dosežki od leta 2000 do leta 2003 poslabšali za 4,23 točke. Če upoštevamo tudi *relativni trud*, je padec dosežkov še očitnejši in sedaj znaša 6,26 točke. Prilagojeni podatki torej razkrijejo večji upad bralnih dosežkov kakor enostavni podatki, ki ne upoštevajo *relativnega truda* učencev.

Pri dečkih v Nemčiji se torej bralni dosežki niso izboljšali, ko smo upoštevali tudi trud, ki so ga vložili v preverjanje. To kaže, da izboljšanja ne moremo pripisati večji bralni pismenosti, ampak boljšemu, bolj pozitivnemu pristopu učencev k reševanju preizkusa. Povsem mogoče je, da pri interpretaciji rezultatov raziskave PISA 2003 v Nemčiji niso upoštevali večje motiviranosti učencev in so boljše rezultate razlagali kot posledico višjih bralnih sposobnosti učencev. Po drugi strani bi bilo izboljšanje v tem ciklu raziskave PISA prav tako mogoče pripisati prenizkim rezultatom pri dečkih iz leta 2000. Glede na pričakovanja, s kakršnimi so v Nemčiji sprejeli rezultate raziskave PISA 2003, bi bilo očitno izboljšanje bralnih dosežkov pri dečkih sicer dobrodošla, a kljub temu zmotna razlaga.

Rezultati za deklice iz Avstralije v Nemčije v letih 2000 in 2003 so prikazani v tabeli 18.

Tabela 18: Enostavna povprečja bralnih dosežkov in povprečja, prilagojena glede na relativni trud, za vse deklice in za deklice iz skupine realnih ocenjevalcev

vsi učenci		2000	2003	realni ocenjevalci		2000	2003
Avstralija	EP	547,085	545,426	Avstralija	EP	553,337	551,781
	PP	564,827	564,950		PP	565,710	565,483
Nemčija	EP	502,198	512,927	Nemčija	EP	515,104	520,186
	PP	530,023	534,045		PP	530,018	533,874

EP – enostavna povprečja bralnih dosežkov

PP – povprečja, prilagojena glede na *relativni trud*

Dosežki vseh deklic v Avstraliji so se od leta 2000 do leta 2003 nekoliko zmanjšali, vendar je razlika majhna (1,659 točke). Isto velja tudi po tem, ko upoštevamo njihov *relativni trud* (razlika v tem primeru znaša 0,123 točke). V Nemčiji so dosežki deklic leta 2003 za 10,729 točke boljši kakor v letu 2000, ko upoštevamo tudi njihov *relativni trud*, pa zvišanje dosežkov glede na leto 2000 znaša le še 4,022 točke. Višje dosežke deklic v Nemčiji leta 2003 je torej mogoče pojasniti z večjim trudom učenk, vendar le deloma.

Slika dosežkov deklic iz skupine *realnih ocenjevalcev* kaže, da se njihovi dosežki v letu 2003 znižajo za 1,556 točke glede na leto 2000, vendar pa ta razlika takorekoč izgine, če upoštevamo tudi *relativni trud* učenk (v tem primeru razlika v dosežkih znaša le še 0,227 točke). Bralni dosežki deklic iz Nemčije so v letu 2003 za 5,082 točke višji kakor leta 2000, ob upoštevanju *relativnega truda* pa se ta razlika zmanjša na 3,856 točke v primerjavi z letom 2000.

V obeh državah so rezultati, prilagojeni *relativnemu trudu* učencev, višji od primerljivih enostavnih vrednosti, saj smo pri prilagajanju rezultatov predpostavili, da se vsi učenci obnašajo kot *prizadevni realisti*. Vendar pa naše ugotovitve tudi v tem primeru niso bile nič drugačne.

V nadaljevanju bomo za obe državi raziskali še trende dosežkov glede na spol. Rezultati za avstralske učence in učenke v letih 2000 in 2003 so prikazani v tabeli 19.

Za Avstralijo je iz leta v leto značilno večanje razlik v bralnih dosežkih glede na spol, pri čemer deklice vse bolj prekašajo dečke. Še več – kot kaže, je razlog za to prej padanje dosežkov pri dečkih, kakor dejansko izboljšanje dosežkov pri deklicah. Četudi vsi povprečni dosežki zrastejo, ko upoštevamo tudi *relativni trud* učencev in učenk, pa splošni vzorec ostaja nespremenjen.

Večanje vrzeli med dosežki dečkov in deklic je sicer opazno tudi v skupini *realnih ocenjevalcev*, vendar pa tu ni tako očitno kakor v skupini, ki jo sestavljajo *vsi učenci*. Zdi se, da v Avstraliji niti v skupini *vseh učencev* niti v skupini *realnih ocenjevalcev* rezultati raziskave PISA na področju bralne pismenosti niso odvisni od tega, koliko se učenci potrudijo pri reševanju preizkusa.

Rezultate za nemške učence in učenke v letih 2000 in 2003 prikazuje tabela 20; vzorec dosežkov glede na spol je v Nemčiji veliko bolj zanimiv. Enostavne vrednosti tukaj tako pri dečkih kot pri deklicah kažejo povišanje dosežkov, ki je pri deklicah nekoliko izrazitejše. Ko pa rezultate prilagodimo glede na *relativni trud* učencev, se raven dosežkov pri učenkah zniža, dosežki učencev pa se spremenijo tako, da razlika v dosežkih med leti dobi negativno vrednost. Prilagojeni rezultati se bolje ujemajo z avstralskimi.

Zgovorna je tudi slika pri skupini *realnih ocenjevalcev*, ki kaže, da dečki v Nemčiji od leta 2000 do 2003 niso izboljšali svojih bralnih dosežkov, kar je še bolj očitno, če pri rezultatih upoštevamo tudi njihov *relativni trud*.

Zaključimo lahko, da trud učencev sicer deloma pojasni povišanje dosežkov pri deklicah, vendar hkrati prikriva upad dosežkov pri dečkih. Rezultati, kjer je upoštevan tudi *relativni trud* učencev, so bolj podobni avstralskim in kažejo, da se razlike med spoloma povečujejo. Ko pri rezultatih upoštevamo tudi trud učencev, se pokaže upad in ne porast bralnih dosežkov.

Tabela 19: Enostavna povprečja bralnih dosežkov in povprečja, prilagojena glede na relativni trud, za vse učence in za učence iz skupine realnih ocenjevalcev v Avstraliji

vsi učenci		2000	2003	realni ocenjevalci		2000	2003
deklice	EP	547,085	545,426	deklice	EP	553,337	551,781
	PP	564,827	564,950		PP	565,710	565,483
dečki	EP	512,614	506,087	dečki	EP	521,778	515,174
	PP	536,198	530,363		PP	536,956	531,138

EP – enostavna povprečja bralnih dosežkov

PP – povprečja, prilagojena glede na *relativni trud*

Tabela 20: Enostavna povprečja bralnih dosežkov in povprečja, prilagojena glede na relativni trud, za vse učence in za učence iz skupine realnih ocenjevalcev v Nemčiji

vsi učenci		2000	2003	realni ocenjevalci		2000	2003
deklice	EP	502,198	512,927	deklice	EP	515,104	520,186
	PP	530,023	534,045		PP	530,018	533,874
dečki	EP	466,437	470,513	dečki	EP	486,600	482,375
	PP	505,863	499,640		PP	505,473	499,215

EP – enostavna povprečja bralnih dosežkov

PP – povprečja, prilagojena glede na *relativni trud*

Zaključek

Podatki o tem, koliko se učenci potrudijo za raziskavo PISA, omogočajo različne primerjave med cikli raziskave in državami. Trud učencev smo izrazili s pomočjo spremenljivke *relativni trud*, ki smo jo določili tako, da je z različnimi stopnjami pokazala, kako zavzeto so se učenci lotili reševanja nalog. Večina učencev je svoj trud za raziskavo PISA precej realno ocenila – ponavadi so menili, da so se pri reševanju nalog za raziskavo PISA potrudili približno toliko ali nekoliko manj, kakor bi se, če bi se rezultati upoštevali pri šolskih ocenah. Le majhen delež učencev je nerealno ocenil svoj trud; ponavadi so bili bralni dosežki teh učencev nižji.

Ključna ugotovitev, da je trud učencev dokaj primerljiv v večini držav, izpodbija trditev, da razlike v tem, koliko zavzeto se učenci lotijo reševanja nalog, onemogočajo mednarodne primerjave. Trud učencev na njihove dosežke vpliva približno toliko kot nekateri drugi dejavniki, na primer enostarševska družina, spol in socialno-ekonomsko ozadje. Države, v katerih učencih pogostejše nerealno ocenjujejo svoj trud, s čimer najbrž želijo dajati socialno zaželene odgovore, imajo ponavadi tudi nižje bralne dosežke.

Razmerje med *relativnim trdom*, bralnimi dosežki in spolom kaže, da so ocene, s katerimi deklice označujejo svoj trud za PISO, višje od ocen dečkov. Korelacija med bralnimi dosežki in spolom se podvoji, ko upoštevamo trud učencev, razlike v bralnih dosežkih med dečki in deklicami pa se zmanjšajo.

Študija primera Avstralije in Nemčije na podatkih raziskave PISA iz let 2000 in 2003 je pokazala, da se učenci v Avstraliji bolj potrudijo kakor v Nemčiji, vendar pa se je trud učencev za raziskavo PISA 2003 v Nemčiji povečal glede na leto 2000. Sodeč po dosežkih tistih učencev, ki so realno ocenili svoj trud, je bila razlika v bralnih dosežkih med državama leta 2000 manjša kakor tri leta pozneje. Nenazadnje bi s trdom učencev lahko pojasnili izboljšanje bralnih dosežkov dečkov v Nemčiji leta 2003.

Viri

- ACER. (2001). *PISA in brief from Australia's perspective*. Camberwell, NSW, Avstralija: Avtor.
- Adams, R. J. in Wu, M. L. (2002). *PISA 2000 technical report*. Pariz: OECD.
- Baumert, J. in Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, 16, 441-462.

- Bracey, G. (1999). The ninth Bracey report on the condition of public education. *Phi Delta Kappan*, 81(2), 147-168.
- Brown, S. M. in Walberg, H. J. (1993). Motivational effects on test scores of elementary students. *Journal of Educational Research*, 86, 133-136.
- Deutsche Presse Agentur (2001, 4 December). *Miserable Noten für deutsche Schüler*. Frankfurter Allgemeine Zeitung.
- Fertig, M. (2003). *Who's to blame? The determinants of German students' achievement in the PISA 2000 study*. Discussion Paper No. 739, Institute for the Study of Labor (IZA) retrieved March, 2003, from <http://www.iza.org>
- Holliday, W. G. in Holliday, B. W. (2003). Why using international comparative math and science achievement data from TIMSS is not helpful. *The Education Forum*, 67(Spring), 250-257.
- Karmos, A. H. in Karmos, J. S. (1984). Attitudes towards standardised achievement tests and their relation to achievement test performance. *Measurement and Evaluation in Counseling and Development*, 17, 56-66.
- King, G., Murray, C., Salomon, J. in Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, 98(1), 191-207.
- Kiplinger, V. L. in Linn, R. L. (1996). Raising the stakes of test administration: The impact on student performance on the National Assessment of Educational Progress. *Educational Assessment*, 3(2), 111-133.
- Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M., et al. (2002). *German scale handbook for PISA 2000*. Berlin, Germany: Max-Planck-Institut für Bildungsforschung.
- Mislevy, R. J. (1995). What can we learn from international assessments? *Educational Evaluation and Policy Analysis*, 4, 419-437.
- OECD. (2001). *Knowledge and skills for life. First results from PISA 2000*. Pariz: Avtor.
- OECD. (2003). *Literacy skills for the world of tomorrow. Further results from PISA 2000*. Pariz: Avtor.
- OECD. (2004a). *First results from PISA 2003. Executive Summary*. Pariz: Avtor.
- OECD. (2004b). *Learning for tomorrow's world. Results for PISA 2003*. Pariz: Avtor.
- OECD. (2005). *Technical report for the OECD programme for international student assessment 2003*. Pariz: Avtor.

- Olafsson, R. F., Halldorsson, A. M. in Bjornsson, J. K. (2006). Gender and the urban-rural differences in mathematics and reading: An overview of PISA 2003 results in Iceland. V: J. Mejdning and A. Roe (ur.), *Northern lights on PISA 2003: A reflection from the Nordic countries* (pp. 185-198). Copenhagen, Den-mark: Nordic Council of Ministers.
- O'Neil, H. F., Jr., Sugrue, B. in Baker, E. L. (1996). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. *Educational Assessment*, 3, 135-157.
- O'Neill, H. F., Sugrue, B., Abedi, J., Baker, E. L. in Golon, S., (1997). *Final report of experimental studies on motivation and NAEP performance* (CSE Technical Report 427). Los Angeles, CA. University of California: CREST.
- O'Neill, H. F., Abedi, J., Lee, C., Myoshi, J. in Mastergeorge, A. (2004). *Monetary incentives for low-stakes tests* (CSE Technical Report 625). Los Angeles, CA. University of California: CREST.
- Pintrich, P. R. in Schunk, D. H. (2002). *Motivation in education: Theory, research, and applications* (2nd ed.). Upper Saddle River, NJ: Merrill Prentice Hall.
- Schulz, W. (2006, April). *Measuring the socio-economic background of students and its effect on achievement in PISA 2000 and 2003*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.
- Turner, R. in Adams, R. J. (2007). The Programme for international student assessment: An overview. *Journal of Applied Measurement*, 8, 237-248.
- Wainer, H. (1993). Measurement problems. *Journal of Educational Measurement*, 30, 1-21.
- Weiner, B. (1986). *An attributional theory of motivation and emotion*. New York: Springer-Verlag.
- Wise, S. L. in DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, 10(1), 1-17.
- Wolf, L. F. in Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, 8, 227-242.

VPLIV METODOLOGIJE IZENAČEVANJA LESTVIC NA TRENDE, UGOTOVLJENE V RAZISKAVI PISA

Eveline Gebhardt
Avstralski inštitut za raziskovanje v izobraževanju

Raymond J. Adams
Avstralski inštitut za raziskovanje v izobraževanju
Univerza v Melbournu

Leta 2005 je PISA objavila kazalnike trendov, ki so med seboj primerjali rezultate raziskav PISA 2000 in PISA 2003. V tem prispevku preučujemo, v kolikšni meri so ugotovitve teh analiz trendov odvisne od izbranih metodologij izenačevanja lestvic, regresijskih modelov in načina povezovanja postavk. Trende je PISA ugotavljala tako, da je preizkuse iz let 2000 in 2003 med seboj uskladila, in sicer z uporabo metodologije, ki temelji na Raschevem modelu: ocenjena linearna transformacija je rezultate iz leta 2003, razvrščene po Raschevi lestvici, primerjala z rezultati iz leta 2000, razvrščenimi po isti lestvici.

V prispevku primerjamo rezultate tega pristopa z alternativnim, ki predpostavlja razvrščanje podatkov raziskav PISA 2000 in PISA 2003 na skupno Raschevo lestvico, ločeno za vsako državo. Opozoriti je treba, da se pri tem pristopu parametri postavk ocenjujejo ločeno za vsako državo, medtem ko so bile pri pristopu linearne transformacije za vse države uporabljene skupne ocene parametrov postavk. PISA je kot primarne kazalnike trenda ugotovila spremembe v povprečnih rezultatih med letoma 2000 in 2003. Te srednje vrednosti niso prilagojene spremembam spremljajočih značilnosti vzorcev raziskav PISA 2000 in 2003 – srednje vrednosti so tako marginalne in ne pogojne. Z uporabo pogojnih srednjih vrednosti namesto marginalnih je mogoče glede trendov priti do drugačnih zaključkov, tako na ravni države kakor tudi znotraj nje.

Uvod

Raziskava PISA primerja in razvršča države po dosežkih, obenem pa po državah ugotavlja tudi trende teh dosežkov. Primerjavo matematičnih, bralnih in naravoslovnih dosežkov med cikloma preverjanj PISA 2000 in PISA 2003 omogoča povezovanje obeh preverjanj. Določen delež vseh nalog, ki so preverjale matematično, bralno in naravoslovno pismenost v raziskavi PISA 2000, je bil uporabljen tudi pri preverjanju v ciklu PISA 2003 – to so tako imenovane vezne naloge. Pri preverjanju matematične pismenosti je bilo takih nalog dvajset, bralne osemindvajset in naravoslovne petindvajset.

V raziskavi PISA 2003 je bila za matematiko določena nova lestvica. Ker je bilo področje matematike v raziskavi PISA 2003 razdeljeno na podpodročja, ki jih v raziskavi PISA 2000 ni bilo, trendi pri matematiki niso predstavljeni (OECD, 2004).

Postopki in modeli, po katerih so razvrščeni podatki raziskav PISA, so podrobneje predstavljeni v Adams in Carstensen (2002) ter Adams, Wu in Carstensen (2006), zato jih v tem prispevku ne obravnavamo. Za razvrščanje podatkov je bila uporabljena aplikacija ConQuest (Wu, Adams in Wilson, 1998) Avstralskega inštituta za raziskovanje v izobraževanju. Način povezovanja lestvic bralne in naravoslovne pismenosti iz raziskav PISA 2000 in PISA 2003 je podrobno predstavljen v OECD (2005b). V grobem ga je mogoče povzeti takole (gl. tudi shemo 1):

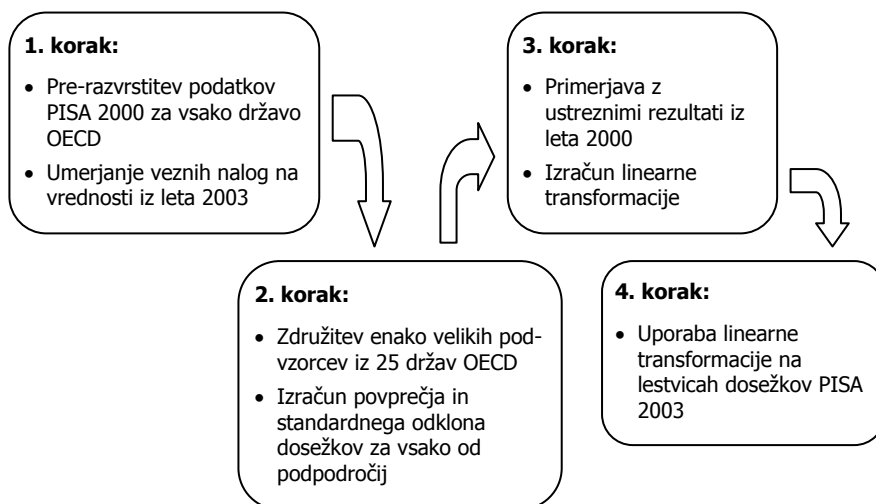
- Korak 1. Podatki raziskave PISA 2000 so bili za vsako državo OECD prerazvrščeni s pomočjo skupne pogojenosti (*full-conditioning*) (gl. Mislevy in Sheehan, 1987; Adams et al. 2006) in preko veznih postavk povezani z mednarodnimi vrednostmi raziskave PISA 2003.¹
- Korak 2. Za vsako področje pismenosti sta bili izračunani srednja vrednost in standardni odklon na kombiniranem nizu podatkov raziskave PISA 2000 za 25 držav OECD. Podatki vseh držav so bili uteženi enako.
- Korak 3. Opravljena je bila primerjava srednjih vrednosti in standardnih odklonov, izračunanih v koraku 2, z ustreznimi srednjimi vrednostmi in standardnimi odkloni prvotne razvrstitve iz raziskave PISA 2000.² Nato so bile izračunane linearne transformacije, ki so rezultate raziskave PISA 2003 preslikale tako, da sta bila srednja vrednost in standardni odklon preslikave enaka srednji vrednosti in standardnemu odklonu rezultatov raziskave PISA 2000.

¹ Ocene mednarodnih parametrov postavk za raziskavo PISA 2003 so objavljene v OECD (2005b).

² Ocene mednarodnih parametrov postavk za raziskavo PISA 2000 so objavljene v Adams in Wu (2002).

Korak 4. Linearna transformacija iz koraka 3 je bila uporabljena na lestvicah raziskave PISA 2003, tako da so bile izenačene z lestvicami raziskave PISA 2000.

Shema 1: Koraki prvotnega povezovanja dosežkov v raziskavah PISA 2000 in PISA 2003



Ugotovljeni trendi so z gledišča sodelujočih držav eden od bolj pomembnih motivov sodelovanja v raziskavi PISA. Pomik po razvrstitveni lestvici vseh sodelujočih držav oziroma spremenjena raven dosežkov neke države sta lahko temeljna kazalnika uspešnosti oziroma izobraževalnega sistema neke države (OECD, 2004). Pri iskanju odgovorov na bolj specifična vprašanja utegnejo sodelujoča gospodarstva zanimati tudi drugi trendi, na primer spremembe v odstopanju dosežkov, vplivi spola, vplivi socialno-ekonomskega statusa in podobno. Namen tega prispevka je preučiti morebiten vpliv izbrane metodologije izenačevanja lestvic na ugotovljene trende, ki zanimajo sodelujoče v raziskavi.

V kontekstu metodologije trenda v prispevku obravnavamo tri vprašanja. Najprej nas je zanimalo, kako so bili ocenjeni parametri postavk. Pri raziskavah PISA je v veljavi način, pri katerem se oceni enoten niz *mednarodnih* parametrov postavk, ki se jih potem uporabi pri oceni zmožnosti učencev v vsaki državi. Z umerjanjem postavk na reprezentativnem pod-vzorcu 30.000 učencev iz tridesetih sodelujočih OECD držav je pridobljen enoten niz ocen parametrov postavk po Raschu (Adams 2002; OECD 2005b). Ta pristop je izbran zato, ker ustreza namenu raziskave: vse učence razvrsti na enotno lestvico, ne glede na državo, iz katere prihajajo.

Potencialna šibkost tega pristopa je v tem, da pri obnašanju postavk spregleda medsebojni vpliv postavke in države. Te, tako imenovane razlike v

delovanju postavk po državah (DIF, *Differential Item Functioning*), so za mednarodne raziskave običajne (Adams in Carstensen, 2002; Adams, Wu in Macaskill, 1998; Mullis in Martin, 1998); obseg medsebojnega vpliva določa veljavnost primerjav med državami.

Namesto ocenjevanja trendov s skupnim nizom parametrov postavk je pri razvrščanju mogoče uporabiti niz parametrov za vsako državo posebej, pri čemer so lestvice med seboj primerljive preko skupne srednje vrednosti niza postavk, ki so uporabljene v vseh državah. V tem primeru so ocene trendov med državami sicer manj primerljive, bi pa za vsako državo posebej temeljile na bolj primernih ocenah parametrov postavk.

Drugo vprašanje, ki ga obravnavamo v tem besedilu, je medsebojna povezanost lestvic raziskav PISA 2000 in PISA 2003. Kot je bilo omenjeno zgoraj in opisano v besedilu OECD (2005b), je transformacija lestvice iz leta 2003 v lestvico iz leta 2000 določena na podlagi prerazvrščanja podatkov iz leta 2000 s postavkami, ki so sidrane pri njihovih vrednostih iz leta 2003. Nato je izračunana transformacija, ki bi ocene iz prerazvrščanja preslikala na isto lestvico, kot je bila uporabljena pri prvotnem razvrščanju. Ta postopek sledi postopku, ki ga v ZDA uporabljajo pri ugotavljanju nacionalnega razvoja v izobraževanju (*US National Assessment of Educational Progress*); je oblika običajnega določanja postavk, ki ga ponekod imenujejo Stocking-Lordovo usklajevanje (*Stocking-Lord equating*) (Johnson in Carlson, 1992; Hedges in Vevea, 2003).

V drugačnih okoliščinah bi bil bolj običajen pristop spajanje dveh nizov podatkov, tako da bi bile postavke umerjene sočasno (Kim in Cohen, 1998). V tem primeru so ocene parametrov postavk in ocene sposobnosti samodejno določene na skupni lestvici, tako da linearne transformacije usklajevanja niso potrebne. Treba pa je poudariti, da ta pristop pri raziskavah, kot je PISA, ni mogoč, ker nujno vključuje določanje trendov, ki temeljijo na rezultatih (v tem primeru na rezultatih iz leta 2000), ki se kaj lahko razlikujejo od tistih, ki so bili objavljeni prvotno.

Tretje, kar ugotavljamo v tem prispevku, je, v kolikšni meri naj bodo trendi prilagojeni demografskim spremembam ciljne populacije in dejanskega vzorca. Predpostavimo, na primer, da imata dva niza podatkov iz različnih ciklov iz določenega razloga različen delež učenk in učencev in da je to posledica razlik v vzorčenju in ne stvarne spremembe v populaciji. V tem primeru lahko kontrola deleža dečkov pri določanju trenda vodi v laže razložljivo (in morda veljavnejšo) izmero trenda v konkretni državi. V drugem primeru vzemimo državo s pozitivnim trendom, ki ni prilagojen možnim spremembam v spremljajočih značilnostih učencev, s stabilnim deležem dečkov med cikloma, toda povečanim številom manjkajočih vrednosti za spol. V primeru, da učenci, ki ne navedejo spola, naloge rešijo slabo, kakor se ponavadi tudi zgodi, bi vključevanje kazalnika za manjkajoče vrednosti v multiplo regresijsko analizo z namenom prilagoditve trenda

spremembam v vzorcu vodilo v precenitev dejanskega trenda. V tem primeru je neprilagojeni trend veljavnejši kazalnik trenda.

V besedilu smo neprilagojene trende poimenovali *marginalni* trendi (*marginal*), prilagojene pa *pogojni* (*conditional*).

Izračunani sta bili dve različni oceni trendov, ki smo ju primerjali med seboj in s prvotno oceno trendov (gl. tabelo 1). Oba trenda temeljita na nacionalnih parametrih postavk in skupnem razvrščanju podatkov raziskav PISA 2000 in PISA 2003. Ena ocena trendov je marginalna razlika (neprilagojena) med cikloma za vsako državo posebej, druga je pogojna razlika pri preverjanju spremljajočih spremenljivk, ki določajo okoliščine učencev, zbranih v obeh ciklih.

Tabela 1: Spremenljivke pri metodologiji za določanje trenda

Parametri postavk	Metoda usklajevanja lestvic	Indikator trendov	
		neprilagojeni: marginalni trendi	prilagojeni: pogojni trendi
Mednarodni parametri postavk	skupno umerjanje postavk linearna transformacija	— (1) v poročilu PISA	—
Nacionalni parametri postavk	skupno umerjanje postavk linearna transformacija	(2) to besedilo —	(3) to besedilo —

Metodologija

Razvrščanje je bilo narejeno na podoben način kakor v izvirni PISA analizi. Ker so za tovrstno razvrščanje potrebni številni koraki, postopek pa je razmeroma zapleten, podrobnega opisa tukaj ne dajemo. Več o tem si lahko preberete v tehničnih poročilih raziskave PISA in priročnikih za analizo podatkov (Adams, 2002; OECD, 2005a; OECD, 2005b). Pri vzorčenju dijakov so bile uporabljene uteži, delovni zvezki so bili prilagojeni, izbrani so bili dosežki v obliki pričakovanih vrednosti (*plausible values*), pri ugotavljanju nepristranih standardnih napak pa je bilo uporabljeno uravnoteženo ponavljajoče se kopiranje (*BRR, balanced repeated replication*). Pri umerjanju postavk nismo upoštevali rezultatov učencev s posebnimi potrebami (ti so imeli prilagojene delovne zvezke), ampak smo njihove dosežke vključili šele pri določanju rezultatov.

Alternativna trenda smo ocenili v treh korakih: najprej smo nacionalne parametre postavk ocenili v enodimenzionalnih modelih, pri čemer smo uporabili združene nacionalne podatke raziskav PISA 2000 in PISA 2003. Upoštevani so bili samo učenci, ki so odgovarjali na postavke določenega področja (ne pa tudi učenci, ki so odgovarjali na prilagojeni delavni zvezek) ter vezne in nevezne postavke. V tem modelu je bila spremenljivka delovnega zvezka uporabljena kot t.i. faceta (*facet*) (Linacre, 1994), z največ (glede na področje preverjanja) devetimi delovnimi zvezki za raziskavo PISA 2000 in največ trinajstimi delovnimi zvezki za raziskavo PISA 2003. Delovni zvezek je moral biti dodan kot faceta, ker

je bilo odstopanje srednjih vrednosti delovnih zvezkov ponekod večje od pričakovanega (Adams in Carstensen, 2002). Vpeljava facete postavke je delno odpravila vpliv pozicije postavke v delovnem zvezku na ocenjene parametre postavke.

V naslednjem koraku smo za področja matematične, bralne in naravoslovne pismenosti s pomočjo tridimenzionalnega Raschevega modela vrednostim, ki so bile ocenjene v prvem koraku, pripisali pričakovane vrednosti parametrov veznih postavk. V modelu smo uporabili vse učence (tudi tiste s posebnimi potrebami) ter vezne in nevezne postavke. Za kodiranje delovnih zvezkov smo uporabili kontrastno kodiranje (*deviation contrast coding*), za kodiranje PISA cikla pa enostavno kodiranje (t. i. *dummy coding*). Spremenljivke so bile uporabljene kot spremenljivke latentne regresije; na ta način smo dobili tridimenzionalni model, ki je matematično ekvivalenten enodimenzionalnemu modelu z delovnimi zvezki kot facetami.³ Facete delovnih zvezkov smo nato pretvorili v kontraste in jih kot take vključili v multidimenzionalni pogojni model (*multi-dimensional conditioning model*). Oblikovani so bili kontrasti delovnih zvezkov, in sicer tako, da so referenčno skupino sestavljajo vsi učenci, ki so reševali naloge z nekega področja. Vse učence smo v skupino vključili zato, ker smo lahko s pomočjo te referenčne skupine zapolnili tiste primere, pri katerih učenci na postavke z nekega področja niso odgovarjali. Ker je bila razlika med srednjimi vrednostmi delovnih zvezkov večja od pričakovane, bi lahko ob upoštevanju samo tistih učencev, ki so reševali samo en, točno določen delovni zvezek, prišlo do precenitve oziroma podcenitve dosežkov učencev, ki na postavke nekega področja niso odgovorili. Poleg tega so bile kot regresijske spremenljivke vključene tudi spremljajoče spremenljivke obeh ciklov. Te so bile spol, najvišji socialno-ekonomski status staršev (HISEI), starost, doma govorjeni jezik, povprečje ocen pri matematiki, branju in naravoslovju, poklic matere in očeta in njihovi kazalniki za manjkajoče vrednosti.

Nazadnje so bile v SPSS opravljene enojne in multiple regresijske analize (z uporabo uteži iz vzorčenja učencev – *student sampling weights* in metode ponavljanja BRR) za izračun marginalnih in pogojnih trendov ter njihovih verjetnosti. V enojni regresijski analizi je bila *spremenljivka* cikla (0=PISA 2000, 1=PISA 2003) edina neodvisna spremenljivka. Regresijski koeficient te *spremenljivke* je bil uporabljen kot marginalni trend na določenem področju v posamezni državi.

V multipli regresijski analizi so bile kot neodvisne spremenljivke upoštevane naslednje:

- starost (v mesecih, manjkajoči podatki pa nadomeščeni s povprečno starostjo v ciklu);
- manjkajoč podatek o starosti (1=manjka, 0=ne manjka);

³ Konfiguriranje aplikacije ConQuest na način, da bi dovoljevala različne facetne učinke za vsako dimenzijo, je precej zapleteno.

- spol (1=ženski, 2=moški ali manjka);
- manjkajoč podatek o spolu (1=manjka, 0=ne manjka);
- HISEI (na lestvici od 19 do 90, manjkajoči podatki pa nadomeščeni s srednjo vrednostjo HISEI v ciklu);
- manjkajoč HISEI (1=manjka, 0=ne manjka);
- jezik, govorjen doma (0=jezik, v katerem je sestavljen test, ali podatek manjka, 1=drug jezik);
- manjkajoč podatek o jeziku, govorjenem doma (1=manjka, 0=ne manjka);
- cikel raziskave (0=PISA 2000, 1=PISA 2003).

Te spremenljivke so bile izbrane, ker predstavljajo ključne nespremenljive spremenljivke, ki so na voljo v podatkovni bazi PISA. Gre za spremenljivke, ki niso podvržene vplivu izobraževalne politike in prakse, zato se pogosto poudarja, da bi bile primerjave med državami in v različnih časovnih obdobjih lahko pravilnejše, če bi bile prilagojene razlikam med učenci glede teh spremenljivk. (Delni) regresijski koeficient spremenljivke cikla raziskave je bil uporabljen kot pogojni trend na posameznem področju v določeni državi.

Opozoriti je treba, da v standardne napake, opisane v tem prispevku, niso bile vključene ocene napak povezovanja (Monseur in Berezner, 2007). Eden izmed razlogov za to je, da o tem vprašanju še ni dosežen konsenz. Drugi, bolj praktičen razlog, pa je, da se v nasprotju s prvotno analizo parametri postavk v posameznih državah razlikujejo, kar pomeni, da je treba napako povezovanja oceniti za vsako državo posebej, kar bi bilo še posebej težavno v luči dejstva, da so nekatere države zaradi napak v prevodih že odstranile vezne naloge.

Rezultati

Za potrebe te študije je bilo v analizo vključenih 28 držav. Sodelujoče države, število uteženih in neuteženih učencev ter srednja vrednost dosežkov v raziskavah PISA 2000 in PISA 2003 so navedeni v tabeli 2. Seznam držav vključuje 27 od 28 držav članic OECD, ki so sodelovale v obeh ciklih, in Rusko federacijo. Grčija, država članica OECD, ki je sodelovala v obeh ciklih, je bila iz analize izključena po nasvetu strokovnjaka za vzorčenje pri raziskavah PISA (K. F. Rust, osebno sporočilo, 8.3.2006), in sicer zaradi opaženih nekonsistentnosti v utežeh učencev. Nizozemska in Velika Britanija nista bili vključeni v uradna poročila raziskave PISA, ker sta imeli nesprejemljivo stopnjo nerešenih nalog v enem izmed ciklov. V pričujoči analizi sta zajeti, vendar je treba rezultate interpretirati previdno.

Tabela 2: Sodelujoče države, število uteženih (weighted) in neuteženih (unweighted) učencev ter srednja vrednost dosežkov v raziskavah PISA 2000 in PISA 2003

	Neutežen N		Utežen N		Branje PISA 2000		Branje PISA 2003		Naravoslovje PISA 2000		Naravoslovje PISA 2000	
	PISA 2000	PISA 2003	PISA 2000	PISA 2003	Povpr.	S.E.	Povpr.	S.E.	Povpr.	S.E.	Povpr.	S.E.
Avstralija	5176	12551	229152	235591	528	(3,5)	526	(2,2)	528	(3,5)	529	(2,1)
Avstrija	4745	4597	71547	85931	507	(2,4)	498	(3,8)	519	(2,5)	494	(3,6)
Belgija	6670	8796	110095	111831	507	(3,6)	509	(2,6)	496	(4,3)	513	(2,4)
Kanada	29687	27953	348481	330436	534	(1,6)	531	(1,8)	529	(1,6)	527	(2,0)
Češka	5365	6320	125639	121183	492	(2,4)	489	(3,6)	511	(2,4)	524	(3,3)
Danska	4235	4218	47786	51741	497	(2,4)	501	(2,8)	481	(2,8)	487	(3,0)
Finska	4864	5796	62826	57884	546	(2,6)	540	(1,7)	538	(2,5)	550	(1,9)
Francija	4673	4300	730494	734579	505	(2,7)	499	(2,6)	500	(3,2)	518	(3,1)
Nemčija	5073	4660	826816	884358	484	(2,5)	495	(3,3)	487	(2,4)	505	(3,6)
Madžarska	4887	4765	107460	107044	480	(4,0)	480	(2,5)	496	(4,2)	498	(2,7)
Islandija	3372	3350	3869	3928	507	(1,5)	494	(1,4)	496	(2,2)	493	(1,5)
Irska	3854	3880	56209	54850	527	(3,2)	520	(2,5)	513	(3,2)	511	(2,6)
Italija	4984	11639	510792	481521	487	(2,9)	471	(3,0)	478	(3,1)	479	(3,2)
Japonska	5256	4707	1446596	1240054	522	(5,2)	507	(3,7)	550	(5,5)	536	(4,2)
Koreja	4982	5444	579109	533504	525	(2,4)	540	(3,1)	552	(2,7)	541	(3,5)
Luksemburg	3404	3923	4138	4080	441	(1,6)	479	(1,2)	443	(2,3)	483	(1,4)
Mehika	4600	29983	960011	1071650	422	(3,3)	394	(4,2)	422	(3,2)	394	(3,7)
Nizozemska	2503	3992	157327	184943	532	(3,4)	516	(2,9)	529	(4,0)	529	(3,2)
Nova Zelandija	3667	4511	46757	48638	529	(2,8)	525	(2,7)	528	(2,4)	525	(2,5)
Norveška	4147	4064	49579	52816	505	(2,8)	497	(2,7)	500	(2,7)	482	(3,0)
Poljska	3654	4383	542005	534900	479	(4,5)	495	(2,9)	483	(5,1)	494	(3,0)
Portugal	4585	4608	99998	96857	470	(4,5)	473	(3,7)	459	(4,0)	472	(3,4)
Rusija	6701	5974	1968131	2153373	462	(4,2)	439	(3,9)	460	(4,7)	483	(4,2)
Španija	6214	10791	399055	344372	493	(2,7)	478	(2,5)	491	(3,0)	479	(2,6)
Švedska	4416	4624	94338	107104	516	(2,2)	515	(2,5)	512	(2,5)	510	(2,8)
Švica	6100	8420	72010	86491	494	(4,2)	505	(3,1)	496	(4,4)	514	(3,6)
Velika Britanija	9340	9535	643041	698579	523	(2,6)	511	(2,4)	532	(2,7)	523	(2,7)
ZDA	3846	5456	3121874	3147089	504	(7,0)	495	(3,0)	499	(7,3)	494	(3,0)

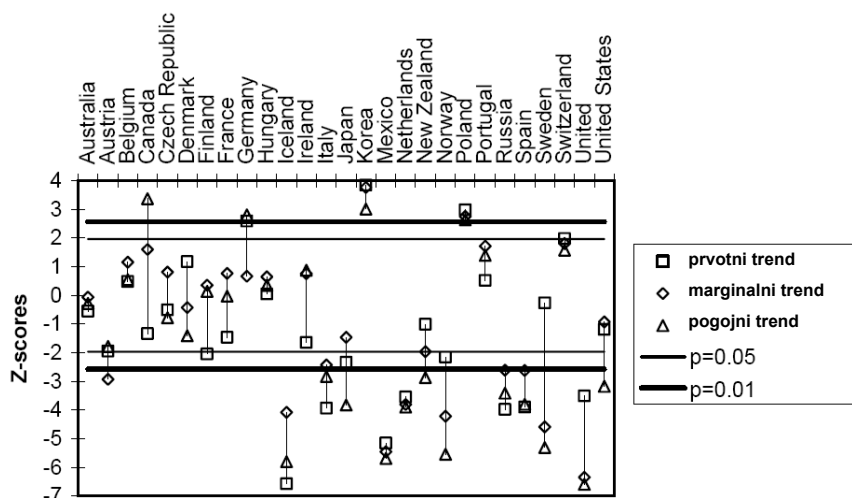
Opravljen je bil primerjava dveh skupin rezultatov: prvotnih rezultatov trendov in rezultatov, dobljenih s pomočjo dveh alternativnih metod izenačevanja dosežkov. Prvotni rezultati, ki so predstavljeni tukaj, in predhodno objavljeni rezultati (OECD, 2001) se nekoliko razlikujejo, ker so bili podatki iz leta 2003 preračunani na lestvico s pomočjo drugačnega kodiranja kontrastov (*deviation contrast coding* namesto *simple contrast coding*). Rezultati, dobljeni s prvo metodo, se imenujejo marginalni trendi, ki predstavljajo razlike v srednjih vrednostih med cikloma za vsako državo. Rezultati, dobljeni z drugo metodo, t.i. pogojni trendi, pa predstavljajo razlike med srednjimi vrednostmi ciklov po opravljeni kontroli spremljajočih spremenljivk, tj. spola, starosti, HISEI, doma govorjenega jezika in njihovih spremenljivk za manjkajoče vrednosti.

V tabeli 3 je predstavljena pomembnost razlik med rezultati raziskav PISA 2000 in PISA 2003 v branju in naravoslovju za vsako državo (za regresijske koeficiente, standardne napake in standardne normalne rezultate glej dodatka A in B). Najprej so predstavljene razlike v *izvirnih* rezultatih, nato *marginalne* razlike in na koncu še *pogojne* razlike. Shemi 2 in 3 vsebujeta grafično predstavitev trendov standardnih normalnih rezultatov. Na teh grafih ima vsaka država tri oznake, po eno za vsako oceno trenda. Kvadratna oznaka predstavlja prvotni trend, karo oznaka marginalni trend in trikotna pogojni trend. Vsaka ocena trenda je enaka razliki v srednjih vrednostih v posamezni državi med letoma 2000 in 2003, deljena s standardno napako razlike, s čimer dobimo standardiziran rezultat. Pozitivne številke označujejo povečanje uspešnosti v posamezni državi med letoma 2000 in 2003, negativne pa njeno zmanjšanje. Vodoravne črte predstavljajo referenčno linijo za preizkus pomembnosti teh sprememb skozi čas. Ocene trendov med tankimi črtami niso relevantne ($p > 0,05$). Povedano drugače, uspešnost v posamezni državi se s časom ni povečala. Ocene trendov nad in pod debelimi črtami so pomembne na ravni 0,01, nad in pod tankimi vodoravnimi črtami pa na ravni 0,05. Fokus teh grafov je razdalja med vsemi tremi oznakami za vsako državo in s tem povezana morebitna sprememba v pomembnosti. Zaradi poenostavitve grafične predstavitve so bile iz grafov izključene izstopajoče ocene trendov za Luksemburg.

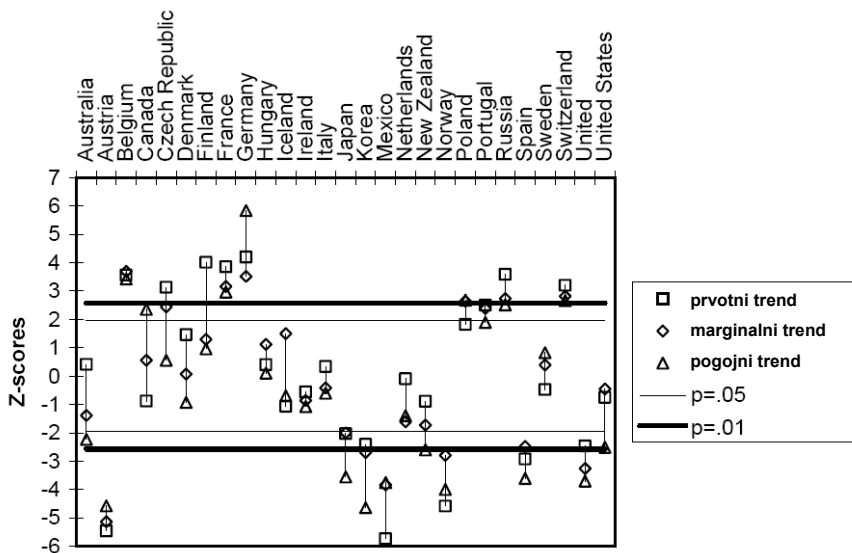
Tabela 3: Primerjava statistične pomembnosti trendov za tri alternativne metode izenačevanja dosežkov

	Branje			Naravoslovje		
	prvotni trend	marginalni trend	pogojni trend	prvotni trend	marginalni trend	pogojni trend
Avstralija	0	0	0	0	0	- -
Avstrija	-	- - -	-	- - -	- - -	- - -
Belgija	0	0	0	+ + +	+ + +	+ + +
Kanada	0	0	+ + +	0	0	+ +
Češka	0	0	0	+ + +	+ +	0
Danska	0	0	0	0	0	0
Finska	- -	0	0	+ + +	0	0
Francija	0	0	0	+ + +	+ + +	+ + +
Nemčija	+ + +	0	+ + +	+ + +	+ + +	+ + +
Madžarska	0	0	0	0	0	0
Islandija	- - -	- - -	- - -	0	0	0
Irska	-	0	0	0	0	0
Italija	- - -	- -	- - -	0	0	0
Japonska	- -	0	- - -	- -	- -	- - -
Koreja	+ + +	+ + +	+ + +	- -	- - -	- - -
Luksemburg	+ + +	+ + +	+ + +	+ + +	+ + +	+ + +
Mehika	- - -	- - -	- - -	- - -	- - -	- - -
Nizozemska	- - -	- - -	- - -	0	0	0
Nova Zelandija	0	- -	- - -	0	-	- - -
Norveška	- -	- - -	- - -	- - -	- - -	- - -
Poljska	+ + +	+ + +	+ + +	+	+ + +	+ + +
Portugal	0	+	0	+ +	+ +	+
Rusija	- - -	- - -	- - -	+ + +	+ + +	+ +
Španija	- - -	- - -	- - -	- - -	- -	- - -
Švedska	0	- - -	- - -	0	0	0
Švica	+ +	+	0	+ + +	+ + +	+ + +
Velika Britanija	- - -	- - -	- - -	- -	- - -	- - -
ZDA	0	0	- - -	0	0	- -
Opomba:	stopnja statistične pomembnosti		2003 višje od 2000	2003 višje od 2000		
	$p > 0,10$		0	0		
	$p < 0,10$		+	-		
	$p < 0,05$		+ +	- -		
	$p < 0,01$		+ + +	- - -		

Shema 2: Trije alternativni trendi med bralnimi dosežki v raziskavah PISA 2000 in PISA 2003 po posameznih državah



Shema 3: Trije alternativni trendi med naravoslovnimi dosežki v raziskavah PISA 2000 in PISA 2003 po posameznih državah



Prvotni in marginalni trendi

Pri prvotnem izračunu trendov so bili parametri postavk ocenjeni s pomočjo mednarodnega umeritvenega/kalibracijskega vzorca, in sicer ločeno za raziskavo PISA 2000 in raziskavo PISA 2003. Nasprotno so marginalni trendi temeljili na nacionalnih parametrih postavk in kombiniranem nizu podatkov iz raziskav PISA

2000 in PISA 2003. Različni metodi sta v nekaterih državah vodili v precejšnje razlike v trendih. Na primer: prvotni rezultati so nakazovali, da so bili švedski učenci leta 2003 enako uspešni v branju kot učenci, ki so sodelovali v raziskavi PISA 2000 (glej shemo 2). Z uporabo skupnega umerjanja (*joint calibration*) in parametrov postavk na ravni države pa je bil uspeh leta 2003 precej slabši kot leta 2000 ($p < 0,01$).

Ta razlika med prvotnim in marginalnim trendom je bila v nekaterih drugih državah precej manjša. Nekaj spremenljivk lahko pojasni variiranje teh razlik med posameznimi državami.

Učenci so v raziskavi PISA 2000 reševali 129 nalog oz. vprašanj, ki sledijo nalogi z uvodnim besedilom. Od teh 129 nalog jih je bilo 28 vključenih v raziskavo leta 2003 (nobena naloga ni bila dodana). Povprečna težavnost teh veznih nalog na mednarodni ravni je bila $-0,03$ (v točkah *logit*), medtem ko je imela preostala 101 naloga povprečno težavnost $0,01$ (v točkah *logit*). Tako so bile vezne naloge pri uporabi mednarodnih parametrov postavk nekoliko *lažje* kot nevezne naloge ($0,04$ točke *logit*). Razlika med težavnostjo veznih in neveznih nalog na ravni posamezne države se imenuje *relativna težavnost veznih nalog* in se med državami razlikuje. Povedano drugače, to je oblika razlike v delovanju preizkusa v različnih državah. Te povprečne težavnosti bralnih nalog in relativna težavnost veznih nalog so predstavljene v tabeli 4, ki ji sledi razlaga števil v tabeli. Drugi stolpec v tabeli 4 vsebuje povprečno težavnost nalog, vključenih le v raziskavo PISA 2000. Tretji stolpec vsebuje povprečno težavnost veznih nalog. Relativna težavnost veznih nalog na ravni posamezne države je razlika med tema dvema stolpcema. Pozitivna vrednost relativne težavnosti veznih nalog pomeni, da so bile vezne naloge lažje od nalog, vključenih le v raziskavo PISA 2000. Povprečna težavnost parametrov veznih postavk na področju branja na Švedskem je bila na primer $-0,20$, povprečje za nevezne postavke pa je znašalo $0,06$. Tako so bile vezne naloge na Švedskem v povprečju za $0,26$ točke *logit* lažje od neveznih.

Tabela 4: Izračun relativne težavnosti veznih nalog na področju branja

	Branje				
	Naloge, vključene le v raziskavo 2000	Povprečna težavnost veznih nalog	Relativna težavnost veznih nalog na ravni posamezne države	(Prilagojena) relativna težavnost na mednarodni ravni	Relativna težavnost veznih nalog
International 2000	0,01	-0,03		0,04	
International 2003		-0,41			
Avstralija	0,01	-0,04	0,05	0,04	0,01
Avstrija	0,02	-0,06	0,07	0,04	0,03
Belgija	-0,01	0,03	-0,04	0,04	-0,08
Kanada	-0,02	0,05	-0,07	0,04	-0,11
Češka	-0,01	0,04	-0,05	0,04	-0,09
Danska	0,03	-0,12	0,15	0,04	0,11
Finska	0,01	-0,04	0,05	0,04	0,01
Francija	-0,02	0,07	-0,09	0,04	-0,13
Nemčija	0,01	-0,05	0,06	0,04	0,02
Madžarska	-0,03	0,09	-0,11	0,03	-0,14
Islandija	-0,01	0,00	-0,01	0,03	-0,04
Irska	-0,01	0,02	-0,03	0,04	-0,07
Italija	-0,03	0,09	-0,12	0,02	-0,14
Japonska	-0,03	0,09	-0,12	0,04	-0,16
Koreja	0,01	-0,04	0,06	0,08	-0,03
Luksemburg	-0,01	0,03	-0,04	0,04	-0,08
Mehika	-0,05	0,18	-0,23	0,04	-0,27
Nizozemska	-0,01	-0,04	0,03	0,03	0,00
Nova Zelandija	0,03	-0,10	0,13	0,04	0,09
Norveška	0,02	-0,09	0,11	0,04	0,08
Poljska	0,00	0,07	-0,07	0,01	-0,08
Portugal	-0,04	0,14	-0,18	0,04	-0,22
Rusija	-0,04	0,11	-0,15	0,05	-0,21
Španija	-0,02	0,08	-0,10	0,05	-0,15
Švedska	0,06	-0,20	0,26	0,05	0,21
Švica	0,01	-0,03	0,04	0,04	0,00
Velika Britanija	0,01	-0,12	0,14	0,04	0,10
ZDA	0,02	0,01	0,01	0,04	-0,03

Relativno težavnost veznih nalog na ravni posamezne države je bilo treba primerjati z vrednostjo na mednarodni ravni, ker so bili mednarodni parametri v prvotnem razvrščanju na lestvico uporabljeni kot sidra, kar bo razloženo v naslednjem odstavku. Relativna težavnost veznih nalog na mednarodni ravni ima vrednost 0,04, kar pomeni, da so bile vezne naloge ob uporabi mednarodnega umeritvenega/kalibracijskega vzorca 0,04 točke *logit* lažje od nalog, vključenih le v raziskavo PISA 2000. Nekatere države so ta korak zakomplicirale z odstranitvijo nekaterih nalog zaradi napak v prevodih, kar je vplivalo na relativno težavnost na mednarodni ravni. Zato je bila vrednost slednje za te države prilagojena (glej stolpec 5). Koreja je na primer odstranila štiri bralne naloge, zato relativna težavnost na mednarodni ravni znaša 0,08. Zadnji stolpec primerja vrednost na ravni države s (prilagojeno) vrednostjo na mednarodni ravni in predstavlja našo zadnjo mero *relativne težavnosti veznih nalog*.

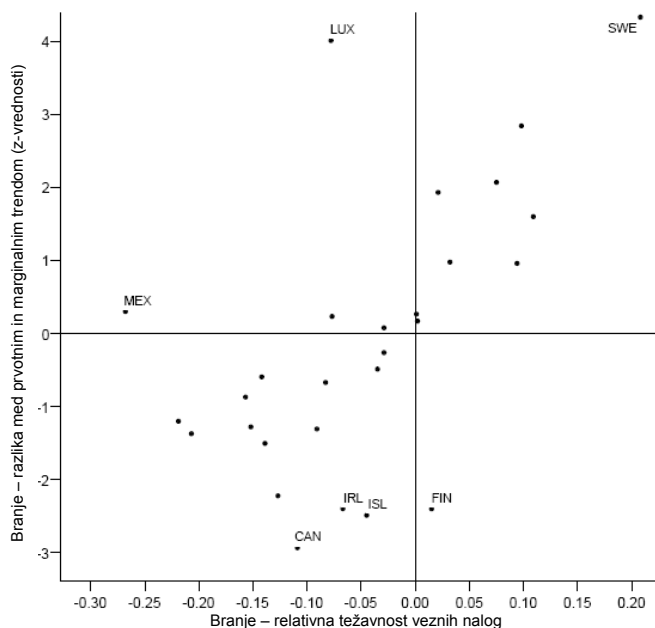
Variiranje relativne težavnosti veznih nalog (zadnji stolpec tabele 4 in tabele 5) je bilo povezano z variiranjem razlike med prvotnimi in marginalnimi trendi, kar ilustrira primer trendov bralne zmožnosti na Švedskem. Kot je bilo že omenjeno, je nacionalno umerjanje pokazalo, da so bile za švedske učence vezne naloge iz

branja za 0,26 točke *logit* lažje od nalog, vključenih le v raziskavo PISA 2000. Ti nacionalni parametri so bili uporabljeni za izračun marginalnega trenda. Pri izračunu prvotnega trenda pa so bili parametri postavk fiksirani na mednarodne vrednosti, kjer so bile vezne naloge le za 0,04 točke *logit* lažje od nalog, vključenih le v raziskavo PISA 2000. Ker so bile vezne naloge edine, ki so bile vključene v raziskavo PISA 2003, so bile zmožnosti švedskih učencev po prvotni razvrstitvi v raziskavi PISA 2003, opravljeni z uporabo mednarodnih parametrov postavk, boljše kot po razvrstitvi na državni ravni za marginalni trend. Zato je bil prvotni trend na Švedskem manj negativen od marginalnega trenda.

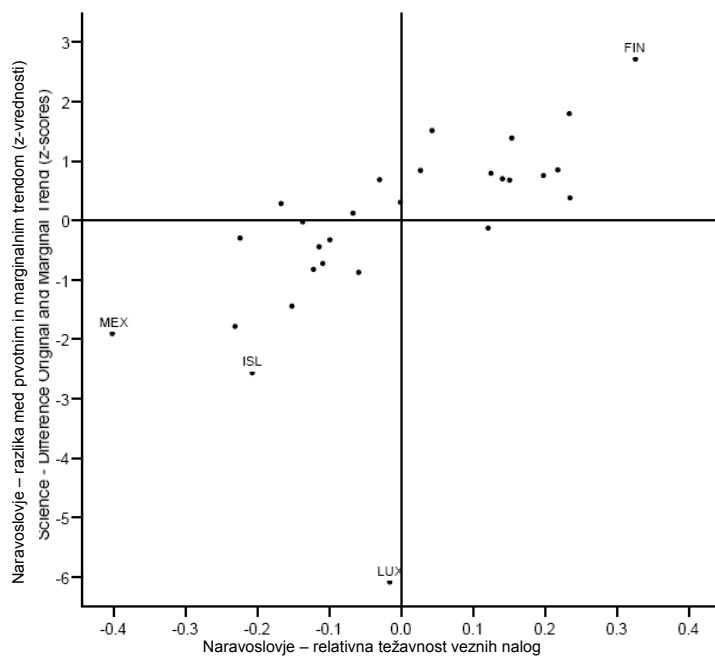
Tako raziskava PISA 2000 kot tudi raziskava PISA 2003 sta na področju naravoslovja ocenjevali tako postavke, vključene samo v eno raziskavo, kot tudi skupne postavke, zaradi česar je izračun relativne težavnosti veznih nalog iz naravoslovja nekoliko kompleksnejši. Četrti stolpec v tabeli 5 predstavlja relativno težavnost veznih nalog iz naravoslovja v raziskavi PISA 2000 (drugi stolpec minus četrti stolpec), peti stolpec pa relativno težavnost veznih nalog iz naravoslovja v raziskavi PISA 2003 (tretji stolpec minus četrti stolpec). Relativna težavnost veznih nalog na ravni posamezne države je razlika med petim in šestim stolpcem in je navedena v sedmem stolpcu. Naslednji stolpec navaja relativno težavnost veznih nalog na mednarodni ravni, prilagojeno tako, da upošteva odstranitev določenih nalog v nekaterih državah. Za mednarodni umeritveni/kalibracijski vzorec so bile vezne naloge 0,10 točke *logit* lažje od nalog, vključenih samo v raziskavo PISA 2000, in 0,02 točke *logit* težje od nalog, vključenih samo v analizo PISA 2003. Rezultat na Norveškem je ravno nasproten, saj so bile vezne naloge 0,07 točke *logit* težje od nalog, vključenih samo v analizo PISA 2000, in 0,04 točke *logit* lažje od nalog, vključenih samo v raziskavo PISA 2003. Tudi tukaj primerjava med vrednostjo na državni in mednarodni ravni daje relativno težavnost veznih nalog in je navedena v zadnjem stolpcu tabele 5.

Učinek relativne težavnosti veznih nalog iz naravoslovja lahko ponazorimo na primeru Danske. Ob uporabi mednarodnih parametrov postavk so bile vezne naloge 0,10 točke *logit* lažje od nalog, vključenih le v raziskavo PISA 2000, kar je bilo zelo podobno parametrom postavk na ravni Danske (0,11 *logit* lažje). V mednarodnem umerjanju raziskave PISA 2003 so učenci na mednarodni ravni ocenili vezne naloge za 0,02 točke *logit* težje od nalog, vključenih samo v raziskavo PISA 2003 ali, povedano drugače, naloge, vključene samo v raziskavo PISA 2003, so bile za 0,02 točke *logit* lažje od veznih nalog. Za danske učence pa so bile naloge, vključene samo v raziskavo PISA 2003, za 0,17 točke *logit* lažje od veznih nalog. Zato je bila z uporabo mednarodnih parametrov postavk raziskave PISA 2003 zmožnost danskih učencev višja kot pri uporabi nacionalnih parametrov postavk v letu 2003. Rezultat je bil bolj pozitiven prvotni trend kot marginalni trend (glej shemo 3). Na shemah 4 in 5 sta predstavljena razsevna grafikona za branje in naravoslovje, ki prikazujeta odvisnost med relativno težavnostjo veznih nalog ter razliko med prvotnimi in marginalnimi trendi.

Shema 4: Razsevni grafikon med relativno težavnostjo veznih bralnih nalog in razliko med prvotnim in marginalnim trendom v bralnih dosežkih



Shema 5: Razsevni grafikon med relativno težavnostjo veznih naravoslovnih nalog in razliko med prvotnim in marginalnim trendom v naravoslovnih dosežkih



Korelacija na področju branja je znašala 0,59, na področju naravoslovja pa 0,58. Luksemburg je na obeh področjih izstopal z ekstremno pozitivnim trendom. Če Luksemburg zanemarimo, bi korelacije znašale 0,67 oziroma 0,82.

Tabela 5: Izračun relativne težavnosti veznih nalog na področju naravoslovja

	Naravoslovje							
	Naloge, vključene le v raziskavo 2000	Naloge, vključene le v raziskavo 2003	Povprečna težavnost veznih nalog	Relativna težavnost veznih nalog v raziskavi 2000	Relativna težavnost veznih nalog v raziskavi 2003	Relativna težavnost veznih nalog na ravni posamezne države	Relativna težavnost (Prilagojena) na mednarodni ravni	Relativna težavnost veznih nalog
International 2000	0,08		-0,03	0,10				
International 2003		0,08	0,10		-0,02		0,13	
Avstralija	0.23	-0.13	-0.04	0.27	-0.10	0.36	0.13	0.23
Avstrija	-0.03	-0.06	0.03	-0.07	-0.09	0.03	0.13	-0.10
Belgija	0.11	-0.13	0.01	0.11	-0.14	0.25	0.13	0.12
Kanada	0.04	0.07	-0.04	0.08	0.11	-0.03	0.13	-0.15
Češka	0.09	-0.01	-0.03	0.12	0.02	0.10	0.13	-0.03
Danska	0.12	-0.16	0.01	0.11	-0.17	0.28	0.13	0.15
Finska	0.13	-0.33	0.07	0.05	-0.40	0.45	0.13	0.33
Francija	0.16	-0.11	-0.02	0.18	-0.09	0.27	0.13	0.14
Nemčija	0.11	-0.25	0.05	0.08	-0.19	0.27	0.12	0.15
Madžarska	-0.03	-0.05	0.03	-0.06	-0.08	0.02	0.13	-0.11
Islandija	-0.05	-0.06	0.04	-0.03	0.13	-0.16	0.05	-0.21
Irska	0.06	-0.07	0.00	0.05	-0.07	0.13	0.13	0.00
Italija	0.17	-0.15	-0.01	0.18	-0.14	0.32	0.13	0.20
Japonska	0.02	0.03	-0.02	0.04	0.05	-0.01	0.13	-0.14
Koreja	-0.04	0.00	0.01	-0.05	-0.01	-0.04	0.13	-0.17
Luksemburg	0.06	-0.05	0.00	0.06	-0.05	0.11	0.13	-0.02
Mehika	-0.03	0.25	-0.08	0.05	0.32	-0.28	0.13	-0.40
Nizozemska	0.10	-0.13	0.01	0.11	-0.06	0.17	0.13	0.04
Nova Zelandija	0.08	-0.07	0.00	0.08	-0.07	0.15	0.13	0.03
Norveška	-0.09	0.01	-0.02	-0.07	0.04	-0.11	0.12	-0.23
Poljska	0.05	0.05	-0.04	0.09	0.09	0.00	0.13	-0.12
Portugal	0.07	0.05	-0.09	0.16	0.14	0.01	0.08	-0.07
Rusija	0.29	-0.09	-0.06	0.30	-0.03	0.33	0.11	0.22
Španija	-0.01	-0.02	0.01	-0.02	-0.03	0.01	0.13	-0.11
Švedska	0.05	-0.02	-0.01	0.06	-0.01	0.07	0.13	-0.06
Švica	0.17	-0.19	0.01	0.17	-0.20	0.36	0.13	0.24
Velika Britanija	0.17	-0.08	-0.03	0.21	-0.05	0.25	0.13	0.12
ZDA	0.07	0.17	-0.09	0.16	0.25	-0.10	0.13	-0.22

Marginalni in pogojni trendi

Če obstajajo razlike med marginalnimi in pogojnimi trendi, so te, kot je bilo že obrazloženo, posledica sprememb v distribuciji vzorca v kategorijah nespremenljivih spremljajočih spremenljivk, ki so vključene v regresijski model. Pregled primerov, v katerih se marginalni in pogojni trend razlikujeta, skladno z našimi pričakovanji vodi do različnih preferenc v posameznih primerih, bodisi ocene marginalnega bodisi pogojnega trenda.

V nadaljevanju pobjže predstavljamo primere, v katerih marginalni in pogojni trend vodita do različnih interpretacij.

AVSTRALIJA

Pogojni trend na področju branja je bil manj negativen od marginalnega trenda.

Čeprav trendi na področju naravoslovja v tabeli 3 izkazujejo različno stopnjo pomembnosti, je iz sheme 3 razvidno, da je bila dejanska razlika precej majhna. Sprememba v Z-vrednosti je bila v veliki meri posledica spremembe v standardni napaki, ne v oceni regresijskega koeficienta – marginalni regresijski koeficient je -0,06 (standardna napaka 0,035), pogojni regresijski koeficient znaša -0,08 (standardna napaka 0,044). Majhna razlika med marginalnimi in pogojnimi trendi pretežno izvira iz povišanja starosti za en mesec (povprečna starost v raziskavi PISA 2000 je bila 188 mesecev, v raziskavi PISA 2003 pa 189 mesecev) in znižanja deleža učencev, ki doma ne govorijo jezika, v katerem je potekalo preverjanje (17 odstotkov v raziskavi PISA 2000 in 9 odstotkov v raziskavi PISA 2003). Če bi morali izbirati med obema trendoma, bi bil pogojni verjetno nekoliko bližje realnemu trendu.

AVSTRIJA

Pogojni trend na področju branja je bil manj negativen od marginalnega trenda.

V raziskavi PISA 2000 je v Avstriji sodelovalo 47 odstotkov dečkov, srednja vrednost HISEI je znašala 50, šest odstotkov otrok pa doma večino časa ni govorilo v jeziku, v katerem je potekalo preverjanje. V raziskavi PISA 2003 so bile te številke naslednje: 50 odstotkov dečkov, srednja vrednost HISEI 47 in 9 odstotkov otrok, ki doma večino časa niso govorili v jeziku, v katerem je potekalo preverjanje. Padec vrednosti HISEI se je zdel glavni razlog bolj negativnega marginalnega trenda, kar kaže, da je realni trend bližje pogojnemu kot marginalnemu trendu.

KANADA

Pogojni trendi na področju naravoslovja in branja so bili bolj pozitivni od marginalnih trendov.

V raziskavi PISA 2003 je pri 7 do 8 odstotkih več učencev kot v raziskavi PISA 2000 manjkala vrednost spremenljivk spola (0 odstotkov v raziskavi PISA 2000, 7 odstotkov v raziskavi PISA 2003), starosti (0 odstotkov v raziskavi PISA 2000, 8 odstotkov v raziskavi PISA 2003) in doma govorjenega jezika (3 odstotki v raziskavi PISA 2000, 11 odstotkov v raziskavi PISA 2003). Odstotek dečkov, povprečna starost in odstotek otrok, ki so jezik, v katerem je potekalo preverjanje, uporabljali tudi doma, je ostal na približno enaki ravni. Posebno močan negativen učinek na rezultate v naravoslovju je imela manjkajoča vrednost doma govorjenega jezika. Zaradi odstranitve nalog je bil pogojni trend zelo podoben

marginalnemu trendu. Ker se porazdelitev veljavnih odgovorov na spremljajoče spremenljivke s časom ni spreminjala in ker ni bilo utemeljenega razloga za neupoštevanje učencev, pri katerih so vrednosti teh spremenljivk manjkale, je bilo zaključiti, da kontrola spremljajočih spremenljivk ni bila pravilna metoda za izračun kanadskih trendov. Marginalni trend se je zdel natančnejši.

ČEŠKA

Pogojni trend na področju naravoslovja je bil manj pozitiven od marginalnega trenda.

Učenci, ki so sodelovali v raziskavi PISA 2003, so bili v poprečju dva meseca starejši od tistih v raziskavi PISA 2000, njihov HISEI pa je bil dve točki višji (na lestvici od 16 do 90). Z odstranitvijo teh regresorjev (in kazalnikov manjkajočih podatkov) je bil pogojni trend enak marginalnemu trendu. Kontrola teh sprememb spremljajočih spremenljivk je verjetno rezultirala v boljši oceni trenda na področju naravoslovja. Spremembe teh spremljajočih spremenljivk so enako učinkovale na trend na področju branja, vendar niti marginalni niti pogojni trend ni bil statistično pomemben.

NEMČIJA

Pogojni trendi so bili bolj pozitivni od marginalnih trendov.

Pri 10 odstotkih učencev, ki so sodelovali v raziskavi PISA 2003, in 3 odstotkih učencev, vključenih v raziskavo PISA 2000, so manjkale vrednosti HISEI. Kazalnik manjkajoče vrednosti HISEI je bil v negativnem razmerju z dosežki v branju in naravoslovju. Z odstranitvijo teh učencev iz analize je prišlo do povečanja dosežkov, tako da so bili pogojni trendi približno enaki marginalnim trendom. Ker ni bilo utemeljenega razloga za odstranitev teh učencev iz analize ali upoštevanje učinka manjkajočih vrednosti, so se marginalni trendi zdeli ustrežnejši od pogojnih.

ISLANDIJA

Pogojni trendi so bili bolj negativni od marginalnih trendov.

Čeprav je bila stopnja pomembnosti obeh trendov enaka, velikost razlike ni bila zanemarljiva (razlika v z-vrednostih je bila 1,72 na področju branja in 2,20 na področju naravoslovja). V letu 2003 pri nobenem učencu ni manjkala vrednost spremenljivk spola in starosti (en odstotek in pet odstotkov v raziskavi PISA 2000). Zaradi neupoštevanja učencev z manjkajočimi vrednostmi teh spremenljivk v raziskavi PISA 2000 je bil pogojni trend približno enak marginalnemu trendu. Ker ni bilo utemeljenega razloga za odstranitev teh učencev iz analize, so bili marginalni trendi boljši pokazatelji od pogojnih trendov.

JAPONSKA

Pogojni trendi so bili bolj negativni od marginalnih trendov.

V raziskavi PISA 2003 je vrednost HISEI manjkala pri 63 odstotkih učencev, v raziskavi PISA 2000 pa samo pri 11 odstotkih. Ker je bila povezava med manjkajočim kazalnikom in dosežki negativna, so bili pogojni trendi nižji od realnega trenda.

NOVA ZELANDIJA

Pogojni trendi so bili bolj negativni od marginalnih trendov.

Učenci, vključeni v raziskavo PISA 2003, so bili v povprečju en mesec starejši od učencev, ki so sodelovali v raziskavi PISA 2000. Odstranitev starosti iz regresijske analize za oceno pogojnih trendov je rezultirala v manj negativnih pogojnih trendih, enakih marginalnim trendom. Zato je sprememba v starosti povzročila razliko med pogojnim in marginalnim trendom in se je kontrola starosti zdela primerna za izračun trendov na Novi Zelandiji. Sprememba je bila prisotna tudi v količini manjkajočih vrednosti za HISEI in doma govorjeni jezik (4 in 5 odstotkov v raziskavi PISA 2000 oziroma 14 in 1 odstotek v raziskavi PISA 2003), vendar sta se ta dva učinka medsebojno izničila.

ZDRUŽENE DRŽAVE AMERIKE

Pogojni trendi so bili bolj negativni od marginalnih trendov.

V primerjavi s podatki raziskave PISA 2000 je bilo v raziskavo PISA 2003 vključenih 4 odstotke več dečkov, 6 odstotkov manj je bilo manjkajočih podatkov o spolu, 5 odstotkov manj manjkajočih podatkov o starosti, učenci so bili v povprečju skoraj dva meseca starejši, vrednost HISEI je bila tri točke višja (na lestvici od 16 do 90), manjkajočih vrednosti HISEI pa je bilo 9 odstotkov manj (glej tabelo 6). Pogojni trendi so se za ZDA zaradi teh sprememb v porazdelitvi spremljajočih spremenljivk zdeli ustrežnejši.

V tabeli 7 so poudarjeni najboljši izmed alternativnih trendov, kjer so bile razlike med marginalnimi in pogojnimi trendi ocenjene kot pomembne.

Tabela 6: Opisne statistike spremljajočih spremenljivk za učence iz ZDA v raziskavah PISA 2000 in PISA 2003

Cikel	Dečki (%)	Manjkajoč spol (%)	Starost (v mesecih)	Manjkajoča starost (%)	HISEI	Manjkajoč HISEI (%)	Jezik	Manjkajoč jezik (%)
P2000	0,46	6	188	5	52	15	10	6
P2003	0,50	0	190	0	55	6	9	4

Razprava

Ta študija prikazuje nekatere zanimive ugotovitve v zvezi z oceno trenda v mednarodnih primerjalnih raziskavah. Trendi so deležni velike pozornosti raziskovalcev, oblikovalcev politik in medijev. V posameznih državah je zanimivo videti, ali se morda dosežki s časom spreminjajo zaradi sprememb v izobraževalnem sistemu ali zaradi metodoloških značilnosti.

Analize v tem prispevku obsegajo skrben in podroben preizkus alternativnih možnosti ocenjevanja trendov in pojasnjevanja razlik med rezultati, dobljenimi s temi metodami, na ravni vsake posamezne države. Rezultati kažejo, da je skupen pristop k ocenjevanju trendov v vseh državah lahko zavajajoč.

Prvič, ugotovljeno je bilo, da je precejšen delež variiranja razlik med dvema skupinama trendov (prvotnim in marginalnim) mogoče pojasniti z obliko medsebojnega vpliva države in postavke. Ena izmed teh skupin trendov je temeljila na mednarodnih vrednostih parametrov postavk, druga pa na nacionalnih parametrih postavk. Relativna povprečna težavnost veznih nalog na določenem področju (v primerjavi s povprečno težavnostjo nalog, vsebovanih le v eni raziskavi) v različnih državah ni stabilna. Uporaba mednarodnih parametrov postavk v nekaterih državah vodi v podcenitev, v drugih pa v precenitev trenda v primerjavi z njihovo oceno trenda na državni ravni. Ocenjevanje trendov s pomočjo uporabe nacionalnih parametrov postavk pa ne reši vseh problemov, ker zmanjša primerljivost rezultatov. Kljub temu lahko to obliko medsebojnega vpliva držav in postavk pojasnimo, ko ocenjujemo trende s pomočjo mednarodnih parametrov, in bo po pričakovanjih izboljšala oceno trendov v posameznih državah.

Drugič, spremembe v porazdelitvi spremljajočih spremenljivk med raziskavami lahko vplivajo na oceno trendov, če so te spremenljivke povezane z dosežki. Ta učinek je lahko bodisi odraz resnične spremembe v populaciji bodisi nečesa drugega, recimo različnega vzorčenja ali spremembe v izvedbi preverjanja. Zato je težko oceniti trende, saj je za zaključek, ali je določeno spremembo v porazdelitvi nespremenljive spremenljivke potrebno prilagoditi ali ne, treba natančno preučiti vzorce v posamezni državi.

Posebnosti v posameznih državah, opisane zgoraj, bi lahko prispevale k večji natančnosti ocenjevanja trendov. Učinek razlik v delovanju postavk v različnih državah na trende in preizkušanje metode za kontrolo tega učinka je pomembno vprašanje za bodoče raziskovanje. Poleg upoštevanja učinka neželenih sprememb spremljajočih spremenljivk bo PISA kmalu zbrala podatke za več kot dve časovni točki, kar bo prav tako zgladilo negotovosti v trendih, ki izvirajo iz vzorčenja in drugih težav.

Tabela 7: Kot Tabela 4, vendar s poudarjenim boljšim izmed dveh alternativnih trendov

	Branje			Naravoslovje		
	prvotni trend	marginalni trend	pogojni trend	prvotni trend	marginalni trend	pogojni trend
Avstralija	0	0	0	0	0	--
Avstrija	-	---	-	---	---	---
Belgija	0	0	0	+++	+++	+++
Kanada	0	0	+++	0	0	++
Češka	0	0	0	+++	++	0
Danska	0	0	0	0	0	0
Finska	--	0	0	+++	0	0
Francija	0	0	0	+++	+++	+++
Nemčija	+++	0	+++	+++	+++	+++
Madžarska	0	0	0	0	0	0
Islandija	---	---	---	0	0	0
Irska	-	0	0	0	0	0
Italija	---	--	---	0	0	0
Japonska	--	0	---	--	--	---
Koreja	+++	+++	+++	--	---	---
Luksemburg	+++	+++	+++	+++	+++	+++
Mehika	---	---	---	---	---	---
Nizozemska	---	---	---	0	0	0
Nova Zelandija	0	--	---	0	-	---
Norveška	--	---	---	---	---	---
Poljska	+++	+++	+++	+	+++	+++
Portugal	0	+	0	++	++	+
Rusija	---	---	---	+++	+++	++
Španija	---	---	---	---	--	---
Švedska	0	---	---	0	0	0
Švica	++	+	0	+++	+++	+++
Velika Britanija	---	---	---	--	---	---
ZDA	0	0	---	0	0	--
Opomba:	stopnja statistične pomembnosti		2003 višje od 2000	2003 višje od 2000		
	$p > 0,10$		0	0		
	$p < 0,10$		+	-		
	$p < 0,05$		++	--		
	$p < 0,01$		+++	---		

Viri

- Adams, R. J., Wu, M. L. in Macaskill, G. (1998). Scaling methodology and procedures for the mathematics and science scales. V: M. O. Martin and D. Kelly (ur.). *TIMSS technical report volume II: Implementation and analysis (primary and middle school years)* (pp. 147-174). Boston: Boston College.
- Adams, R. J. (2002). Scaling PISA cognitive data. V: R. J. Adams and M. Wu (ur.), *Programme for international student assessment: PISA 2000 technical report* (pp. 99-108). Pariz: OECD Publications.
- Adams, R. J. in Carstensen, C. (2002). Scaling outcomes. V: R. J. Adams and M. Wu (ur.), *Programme for international student assessment: PISA 2000 technical report* (pp. 149-162). Pariz: OECD Publications.
- Adams, R. J. and Wu, M. L. (2002). (ur.), *Programme for international student assessment: PISA 2000 technical report* (pp. 99-108). Pariz: OECD Publications.
- Adams, R. J., Wu, M. L. in Carstensen, C. H. (2006). Application of multivariate Rasch models in international large scale educational assessment. V: M. v. Davier and C. H. Carstensen (ur.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 271-280). New York: Springer Verlag.
- Ganzeboom, H. B. G., de Graaf, P. M. in Treiman, D. J. (1992). A standard international socioeconomic index of occupational status. *Social Science Research*, 21, 1-56.
- Hedges, L. V. in Vevea, J. L. (2003). *NAEP validity studies: A study of equating in NAEP (NCES 2003-13)*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.
- Johnson, E. G. in Carlson, J. E. (1992). *NAEP technical report*. Washington, DC: National Center for Educational Statistics.
- Judkins, D. R. (1990). Fay's method of variance estimation. *Journal of Official Statistics*, 3, 223-239.
- Kim, S. H. in Cohen, A. S., (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, 22(2), 131-143.
- Linacre, J. M. (1994). *Many-facet Rasch measurement*. Chicago: MESA Press.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.

- Michaelides, M. P. in Haertel, E. H. (2004). *Sampling of common items: An unrecognized source of error in test equating* (CSE Report 636). Los Angeles: The Regents of the University of California.
- Mislevy, R. J. in Sheehan, K. M. (1987). Marginal estimation procedures. V: A. E. Beaton (Ed.), *The NAEP 1983-1984 technical report*. (pp. 293-360). Princeton, NJ: Educational Testing Service.
- Monseur, C. in Berezner A. (2007). The computation of equating errors in international surveys in education. *Journal of Applied Measurement*, 8, 323-335.
- Mullis, I. V. S. in Martin, M. (1998). Item analysis and review. V: M. O. Martin and D. Kelly (ur.). *TIMSS technical report volume II: Implementation and analysis (Primary and middle school years)* (pp. 111-146). Boston: Boston College.
- OECD (2001). *Knowledge and skills for life: First results from PISA 2000*. Pariz: OECD Publications.
- OECD (2004). *Learning for tomorrow's world: First results from PISA 2003*. Pariz: OECD Publications.
- OECD (2005a). *PISA 2003 data analysis manual—SPSS users*. Pariz: OECD Publications.

Dodatek

A – branje							
Država	Prvotni trend	Marginalni trend			Pogojni trend		
	Z-vrednost	nestandardiziran regresijski koeficient	S.E.	Z-vrednost	nestandardiziran regresijski koeficient	S.E.	Z-vrednost
Avstralija	-0,554	-0,003	0,050	-0,064	-0,011	0,039	-0,284
Avstrija	-1,948	-0,142	0,049	-2,924	-0,071	0,040	-1,781
Belgija	0,481	0,063	0,055	1,154	0,026	0,048	0,546
Kanada	-1,339	0,043	0,027	1,596	0,077	0,023	3,369
Češka	-0,505	0,043	0,053	0,805	-0,028	0,036	-0,786
Danska	1,174	-0,017	0,040	-0,423	-0,051	0,036	-1,413
Finska	-2,049	0,014	0,039	0,354	0,005	0,037	0,136
Francija	-1,465	0,035	0,046	0,760	-0,001	0,036	-0,028
Nemčija	2,593	0,033	0,050	0,664	0,122	0,044	2,813
Madžarska	0,045	0,035	0,055	0,642	0,039	0,115	0,342
Islandija	-6,572	-0,108	0,027	-4,080	-0,155	0,027	-5,796
Irska	-1,647	0,036	0,048	0,752	0,035	0,041	0,874
Italija	-3,933	-0,118	0,048	-2,426	-0,126	0,045	-2,831
Japonska	-2,338	-0,106	0,073	-1,466	-0,279	0,073	-3,826
Koreja	3,842	0,166	0,044	3,767	0,114	0,038	3,013
Luksemburg	18,553	0,381	0,026	14,545	0,300	0,025	11,789
Mehika	-5,154	-0,338	0,062	-5,452	-0,292	0,051	-5,690
Nizozemska	-3,546	-0,193	0,051	-3,809	-0,157	0,040	-3,896
Nova Zelandija	-1,009	-0,086	0,044	-1,966	-0,117	0,041	-2,861
Norveška	-2,155	-0,203	0,048	-4,222	-0,244	0,044	-5,547
Poljska	2,982	0,172	0,062	2,751	0,144	0,055	2,638
Portugal	0,520	0,122	0,071	1,723	0,082	0,059	1,390
Rusija	-3,984	-0,174	0,067	-2,611	-0,203	0,059	-3,413
Španija	-3,900	-0,118	0,045	-2,619	-0,137	0,036	-3,796
Švedska	-0,261	-0,177	0,038	-4,589	-0,164	0,031	-5,308
Švica	1,980	0,115	0,063	1,812	0,085	0,054	1,570
Velika Britanija	-3,506	-0,245	0,039	-6,346	-0,214	0,032	-6,594
ZDA	-1,189	-0,088	0,096	-0,924	-0,212	0,067	-3,173

B – naravoslovje							
Država	Prvotni trend	Marginalni trend			Pogojni trend		
	Z-vrednost	nestandardiziran regresijski koeficient	S.E.	Z-vrednost	nestandardiziran regresijski koeficient	S.E.	Z-vrednost
Avstralija	0,414	-0,061	0,044	-1,384	-0,078	0,035	-2,234
Avstrija	-5,457	-0,239	0,047	-5,131	-0,172	0,038	-4,570
Belgija	3,564	0,185	0,050	3,690	0,152	0,044	3,421
Kanada	-0,879	0,015	0,027	0,562	0,053	0,022	2,359
Češka	3,127	0,108	0,044	2,439	0,020	0,037	0,558
Danska	1,459	0,003	0,040	0,071	-0,031	0,033	-0,927
Finska	4,014	0,040	0,031	1,295	0,028	0,029	0,967
Francija	3,856	0,149	0,047	3,154	0,119	0,040	2,963
Nemčija	4,202	0,172	0,049	3,522	0,243	0,042	5,842
Madžarska	0,399	0,059	0,053	1,121	0,012	0,109	0,106
Islandija	-1,059	0,034	0,023	1,503	-0,017	0,024	-0,695
Irska	-0,556	-0,038	0,044	-0,863	-0,038	0,036	-1,074
Italija	0,341	-0,019	0,046	-0,416	-0,026	0,043	-0,604
Japonska	-2,025	-0,157	0,079	-2,001	-0,294	0,083	-3,551
Koreja	-2,409	-0,135	0,050	-2,695	-0,223	0,048	-4,635
Luksemburg	14,887	0,460	0,022	20,970	0,388	0,023	16,953
Mehika	-5,735	-0,199	0,052	-3,833	-0,160	0,043	-3,754
Nizozemska	-0,094	-0,084	0,052	-1,608	-0,062	0,044	-1,396
Nova Zelandija	-0,883	-0,067	0,039	-1,726	-0,096	0,037	-2,599
Norveška	-4,579	-0,120	0,043	-2,797	-0,157	0,039	-3,989
Poljska	1,823	0,161	0,061	2,645	0,144	0,054	2,676
Portugal	2,505	0,140	0,059	2,380	0,096	0,051	1,897
Rusija	3,588	0,199	0,073	2,735	0,168	0,067	2,512
Španija	-2,924	-0,110	0,044	-2,483	-0,126	0,035	-3,607
Švedska	-0,475	0,015	0,038	0,397	0,027	0,032	0,829
Švica	3,201	0,183	0,065	2,818	0,148	0,056	2,657
Velika Britanija	-2,462	-0,123	0,038	-3,257	-0,113	0,031	-3,696
ZDA	-0,751	-0,041	0,089	-0,456	-0,158	0,063	-2,507

IZRAČUN NAPAK IZENAČEVANJA LESTVIC DOSEŽKOV V MEDNARODNIH RAZISKAVAH NA PODROČJU IZOBRAŽEVANJA

Christian Monseur
Univerza v Liègeu, Belgija
Avstralski inštitut za raziskovanje v izobraževanju, Avstralija

Alla Berezner
Avstralski inštitut za raziskovanje v izobraževanju, Avstralija

Od *Tretje mednarodne raziskave znanja matematike in naravoslovja* IEA je bil eden glavnih ciljev mednarodnih raziskav na področju izobraževanja ugotavljanje trendov v dosežkih. Ta cilj se odraža tudi v poimenovanju dveh trenutnih raziskav IEA: *Mednarodna raziskava trendov v znanju matematike in naravoslovja (TIMSS)* in *Mednarodna raziskava bralne pismenosti (PIRLS)*. Tudi raziskava PISA, ki poteka pod okriljem OECD, je usmerjena v analizo trenda rezultatov skozi čas. Da bi omogočile tovrstno analizo, te študije povezujejo svoje preizkuse znanja s pomočjo skupnih postavk v povezavi z metodami teorije odgovora na postavko.

Raziskave IEA in raziskava PISA se razlikujeta v načinu podajanja napake, povezane s trendi. V raziskavah IEA standardna napaka ocene trenda ne vsebuje negotovosti, povezane s postopkom povezovanja, medtem ko PISA vključuje komponento napake povezovanja v standardno napako ocene trenda. Povedano drugače, PISA implicitno priznava, da so ocene trendov delno odvisne od izbire skupnih postavk, medtem ko raziskave IEA ne upoštevajo tega vira napake.

Posledica neupoštevanja napake povezovanja je podcenjena standardna napaka, kar povečuje stopnjo napake tipa I in vodi k zaključku, da so razlike v dosežkih statistično pomembne, ko v resnici niso. Naraščajoče zanimanje oblikovalcev politik za kazalnike trenda in njihov vpliv na oceno izobraževalnih reform se zdi neskladen s tovrstnim podcenjevanjem standardne napake.

Vseeno pa postopek, ki ga uporablja PISA, vzbuja določene pomisleke o osnovnih predpostavkah za računanje napak pri izenačevanju lestvic dosežkov (equating error).

Po kratkem uvodu bo v pričujočem prispevku predstavljen postopek, po katerem PISA izračunava napako povezovanja, in njegove osnovne predpostavke. V sklepnem delu bo opisana alternativna metoda, ki temelji na tehnikah ponavljanja (replication techniques), in sicer najprej na simuliranem primeru, nato pa z uporabo podatkov raziskave PISA 2000.

Uvod

Zaradi interesa oblikovalcev politik za monitoring izobraževalnih sistemov in ocenjevanje učinkov izobraževalnih reform se pri oblikovanju raziskav na področju izobraževanja vedno bolj poudarjajo indikatorji trenda. Ti namreč oblikovalcem politik povedo, kako se spreminja nivo dosežkov učencev v njihovi državi v primerjavi z ostalimi državami, pa tudi, kako se razlike v rezultatih v posameznih delih države, npr. različni dosežki po spolu, s časom razvijajo. Premik poudarka h kazalnikom trenda predstavlja velik izziv za raziskave dosežkov.

Danes so raziskave TIMSS, PIRLS in PISA reden vir podatkov o trendih dosežkov (glej Beaton et al., 1996a; Beaton et al., 1996b; Mullis et al., 2000; Martin et al., 2000; Mullis et al., 2004; Martin et al., 2004; Martin et al., 2003; OECD, 2004).

Raziskavi IEA in PISA se razlikujeta po načinu ugotavljanja standardne napake trenda dosežkov. Le-ta pri raziskavah IEA temelji na standardnih napakah, povezanih z oceno dveh neodvisnih parametrov populacije (population parameter estimates). Ta standardna napaka torej temelji na dveh variancah vzorčenja in dveh variancah napake merjenja.¹ V raziskavah PISA pa standardna napaka ocene trenda vsebuje še tretjo komponento, imenovano napaka povezovanja, ki bo opisana v nadaljevanju (za obširnejši opis gl. tehnično poročilo raziskave PISA 2003, OECD, 2005).

Ta tretja komponenta standardne napake odraža razpršenost ocen trenda, povezano z izbiro skupnih postavk. Upoštevajoč Rascheve predpostavke sestavljanja lestvic (Rasch scaling assumptions) bi bil rezultat izenačevanja dosežkov enak ne glede na to, katere skupne postavke bi uporabili, ker bi bile značilnosti posamezne postavke v celoti zajete v parametrih postavk, ocenjenih po Raschevem modelu. Vendar se napake v specifikaciji modela vedno dogajajo: majhne spremembe v nalogah, učinki umestitve naloge in kurikula lahko vplivajo na obnašanje postavke v modelu. Tako bi drugačni nizi skupnih postavk ustvarili drugačne transformacije izenačevanja (equating transformations), celo pri zelo velikem vzorcu preizkušancev.

Kot navajata Sheehan in Mislevy (1988, str. 2), je »ignoriranje negotovosti, povezane s postopkom povezovanja, postalo stalnica pri oblikovanju zaključkov, ki zajemajo postavke iz različnih podtestov, kar se pojavlja npr. pri merjenju sprememb.« V študiji z uporabo podatkov NAEP sta Sheehan in Mislevy (1998) za merjenje negotovosti povezave uporabila metodo *jackknife* (Wolter, 1985; Rust in Rao, 1996). Avtorja sta zaključila: »Medtem ko se je standardna napaka ocene sprememb pri posamezniku povečala le za približno dva odstotka, je povečanje standardne napake pri skupinah okoli 200-odstotno [...] Komponenta napake, ki

¹ Ker so dosežki učencev navajani s pomočjo pričakovanih vrednosti, ta varianca merjenja ustreza varianci vračunavanja (imputation variance).

izvira iz povezovanja, predstavlja v povprečju približno 90 odstotkov skupne nepojasnjene variance.« (Sheehan in Mislevy, 1988, str. 18–19). Podobno sta s primerjavo analitičnih rešitev po metodi *bootstrap* ugotovila tudi Michaelides in Haertel (2004, str. 24–25): »Negotovost glede natančnosti srednje vrednosti zaradi napak vzorčenja in merjenja [...] je zelo majhna pri velikih vzorcih preizkušancev, in ocene povprečnih rezultatov so pri velikih vzorcih ponavadi zelo natančne. Zato so napake izenačevanja dosežkov, posebej napaka zaradi izbiranja skupnih postavk, relativno velike v primerjavi s standardno napako srednje vrednosti. Napaka vzorčenja skupnih postavk predstavlja 82,6 odstotka skupne variance, kar je veliko več od ostalih virov napak, ki so povezani z velikostjo vzorca.«

Omenjena tretja komponenta standardne napake utegne biti, kot navajajo ti avtorji, precejšnja. Neupoštevanje napake povezovanja bo namreč vodilo v precejšnjo podcenitev standardne napake. Sheehan in Mislevy (1988, str. 19) v svoji študiji ugotavljata, da »upad povprečnih bralnih sposobnosti pri devetletnikih znaša približno tri standardne napake, če negotovost postopka povezovanja ni upoštevana, in samo eno standardno napako, če je.« Povedano drugače, izboljšanje ali upad dosežkov bo označen kot pomemben, ko v resnici ne bo, kar bo vodilo k povečanju napake tipa I.

TIMSS je trenutno edina raziskava, ki je bila izvedena več kot dvakrat (tj. v letih 1995, 1999 in 2003). Raziskava TIMSS/REPEAT (Martin et al., 2000; Mullis et al., 2000) je pri učencih osmega razreda ugotovila štiri pomembne spremembe v dosežkih pri matematiki: v Latviji (+17), Kanadi (+10), na Cipru (+9) in Češkem (-26). Žal Kanada in Češka nista sodelovali v raziskavi TIMSS 2003. Izmed ostalih držav s pomembno spremembo med leti 1995 in 1999 je bila ocena srednje vrednosti v Latviji enaka kot leta 1999, medtem ko se je na Cipru občutno znižala, in sicer za 17 točk (Mullis et al., 2004). Povedano drugače, IEA je najprej ugotovila pomembno povečanje, štiri leta kasneje pa pomembno znižanje dosežkov. Povedati je treba, da je izmed 20 držav, ki so sodelovale v vseh treh raziskavah, Ciper edina s takšnim vzorcem rezultatov. Ta je lahko odraz realnega razvoja dosežkov učencev, posledica neznanih kontekstualnih učinkov med zbiranjem podatkov (npr. stavke), lahko pa tudi izbire sidrskih postavk. Neupoštevanje napake povezovanja povečuje možnost nastanka takšnega nepojasnjene vzorca, kar bi lahko dolgoročno ogrozilo verodostojnost kazalnikov trenda za oblikovalce politik.

Zdaj ko smo predstavili pomembnost napake povezovanja, se bomo v prispevku ukvarjali z metodo njenega izračunavanja, ki jo uporablja OECD, in predlagali alternativno metodo. Prispevek je sestavljen iz štirih glavnih delov:

- (i) opisali bomo metodo OECD/PISA 2003 za izračun napake povezovanja;
- (ii) našli in obdelali bomo osnovne predpostavke programa OECD/PISA 2003 za izračun te napake, tj. slučajno (stochastic) neodvisnost postavk, populacijo

postavk (item population), kategorije odgovorov pri posameznih postavkah in uniformnost napake povezovanja;

- (iii) predstavili bomo alternativno metodo, temelječo na ponavljanju, in preizkusili njeno učinkovitost na simuliranem primeru;
- (iv) to metodo ponavljanja bomo aplicirali na podatke preverjanja bralne pismenosti v raziskavi PISA 2003.

Postopek sidranja

To poglavje opisuje postopek, ki se pri mednarodnih raziskavah na področju izobraževanja običajno uporablja za umeščanje podatkov o kognitivnih zmožnostih učencev na obstoječo lestvico.

Bralna pismenost je bila glavno področje v raziskavi PISA 2000, medtem ko je bila v raziskavi PISA 2003 stransko področje. Da bi povezali obe študiji, je bilo 28 od skupno 138 postavk oziroma vprašanj, uporabljenih leta 2000, vključenih tudi v raziskavo, izvedeno leta 2003. Ta vezne postavke (*linking items*) so bile uporabljene z namenom umestiti podatke o bralni pismenosti iz leta 2003 na lestvico bralne pismenosti iz leta 2000.

Koraki za umestitev podatkov o bralni pismenosti iz raziskave PISA 2003 na kombinirano lestvico bralne pismenosti iz leta 2000 so bili (OECD, 2005):

1. Umerjanje podatkov o bralni pismenosti iz leta 2003 z namenom pridobiti parametre postavk iz raziskave PISA 2003, tj. relativno težavnost nalog po Raschevi lestvici;
2. Določitev pričakovanih vrednosti (*plausible values*) bralne pismenosti po podatkih raziskave PISA 2003 na podlagi omenjenih parametrov postavk (za več informacij gl. npr. Wu, 2005);
3. Določitev pričakovanih vrednosti bralne pismenosti po podatkih raziskave PISA 2000 na podlagi parametrov veznih postavk iz prvega koraka. Sedaj sta za raziskavo PISA 2000 na voljo dve skupini pričakovanih vrednosti: (i) prvotna skupina pričakovanih vrednosti, vključena v podatkovno bazo PISA 2000 in (ii) skupina pričakovanih vrednosti, temelječa na parametrih postavk iz raziskave PISA 2003. Žal se srednja vrednost in standardni odklon nove skupine pričakovanih vrednosti nekoliko razlikujeta od prvotnih pričakovanih vrednosti iz raziskave PISA 2000. Te razlike odražajo spremembe v težavnosti veznih nalog med letoma 2000 in 2003.
4. Ocena linearne transformacije, ki zagotovi, da srednja vrednost nove skupine pričakovanih vrednosti po podatkih raziskave PISA 2000 znaša 500, standardni odklon pa 100. To linearno transformacijo lahko zapišemo kot:

$$PV_{cal_2000} = a + bPV_{cal_2003} ,$$

$$z\ b = \frac{S_{cal_2000}}{S_{cal_2003}} ,$$

$$in\ a = m_{cal_2000} - bm_{cal_2003} .$$

Na zgornjem primeru je $b = (100 / 110) = 0,909$ in

$$a = (500 - (0,909 \cdot 505)) = 40,955 .$$

5. Aplikacija te linearne transformacije na pričakovane vrednosti iz raziskave PISA 2003, kar zagotovi primerljivost rezultatov učencev v letu 2003 z njihovimi rezultati v letu 2000.

Kot je bilo že omenjeno, bi bila ocena linearne transformacije pri drugačnem izboru veznih postavk drugačna, zato obstaja določena negotovost transformacije, ki je odvisna od vzorčenja veznih postavk.

Izračun napake povezovanja za kazalnike trenda bralnih sposobnosti v raziskavi PISA 2000–2003

Za vsako vezno postavko imamo dve oceni parametra postavke, ki sta sedaj primerljivi: (i) oceno parametra postavke iz leta 2000 in (ii) oceno parametra postavke iz leta 2003. Obe skupini parametrov postavk se centrirata in nato primerjata. Pri nekaterih veznih postavkah se njihova relativna težavnost poveča, pri drugih pa se zmanjša. To pomeni, da se nekatere vezne naloge v letu 2003 zdijo relativno težje kot v letu 2000 in obratno.

V raziskavi PISA 2003 je bila napaka povezovanja izračunana po naslednji formuli:

$$s_{link} = \sqrt{\frac{s^2}{n}} , \quad (1)$$

kjer je s^2 varianca razlik med parametri postavk, n pa število uporabljenih postavk. Ta formula je skladna z zaključki Michaelidesa in Haertela (2004, str. 1–2), da »napaka zaradi vzorčenja skupnih postavk ni odvisna od velikosti vzorca preizkušancev, ampak od števila skupnih postavk«.

Predpostavimo, da se parametri postavk po umerjanju iz leta 2003 popolnoma ujemajo s parametri postavk po umerjanju iz leta 2000. Povedano drugače, relativna težavnost veznih postavk je nespremenjena. V tem primeru bi bile vse razlike med relativno težavnostjo v letih 2000 in 2003 enake nič, zato bi bila tudi napaka povezovanja enaka nič.

Z večanjem razlik v parametrih postavk se povečuje varianca teh razlik in posledično tudi napaka povezovanja. Logično je, da je negotovost glede trenda sorazmerna s spremembami parametrov postavk.

Prav tako je negotovost glede trenda obratno sorazmerna s številom veznih postavk. S teoretičnega vidika je potrebna ena sama postavka. Če trend temelji na samo eni postavki, bo negotovost verjetno velika, a neocenljiva. S povečevanjem števila veznih postavk se negotovost zmanjšuje.

Tabela 1 prikazuje centrirane ocene parametrov veznih postavk v raziskavah bralne pismenosti PISA 2000 in PISA 2003 ter razliko med obema ocenama za vsako postavko.

Varianca razlike je enaka 0,047486. Napaka povezovanja je torej enaka

$$s_{link} = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{0,047486}{28}} = 0,041 \quad (2)$$

Na PISA lestvici bralne pismenosti, ki ima srednjo vrednost 500 in standardni odklon 100, vrednost 0,041 ustreza 3,75 točke. Več informacij o napaki povezovanja je dostopnih v tehničnem poročilu raziskave PISA 2003 (OECD, 2005).

Tabela 1: Ocene parametrov nalog oz. vprašanj bralne pismenosti v letih 2000 in 2003

Ime naloge oz. vprašanja	Centrirana ocena parametra v letu 2003	Centrirana ocena parametra v letu 2000	Razlika
R055Q01	-1,28	-1,347	-0,072
R055Q02	0,63	0,526	-0,101
R055Q03	0,27	0,097	-0,175
R055Q05	-0,69	-0,847	-0,154
R067Q01	-2,08	-1,696	0,388
R067Q04	0,25	0,546	0,292
R067Q05	-0,18	0,212	0,394
R102Q04A	1,53	1,236	-0,290
R102Q05	0,87	0,935	0,067
R102Q07	-1,42	-1,536	-0,116
R104Q01	-1,47	-1,205	0,268
R104Q02	1,44	1,135	-0,306
R104Q05	2,17	1,905	-0,267
R111Q01	-0,19	-0,023	0,164
R111Q02B	1,54	1,395	-0,147
R111Q06B	0,89	0,838	-0,051
R219Q01T	-0,59	-0,520	0,069
R219Q01E	0,10	0,308	0,210
R219Q02	-1,13	-0,887	0,243
R220Q01	0,86	0,815	-0,041
R220Q02B	-0,14	-0,114	0,027
R220Q04	-0,10	0,193	0,297
R220Q05	-1,39	-1,569	-0,184
R220Q06	-0,34	-0,142	0,196
R227Q01	0,40	0,226	-0,170
R227Q02T	0,16	0,075	-0,086
R227Q03	0,46	0,325	-0,132
R227Q06	-0,56	-0,886	-0,327

Predpostavke izračuna napake povezovanja

Formula (1) je podobna formuli, ki se uporablja za izračun standardne napake srednje vrednosti enostavnega vzorca, izbranega iz neskončne populacije (Cochran, 1977). Vendar je treba v zvezi z uporabo te formule v kontekstu napak povezovanja opozoriti na dve pomembni vprašanji: slučajno neodvisnost izbranih enot in predpostavko neskončnosti populacije.

Poleg tega uporaba napake povezovanja vzbuja še tri potencialne pomisleke: učinek nalog oz. vprašanj, ki sledijo nalogi z uvodnim stimulusom, pri katerih so možni delno pravilni odgovori; predpostavka, da je napaka povezovanja konstantna v vseh sodelujočih državah; in napake v specifikaciji v posamezni državi.

Slučajna neodvisnost

Formula (1) je pravilna, če predpostavimo, da so ocene razlik v parametrih vsake postavke med seboj neodvisne. Vendar v praksi ta predpostavka redko drži. Eden izmed razlogov za to je, da so v raziskavah PISA 2000 in PISA 2003 naloge razvrščene v skupine (Adams in Wu, 2002; OECD, 2005). V raziskavi PISA 2000 je bilo na primer 138 nalog razvrščenih v 37 skupin, od katerih je vsaka temeljila na skupnem uvodnem besedilu; vezne naloge je bilo treba izbrati skupaj z ostalimi nalogami, ki so sestavljale posamezno skupino.

Po Kishevi formuli (1965) je varianca vzorčnih ocen srednje vrednosti vzorca, razvrščenega v skupine, enaka:

$$s_{link}^2 = \frac{s_B^2}{n_B} + \frac{s_W^2}{n_B n_W}, \quad (3)$$

za neskončne populacije, kjer je s_B^2 varianca med skupinami, s_W^2 varianca znotraj skupine, n_B število skupin v vzorcu in n_W število postavk v posamezni skupini, za katero se predpostavlja, da je v vseh skupinah enako.

Varianco razlike v parametrih postavk med letoma 2000 in 2003 med skupinami in znotraj skupine je mogoče oceniti s pomočjo ANOVA analize.

Tabela 2: ANOVA tabela razlik v nalogah bralne pismenosti

	SS	DF	MS
Med skupinami	0,7532	7	0,1076
Znotraj skupine	0,5288	20	0,0264
Skupaj	1,2820	27	0,0475

Skupno povprečje kvadratov (Total mean square) je nepristrana ocena skupne variance, saj je srednja napaka kvadratov (ali srednja vrednost kvadratov znotraj vzorca) nepristrana ocena nepojasnjene variance.

Nepristranka ocena variance med skupinami se lahko izračuna na naslednji način:

$$\begin{aligned}s_B^2 &= \frac{MS_B - MS_W}{\bar{n}} \\&= \frac{0,1076 - 0,0264}{(28 / 8)} \\&= \frac{0,0812}{3,5} = 0,0232 .\end{aligned}$$

Kot je prikazano v tabeli 2, je varianca premikov parametrov postavk med skupinami enako velika kot varianca znotraj skupine.

V skladu s formulo (3) je ocena napake povezovanja:

$$\begin{aligned}s_{link} &= \frac{s_B^2}{n_B} + \frac{s_W^2}{n_B n_W} \\&= \sqrt{\frac{0,0232}{8} + \frac{0,0264}{28}} = 0,062 ,\end{aligned}$$

kar nakazuje, da je napaka povezovanja, vsebovana v tehničnem poročilu raziskave PISA 2003 (OECD, 2005) podcenjena zaradi združevanja nalog v skupine.

Populacija postavk

Vse zgornje formule temeljijo na predpostavki, da se trendi ocenjujejo na podlagi vzorca postavk, izbranega iz neskončno velike banke nalog. Sedaj pa preučimo napako povezovanja v primeru, če se vse postavke uporabijo kot skupne. Leta 2001 je IEA izvedla ponovitevno raziskavo bralne pismenosti (Martin et al., 2003) in devet držav, ki so sodelovale v raziskavi bralne pismenosti IEA leta 1991, je učencem ponovno predložilo v reševanje isti preizkus, ne da bi dodali ali odstranili katerokoli nalogo. Podobno so bile tudi v raziskavi PISA 2006 uporabljene vse bralne naloge iz raziskave PISA 2003. Ali je v teh okoliščinah sploh smiselno upoštevati negotovost, ki izvira iz vzorčenja skupnih postavk, če se pravzaprav vse naloge uporabijo kot skupne postavke? Povedano drugače, bi morali populacijo, iz katere izberemo skupne postavke, šteti za končno obstoječo populacijo ali za enostavni naključni vzorec iz hipotetične neskončne populacije?

S formulo (1) dobimo oceno variance vzorčne srednje vrednosti ob predpostavki enostavnega naključnega vzorca, izbranega iz neskončne populacije. Če štejemo banko nalog za končno populacijo, je treba formulo (1) zamenjati z naslednjo:

$$s_{(m)}^2 = \frac{s^2}{n} \left(\frac{N - n}{N - 1} \right) \quad (4)$$

kjer je N velikost banke nalog, n pa število skupnih postavk. V primeru študije bralne pismenosti IEA, kjer so bile v letu 2001 kot skupne postavke uporabljene vse naloge iz leta 1991, je $N = n$, zato bi bila, ker bi bil »koeficient izčrpnosti« (exhaustivity coefficient)

$$\left(\frac{N - n}{N - 1} \right)$$

enak nič, tudi napaka povezovanja enaka nič. V raziskavi PISA 2003 bi bil koeficient izčrpnosti enak

$$\left(\frac{138 - 28}{138 - 1} \right) = 0,80292.$$

To pomeni, da bi uporaba te korekcije končne populacije zmanjšala standardno napako za faktor 0,896.

Kot navajata Michaelides in Haertel (2004, str. 6), so »skupne postavke izbrane iz hipotetične neskončne banke možnih nalog.« Temu stališču pritrdjujejo tudi Cronbach, Linn, Brennan in Haertel (1997). Rezultat testa temelji na dosežku preizkušanca pri določenem testnem vprašalniku, sestavljenem iz določenih nalog. Najbolj zanimivo pri tem ni, kako dobro je preizkušane rešil določene naloge ob določenih priložnostih, ampak sklep, kako dobro bi lahko glede na svoj trenutni rezultat rešil naloge, ki bi zahtevale uporabo enakih veščin in znanja.

Tako lahko štejemo, da so skupne postavke izbrane iz hipotetične neskončne populacije, zato ni treba korekcij zaradi deleža veznih nalog. Če upoštevamo ta argument, bi morali, ko primerjamo bralne dosežke raziskav PISA 2003 in 2006 ali rezultate raziskave bralne pismenosti IEA iz leta 1991 s tistimi iz leta 2001, vključiti napako povezovanja v standardno napako ocene trenda.

Politomne in dihotomne naloge

Šest od 28 skupnih postavk na področju bralne pismenosti v raziskavah PISA 2000 in 2003 je politomnih. Iz tabele 1 je razvidno, da napaka povezovanja temelji na varianci razlik v oceni težavnosti posameznih postavk. Napake povezovanja v raziskavah PISA ne vključujejo morebitnih sprememb v parametrih posamezne *kategorije* postavk. Poleg tega je v primeru predstavljanja politomnih

postavk z njihovo splošno težavnostjo v napaki povezovanja podcenjen njihov prispevek k transformaciji izenačevanja.

Uniformnost napake povezovanja

Monseur, Sibberns in Hastedt (2007) so s ponovno analizo ponovitvene raziskave bralne pismenosti IEA ugotovili, da so se napake izenačevanja dosežkov od države do države precej razlikovale. Poleg tega so opazili visoko korelacijo med absolutno vrednostjo trendov in velikostjo napake povezovanja, vendar ti koeficienti korelacije temeljijo na rezultatih le osmih držav.

Različnost rezultatov nakazuje, da bi bilo treba napako povezovanja izračunavati v vsaki državi, in ne le na mednarodni ravni, kot je bilo v raziskavi PISA 2003.

Napake v specifikaciji na ravni posamezne države

Kot v vsaki mednarodni študiji se tudi v raziskavi PISA 2003 sprememba v parametrih postavk med dvema zbirkama podatkov razlikuje od države do države. Zato se lahko učinek odstranitve določene skupine postavk (tj. določenega števila vsebinsko povezanih postavk) na oceno srednje vrednosti v letu 2003 med državami razlikuje.

Metoda ponavljanja *jackknife* za izračun napake povezovanja: simulirana študija

Rust in Rao sta zapisala (1996, str. 285), da se »metode ponavljanja pogosto označujejo kot tehnike samovzorčenja (resampling). Skupni princip teh metod je uporaba zmogljivih računalniških postopkov, s katerimi premagajo težave, ki se pojavljajo pri uporabi analitičnih rešitev problema. Povedano na kratko, pristop ponavljanja pomeni, da ocenimo varianco določenega parametra populacije tako, da izračunamo ta parameter z uporabo velikega števila različnih podvzorcev (ali različnih uteži pri vzorčenju). Variabilnost dobljenih ocen je podlaga za oceno napake vzorčenja pri prvotnem, popolnem vzorcu.«

Uporaba tehnik samovzorčenja v kontekstu napake povezovanja pomeni, da se izračuna ocena trenda v dosežkih pri podtestih tako, da se vsakič odstrani ena postavka ali niz postavk, nato pa se na podlagi variabilnosti teh ocen trenda oceni napaka povezovanja. Metoda *jackknife* v vsaki ponovitvi zahteva odstranitev ene postavke, če gre za test s samimi samostojnimi nalogami, ali ene skupine postavk v primeru, ko so vprašanja združena v skupine, kot v preizkusu PISA. Če na primer test sestoji iz 30 neodvisnih nalog, bo metoda *jackknife* obsegala 30 ponovitev, od katerih bo vsaka temeljila na 29 nalogah. Če pa je test sestavljen iz 10 nalog z uvodnim besedilom (kot je na primer izhodiščno besedilo pri testu bralnega razumevanja) in se na vsako uvodno besedilo nanaša več vprašanj, bo

metoda *jackknife* obsegala 10 ponovitev, od katerih bo vsaka temeljila na 9 besedilih in vprašanjih, povezanih z njimi.

Da bi ocenili natančnost postopka *jackknife* za oceno napake povezovanja, je bilo izvedenih več simulacij.

V izogib drugim vplivom na varianco vzorčnih ocen je bila v simulaciji dvakrat preizkušena skupina 10.000 učencev, pri čemer preizkusa označujemo kot T1 in T2. Zmožnosti teh učencev, ki jih želimo izmeriti, so normalno porazdeljene, s srednjo vrednostjo nič in standardnim odklonom ena.

V preizkusu T1 je bilo učencem predloženih 50 nalog s težavnostjo, naključno izbrano iz $N(0,1)$. Verjetnost, da bo vsak učenec uspešno rešil vsako izmed 50 nalog, se v tem primeru izračuna po Raschevem modelu kot:

$$P(X_{ij} = 1 | b_i, d_j) = \frac{\exp(b_i - d_j)}{1 + \exp(b_i - d_j)}.$$

Ta verjetnost se nato primerja z uniformno porazdelitvijo. Če je verjetnost, da bo učenec nalogo pravilno rešil, večja od vrednosti uniformne porazdelitve, se šteje, da je bil učenec pri nalogi uspešen, v nasprotnem primeru pa, da je bil neuspešen. Podatkovna matrica se nato obdela v aplikaciji ConQuest Avstralskega inštituta za raziskovanje v izobraževanju (Wu, Adams in Wilson, 1997), ki izračuna pričakovane vrednosti.

V preizkusu T2 pa je 50 nalog razporejenih v 10 skupin: naloge od 1 do 5 sestavljajo prvo skupino, naloge od 6 do 10 drugo skupino itd. Nato se uvedejo spremembe v dejanski težavnosti prvotnih nalog, in sicer najprej na nivoju skupine oz. naloge z vsemi vprašanji, ki ji sledijo, nato pa še na nivoju nalog oz. vprašanj, ki sledijo uvodnemu besedilu. Skupen premik težavnosti je porazdeljen normalno, in sicer s srednjo vrednostjo nič in standardnim odklonom 0,5. Nato so bile izvedene tri razčlenbe skupne variance na nivo skupine in posameznih nalog:

- (i) 25 odstotkov na nivo skupine in 75 odstotkov na nivo naloge;
- (ii) 50 odstotkov na nivo skupine in 50 odstotkov na nivo naloge; in
- (iii) 75 odstotkov na nivo skupine in 25 odstotkov na nivo naloge.

Za oceno trenda so potrebne:

- (i) lestvica populacije v preizkusu T1 z nalogami v preizkusu T1, tj. pred spremembo težavnosti nalog;
- (ii) lestvica populacije v preizkusu T2 z nalogami v preizkusu T2, tj. po spremembi težavnosti nalog;
- (iii) lestvica populacije v preizkusu T1 s težavnostjo nalog, kot je bila izračunana v preizkusu T2.

Za vsako razčlenbo variance je bilo izdelanih 100 možnih preizkusov, tj. parametrom originalnih nalog je bilo dodanih 100 premikov. Teh 100 ocen trenda bo kasneje uporabljenih za empirično določitev porazdelitve napake povezovanja.

Če želimo oceniti natančnost predvidene napake povezovanja po metodi *jackknife*, moramo za vsak možen preizkus oceniti napako povezovanja. Če je metoda *jackknife* primerna za ocenjevanje napake povezovanja, bi morale biti povprečje ocen napake povezovanja po metodi *jackknife* identično standardni porazdelitvi empirično ugotovljene porazdelitve napake povezovanja.

Za vsak možen preizkus je treba ustvariti deset podpreizkusov, ki vsebujejo po devet skupin nalog, in za vsakega od podpreizkusov izvesti postopek povezovanja, opisan zgoraj. Tako dobimo za vsak preizkus oceno trenda na podlagi celotnega preizkusa in deset ocen trenda na podlagi podpreizkusov.

Ocena napake povezovanja se izračuna:

$$s_{(q)} = \sqrt{\frac{(G-1)}{G} \sum_{i=1}^G (q_i - q)^2}, \quad (5)$$

kjer je q ocena trenda celotnega izbora nalog, q_i ocena trenda enega izmed podpreizkusov in G število ponavljanj, v simulaciji torej deset.

V tabeli 3 so zbrani rezultati teh simulacij za vsako razčlenbo variance.

Tabela 3: Ocene napak povezovanja po analitični, empirični in metodi jackknife

% zamenjave nalog na nivoju skupine	% zamenjav nalog na nivoju naloge	št. simulacij	matematična napaka povezovanja	ugotovljena napaka povezovanja	povprečje ocen napake povezovanja po metodi jackknife
25	75	100	10,00	8,98 (0,63)	9,44 (0,20)
50	50	100	12,25	11,85 (0,84)	11,63 (0,32)
75	25	100	14,14	13,50 (1,00)	13,79 (0,33)

Pričakovana napaka povezovanja za tretjo razčlenbo variance, tj. 75 odstotkov na nivoju skupine nalog in 25 odstotkov na nivoju naloge, je enaka:

$$\begin{aligned}
 s_{link} &= \sqrt{\frac{s_B^2}{n_B} + \frac{s_W^2}{n_B n_W}} 100 \\
 &= \sqrt{\frac{(0,75)(0,50^2)}{10} + \frac{(0,25)(0,50^2)}{50}} 100 = 14,14.
 \end{aligned}$$

Ugotovljena napaka povezovanja ustreza standardnemu odklonu 100 ocen trenda, izračunanih na celotnem preizkusu. Zadnji stolpec ustreza srednji vrednosti 100 standardnih napak po metodi *jackknife*.

Kot je prikazano v tabeli 3, so ocene napake povezovanja po metodi *jackknife* zelo blizu ocenam ugotovljene napake povezovanja in matematičnim rešitvam po formuli (3). Vendar je treba opozoriti, da se ocena po metodi *jackknife* pomembno razlikuje od analitične rešitve za prvo razčlenbo variance (25 odstotkov na nivo skupine nalog in 75 odstotkov na nivo naloge), blizu pomembni razliki pa je tudi pri drugi razčlenbi variance (50 odstotkov na nivo skupine nalog in 50 odstotkov na nivo naloge). V obeh primerih so ocene po metodi *jackknife* nižje od analitične rešitve.

Za analizo natančnosti ocen napake povezovanja po metodi *jackknife*, ki je odvisna od števila postavk in kombiniranega standardnega odklona sprememb parametrov postavk, bi bilo treba izvesti nadaljnje simulacije. Vendar ti preliminarni rezultati potrjujejo, da lahko metodo *jackknife* štejemo za obetavno metodo izračuna napake povezovanja.

Uporaba metode *jackknife* za izračun napake povezovanja: PISA 2000 in PISA 2003

Kot je bilo že omenjeno, so metode ponavljanja posebej zanimive takrat, ko ni na voljo analitične rešitve ali je ta zelo kompleksna. Spomniti je treba, da je bila simulacija izvedena samo na enem nizu podatkov, zato ocene parametrov nalog povsem ustrezajo podatkom. V mednarodnih raziskavah na področju izobraževanja pa umerjanje nalog temelji na mednarodnem vzorcu z enakim številom meritev v vsaki državi. Nato se za vsako državo postavijo ocene uspešnosti učencev na osnovi celotnega vzorca posamezne države. Umerjanje nalog na mednarodni ravni morda ni najprimernejša metoda, gledano z vidika posamezne države, saj lahko nacionalne in kulturne razlike v kurikulumih vplivajo na relativno težavnost posameznih nalog. Razlike v ocenah parametrov nalog na mednarodni in državni ravni oziroma medsebojni vpliv naloge in države so dodaten vir napake, ki vpliva na ocene trenda. Na primer, če pri določeni nalogi obstaja velika razlika med oceno težavnosti na mednarodni in državni ravni, bo imela odstranitev te naloge za tisto državo večji vpliv kot odstranitev naloge, kjer ni take razlike. Uporaba metode *jackknife* za oceno napake povezovanja bo zato odražala napake v specifikaciji na ravni posamezne države in negotovost postopka povezovanja na mednarodni ravni.

Metoda *jackknife* za oceno napake povezovanja je bila uporabljena v podatkovnih bazah PISA. Izvedeni so bili enaki koraki, kot so bili že opisani. Povedano na kratko, metoda sestoji iz ponavljanja korakov, opisanih zgoraj, na osmih podskupinah po sedem bralnih enot. Prva ponovitev temelji na enotah dve do osem, druga na enoti ena in enotah tri do osem itd.

Tabela 1 v dodatku in shema 1 prikazujeta spremembo v oceni srednje vrednosti za posamezno državo na kombinirani lestvici bralne pismenosti raziskave PISA 2003 v primerjavi z ocenami srednje vrednosti, izračunanih na celotni skupini

sidrnih postavk na področju bralne pismenosti. Kot primer lahko navedemo, da bi odstranitev prve enote iz skupine sidrnih postavk imela za posledico, da bi se ocena srednje vrednosti v Avstraliji povečala za 1,82 točke, odstranitev šeste enote pa bi avstralski rezultat na lestvici PISA znižala za 6,17 točke.

Na ravni OECD bi odstranitev prve enote povečala oceno srednje vrednosti za 4,14 točke, odstranitev druge enote pa bi jo znižala za 3,96 točke. Te spremembe v ocenah srednje vrednosti na ravni OECD so skladne s spremembami parametrov postavk, predstavljenih v tabeli 1.

Shema 1: Sprememba v oceni srednje vrednosti za posamezne države na lestvici bralne pismenosti, če je ena bralna enota odstranjena iz skupine osmih sidrnih enot

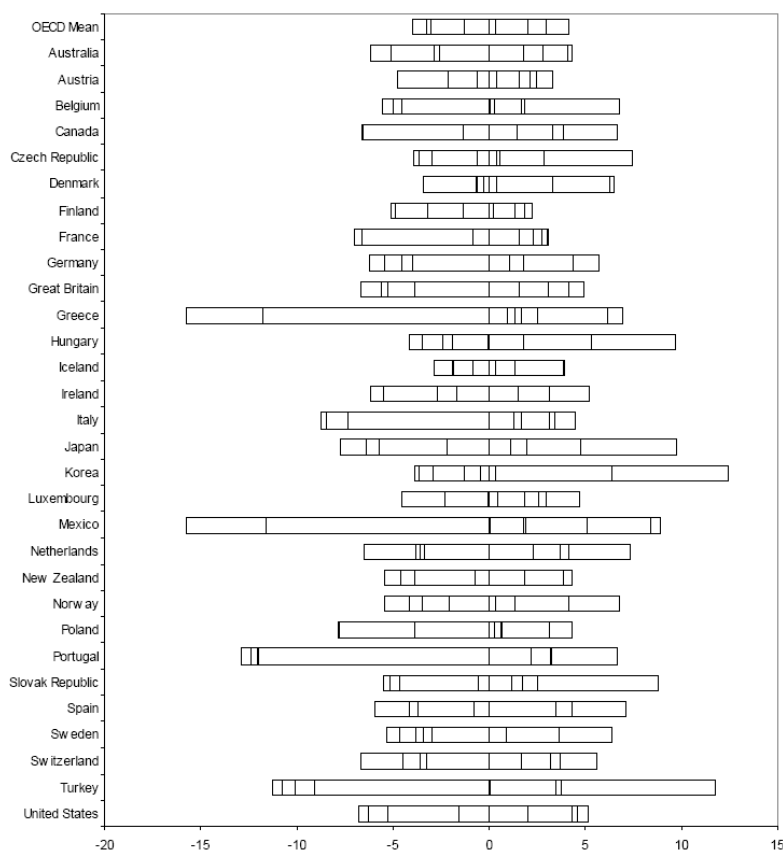


Tabela 4: Ocena napake povezovanja v posameznih državah

država	napaka povezovanja	država	napaka povezovanja
Avstralija	9,78	Japonska	13,85
Avstrija	7,93	Koreja	14,15
Belgija	8,60	Luksemburg	8,71
Kanada	12,83	Mehika	19,86
Češka	8,92	Nizozemska	11,74
Danska	10,68	Nova Zelandija	9,15
Finska	6,82	Norveška	9,84
Francija	10,22	Poljska	9,66
Nemčija	9,54	Portugal	14,38
Velika Britanija	9,78	Slovaška	11,20
Grčija	17,85	Španija	12,21
Madžarska	10,78	Švedska	7,78
Irska	9,84	Švica	10,58
Islandija	6,42	Turčija	16,13
Italija	10,47	ZDA	11,31

V določenih državah so pri nekaterih ponovitvah prisotne velike spremembe, ki ustrezajo parametrom postavk na nacionalni ravni. Na primer, parametri treh postavk v šesti enoti na mednarodni ravni znašajo -0,520, 0,308 in -0,887, v Avstraliji pa -0,470, -1,243 in -1,772. Medsebojni vpliv naloge in države pri drugi postavki šeste enote torej znaša približno 1,5 točk *logit*, pri zadnji pa 1 točko *logit*. V Grčiji so parametri postavk v drugi enoti približno eno točko *logit* nižji od parametrov na mednarodni ravni. Vpliv odstranitve te naloge bi v Grčiji znašal 15,79 točke *logit*.

V tabeli 4 je predstavljena napaka povezovanja, ocenjena z metodo ponavljanja *jackknife*. Te ocene napake povezovanja so precej višje od tistih, pridobljenih z metodo ocenjevanja, ki je bila uporabljena v raziskavi PISA 2003, kar je mogoče pripisati več dejavnikom:

- (i) združitev nalog v skupine v raziskavi PISA;
- (ii) naloge z možnimi delno pravilnimi odgovori;
- (iii) napake v specifikaciji na ravni posamezne države ali medsebojni vpliv naloge in države.

Največji upad bralne pismenosti je bil ugotovljen na Japonskem, kjer je srednja vrednost v raziskavi PISA 2000 znašala 522, leta 2003 pa 498 točk. Ta razlika je bila v začetnem poročilu raziskave PISA 2003 ugotovljena kot statistično izredno pomembna. Če napako povezovanja ocenimo z metodo *jackknife*, pa razlike ne bi več mogli šteti kot statistično pomembne.

Zaključki

Z vključitvijo napake izenačevanja dosežkov, ki jo je v nekaterih primerih treba prišteti varianci vzorčnih ocen in varianci vračunavanja (imputation variance), je PISA v preverjanja širokega obsega uvedla pomembno novost. Vendar pričujoč prispevek opozarja na štiri vprašanja, ki bi jih bilo treba raziskati pred vključitvijo napake izenačevanja dosežkov v prihodnje raziskave PISA. Ta vprašanja so:

1. hierarhična struktura nalog v raziskavah PISA;
2. končne in neskončne populacije;
3. politomne naloge; in
4. uniformnost napake izenačevanja dosežkov.

Zdi se, da metode ponavljanja (metodi *jackknife* in *bootstrap*) ponujata zanimivo alternativo analitičnim rešitvam, ki lahko zlahka premosti hierarhično strukturo nalog v raziskavi PISA. Rezultati simulacije so kar obetavni.

Zaradi variabilnosti napake izenačevanja dosežkov je treba izračunati specifične napake izenačevanja dosežkov za vsako državo. To je tudi prednost vključevanja napak v specifikaciji na ravni posamezne države v izračun napake povezovanja, čeprav to ne reši problema primerjav znotraj posamezne države.

Prispevek opozarja tudi na pomembnost števila sidrnih postavk, še posebej pri oblikovanju postavk v enote. Povečanje števila postavk bo zmanjšalo učinek napak v specifikaciji na ravni posamezne države.

Viri

- Adams, R. J. in Wu, M. (ur.) (2002). *PISA 2000 technical report*. Pariz: OECD Publications.
- Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Kelly, D. L. in Smith, T. A. (1996a). *Mathematics achievement in the middle school years: IEA's third international mathematics and science study*. Chestnut Hill, MA: Boston College.
- Beaton, A. E., Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Smith, T. A. in Kelly, D. L. (1996b). *Science achievement in the middle school years: IEA's third international mathematics and science study*. Chestnut Hill, MA: Boston College.
- Cochran, W. G. (1977). *Sampling techniques*. New York: Wiley and Sons.

- Cronbach, L. J., Linn, R. L., Brennan, R. L. in Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement*, 57, 373-399.
- Dagnelie, P. (1998). *Statistique théorique et appliquée. Tome 1: Statistique descriptive et bases de l'inférence statistique*. Paris and Bruxelles: De Boek and Larcier.
- Harris, D. J. in Crouse, J. D. (1993). A study of criteria used in equating. *Applied Measurement in Education*, 6, 195-240.
- Kish, L. (1995). *Survey sampling*. New York: Wiley and Sons.
- Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Gregory, K. D., Smith, T. A., Chrostowski, S. J., et al. (2000). *TIMSS 1999 international science report*. Chestnut Hill, MA: Boston College.
- Martin, M. O., Mullis, I. V. S., Gonzalez, E. J. in Kennedy, A. M. (2003). *PIRLS trends in children's reading literacy achievement 1991-2001. IEA's repeat in nine countries of the 1991 reading literacy study*. Chestnut Hill, MA: Boston College.
- Martin, M. O., Mullis, I. V. S., Gonzalez, E. J. in Chrostowski, S. K. (2004). *TIMSS 2003 international science report: Findings from the IEAs trends in international mathematics and science study at the fourth and eighth grades*. Chestnut Hill, MA: Boston College.
- Michaelides, M. P. in Haertel, E. H. (2004). *Sampling of common items: An unrecognized source of error in test equating* (Technical Report). Los Angeles: Center for the Study of Evaluation and National Center for Research on Evaluation, Standards, and Student Testing.
- Monseur, C., Sibberns, H. in Hastedt, D. (2006, November). *Equating errors in international survey in education*. IEA Second International Research Conference, Washington, DC.
- Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Gregory, K. D., Garden, R. A., O'Connor, K. M., et al. (2000). *TIMSS 1999 international mathematics report*. Chestnut Hill, MA: Boston College.
- Mullis, I. V. S., Martin, M. O., Gonzalez, E. J. in Chrostowski, S. J. (2004). *TIMSS 2003 international mathematics report: Findings from the IEAs trends in international mathematics and science study at the fourth and eighth grades*. Chestnut Hill, MA: Boston College.
- Rust, K. F. in Rao, J. N. K. (1996). Variance estimation for complex surveys using replication technique. *Statistical Methods in Medical Research*, 5, 283-310.

- OECD (1999). *Measuring student knowledge and skills: A new framework for assessment*. Pariz: OECD Publications.
- OECD (2004). *Learning for tomorrow's world: First results from PISA 2003*. Pariz: OECD Publications.
- OECD (2005). *PISA 2003 technical report*. Pariz: OECD Publications.
- Sheehan, K. M. in Mislevy, R. J. (1988). *Some consequences of the uncertainty in IRT linking procedures*. (Report No: ETS-RR-88-38-ONR) Princeton, NJ: Education Testing Service.
- Wolter, K. M. (1985). *Introduction to variance estimation*. New York: Springer-Verlag.
- Wonnacott, T. H. in Wonnacott, R. J. (1991). *Statistique: économie-gestion-sciences-médecine*. Pariz: Economica.
- Wright, B. D. in Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.
- Wu, M. L., Adams, R. J. in Wilson, M. R. (1997). ConQuest: Multi-aspect test software [Computer program]. Camberwell: Australian Council for Educational Research.
- Wu, M. (2005). The role of plausible values in large scale surveys. *Studies in Educational*

Dodatek

Tabela 1: Sprememba v oceni srednje vrednosti za posamezne države na lestvici bralne pismenosti, če je ena bralna enota odstranjena iz skupine osmih sidrnih enot

država	Sprememba v srednji vrednosti za posamezno državo brez...							
	enote 1	enote 2	enote 3	enote 4	enote 5	enote 6	enote 7	enote 8
Avstralija	1,82	-3,32	-1,09	4,07	-3,60	-6,17	4,30	2,77
Avstrija	3,34	-2,66	2,48	1,58	0,40	-4,17	-4,79	2,13
Belgija	6,79	-1,03	0,07	1,87	1,67	-0,53	-5,54	0,24
Kanada	-0,06	-6,64	1,41	3,87	-6,64	-5,31	3,32	6,62
Češka	0,60	-3,31	-0,21	-0,93	0,37	2,82	-3,90	7,44
Danska	6,52	6,31	-2,79	3,34	0,37	-3,11	-3,41	-2,76
Finska	1,87	1,36	-1,85	0,23	-0,13	-3,70	-5,05	2,24
Francija	3,06	3,01	2,30	2,71	-0,36	-6,11	-6,97	1,53
Nemčija	5,73	-0,79	1,07	4,39	1,77	-2,25	-6,24	-1,70
Velika Britanija	1,53	-2,82	-1,07	3,07	-1,42	-6,68	4,95	4,15
Grčija	6,95	-15,79	1,37	6,16	-4,04	0,91	2,54	1,67
Madžarska	9,04	-3,49	1,12	4,65	-1,06	-3,44	-1,62	-0,65
Islandija	3,86	-1,06	-2,07	1,34	3,94	-0,99	-2,89	0,31
Irska	0,02	-4,48	-0,65	1,50	-3,52	-6,18	3,16	5,23
Italija	4,49	-8,77	1,68	3,39	3,17	1,33	-1,44	-0,30
Japonska	9,76	1,13	-1,37	4,74	-7,72	-5,55	1,96	-2,03
Koreja	6,43	-3,83	0,31	-0,93	-0,16	-2,53	12,42	-3,34
Luksemburg	2,95	-4,48	1,88	2,61	4,68	-2,2	-4,54	0,45
Mehika	5,10	-15,79	1,79	8,42	-4,23	1,92	0,09	8,93
Nizozemska	7,30	3,68	-2,75	4,11	2,31	-3,13	-2,91	-6,52
Nova Zelandija	1,88	-1,56	-0,84	3,90	-4,71	-5,43	1,83	4,32
Norveška	6,75	0,33	-3,35	4,15	-1,91	-5,42	1,35	-1,25
Poljska	0,65	-7,83	-4,00	3,14	0,26	0,70	-0,03	4,31
Portugal	-0,84	-12,85	3,26	2,15	3,22	-0,76	-0,44	6,65
Slovaška	-0,82	-5,48	2,54	-0,35	1,76	1,16	-4,92	8,80
Španija	4,33	-5,94	3,43	4,31	-5,13	-2,22	-1,77	7,17
Švedska	5,77	-1,21	0,24	2,97	-0,84	-4,66	-1,64	-0,65
Švica	5,60	-3,40	3,68	3,21	1,68	-2,21	-6,66	-3,05
Turčija	11,75	-11,26	3,75	0,05	-0,48	-1,13	-2,21	3,45
ZDA	2,07	-6,78	-0,45	4,61	-1,53	-5,18	4,29	5,19
OECD povprečje	4,14	-3,96	0,33	2,95	-0,73	-2,67	-0,89	2,04