Overview of ILSA Analysis Issues

Anna Katyn Chmielewski, University of Toronto

Elizabeth Dhuey, University of Toronto

June 17, 2016

Workshop Series on International Large-Scale Assessment (ILSA)

National Academy of Education

Washington, DC

Introduction

- Advantages of ILSA data:
  - Large variation can be found across different education systems
  - Allows analysis of different kinds of education systems *and* validation of whether findings from one country hold in different contexts

- But ILSAs have been underutilized by US education researchers. Why?

Introduction

- Challenges/limitations of ILSA data:
  - Cultural/historical/contextual differences
  - Background questionnaires limited to info that can be collected comparably across countries
  - Lacking in longitudinal data – hard to estimate causal inference

- Recent improvements and larger number of years of assessments have allowed a small but growing number of scholars to design studies that allow for causal inference (or close to it)

LITERATURE REVIEW

Variation used in data analysis

At what level does “treatment” occur?

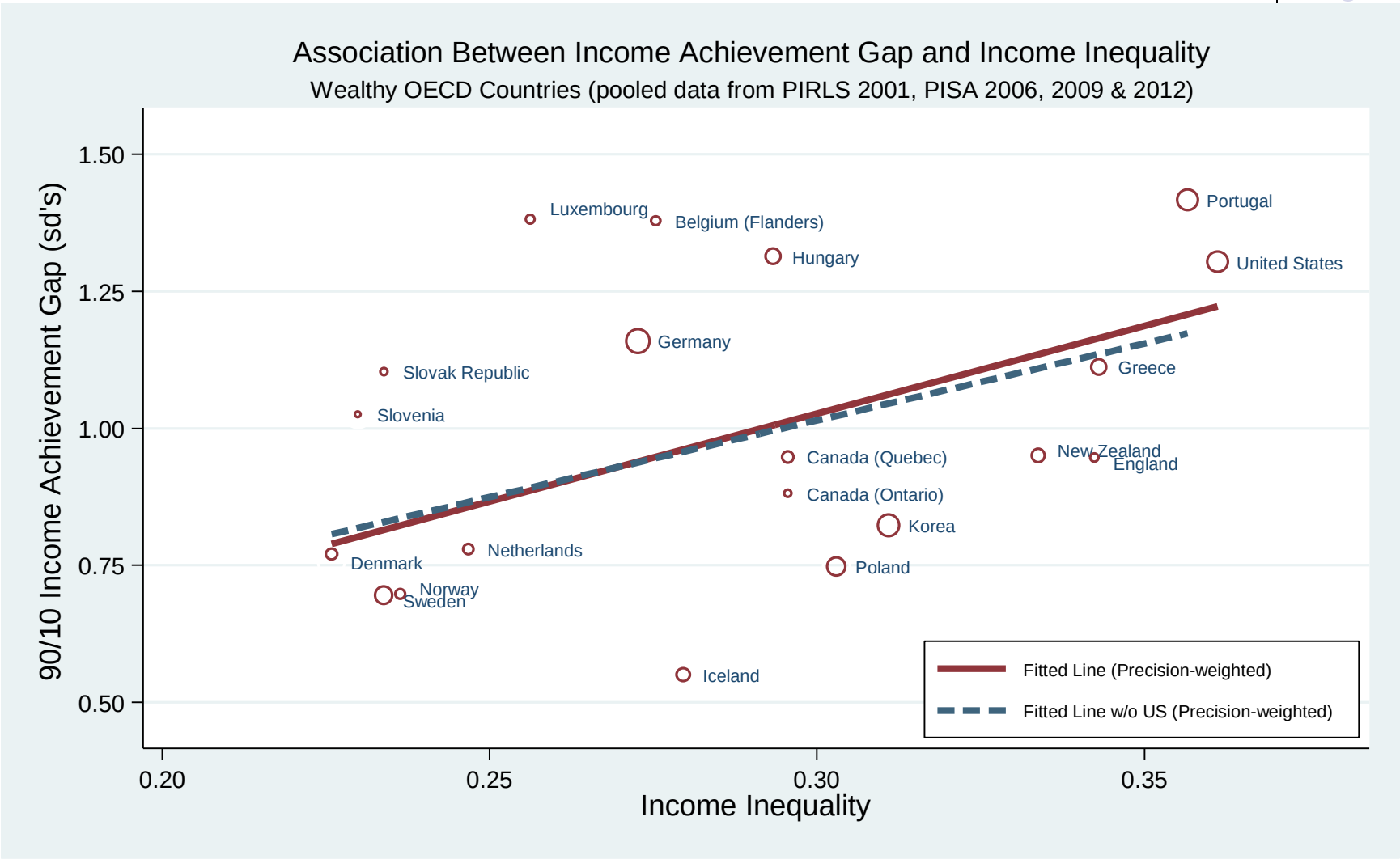
1. Cross-national variation
2. Within-country variation using cross-sectional data
3. Variation across birth cohorts within countries using repeated cross-sectional data
4. Variation in age within country-cohorts using synthetic cohorts
5. Variation in age using longitudinal follow-ups or longitudinal aspects of particular ILSAs

Variation used in data analysis

At what level does “treatment” occur?

1. Cross-national variation
2. Within-country variation using cross-sectional data
3. Variation across birth cohorts within countries using repeated cross-sectional data
4. Variation in age within country-cohorts using synthetic cohorts
5. Variation in age using longitudinal follow-ups or longitudinal aspects of particular ILSAs

1. Correlational cross-national variation



Source: Chmielewski & Reardon 2016

1. Correlational cross-national variation

- Correlation between national policies/characteristics and educational outcomes (non-causal)
- Gives a descriptive sense of what variation exists, suggests future areas of research
- Other examples (a *small* selection of many):
  - **Country GDP per capita:** Heyneman & Loxley 1983 (FISS 1970); Baker, Goesling & LeTendre 2002 (TIMSS 1995); Chudgar & Luschei 2009 (TIMSS 2003)
  - **Country income inequality:** Chudgar & Luschei 2009 (TIMSS 2003); Marks 2005 (PISA 2000)
  - **Country duration of democracy:** Torney-Purta et al. 2008 (CIVED 1999)
  - **Country tracking/ability grouping policies:** Buchmann & Park 2009 (PISA 2003); Chmielewski 2014 (PISA 2003); Chmielewski, Dumont & Trautwein 2013 (PISA 2003); Marks 2005 (PISA 2000); Pfeffer 2008 (IALS 1994); Schmidt et al. 2015 (PISA 2012)
  - **Country level of SES segregation between schools:** Willms 2006 (PISA 2006)

1. Correlational cross-national variation

- Correlation between national policies/characteristics and educational outcomes (non-causal)
- Gives a descriptive sense of what variation exists, suggests future areas of research
- Other examples (a *small* selection of many):
  - **Country GDP per capita:** Heyneman & Loxley 1983 (FISS 1970); Baker, Goesling & LeTendre 2002 (TIMSS 1995); Chudgar & Luschei 2009 (TIMSS 2003)
  - **Country income inequality:** Chudgar & Luschei 2009 (TIMSS 2003); Marks 2005 (PISA 2000)
  - **Country duration of democracy:** Torney-Purta et al. 2008 (CIVED 1999)
  - **Country tracking/ability grouping policies:** Buchmann & Park 2009 (PISA 2003); Chmielewski 2014 (PISA 2003); Chmielewski, Dumont & Trautwein 2013 (PISA 2003); Marks 2005 (PISA 2000); Pfeffer 2008 (IALS 1994); Schmidt et al. 2015 (PISA 2012)
  - **Country level of SES segregation between schools:** Willms 2006 (PISA 2006)

1. Correlational cross-national variation

- Correlation between national policies/characteristics and educational outcomes (non-causal)
- Gives a descriptive sense of what variation exists, suggests future areas of research
- Other examples (a *small* selection of many):
  - **Country GDP per capita:** Heyneman & Loxley 1983 (FISS 1970); Baker, Goesling & LeTendre 2002 (TIMSS 1995); Chudgar & Luschei 2009 (TIMSS 2003)
  - **Country income inequality:** Chudgar & Luschei 2009 (TIMSS 2003); Marks 2005 (PISA 2000)
  - **Country duration of democracy:** Torney-Purta et al. 2008 (CIVED 1999)
  - **Country tracking/ability grouping policies:** Buchmann & Park 2009 (PISA 2003); Chmielewski 2014 (PISA 2003); Chmielewski, Dumont & Trautwein 2013 (PISA 2003); Marks 2005 (PISA 2000); Pfeffer 2008 (IALS 1994); Schmidt et al. 2015 (PISA 2012)
  - **Country level of SES segregation between schools:** Willms 2006 (PISA 2006)

1. Correlational cross-national variation

- Correlation between national policies/characteristics and educational outcomes (non-causal)
- Gives a descriptive sense of what variation exists, suggests future areas of research
- Other examples (a *small* selection of many):
  - **Country GDP per capita:** Heyneman & Loxley 1983 (FISS 1970); Baker, Goesling & LeTendre 2002 (TIMSS 1995); Chudgar & Luschei 2009 (TIMSS 2003)
  - **Country income inequality:** Chudgar & Luschei 2009 (TIMSS 2003); Marks 2005 (PISA 2000)
  - **Country duration of democracy:** Torney-Purta et al. 2008 (CIVED 1999)
  - **Country tracking/ability grouping policies:** Buchmann & Park 2009 (PISA 2003); Chmielewski 2014 (PISA 2003); Chmielewski, Dumont & Trautwein 2013 (PISA 2003); Marks 2005 (PISA 2000); Pfeffer 2008 (IALS 1994); Schmidt et al. 2015 (PISA 2012)
  - **Country level of SES segregation between schools:** Willms 2006 (PISA 2006)

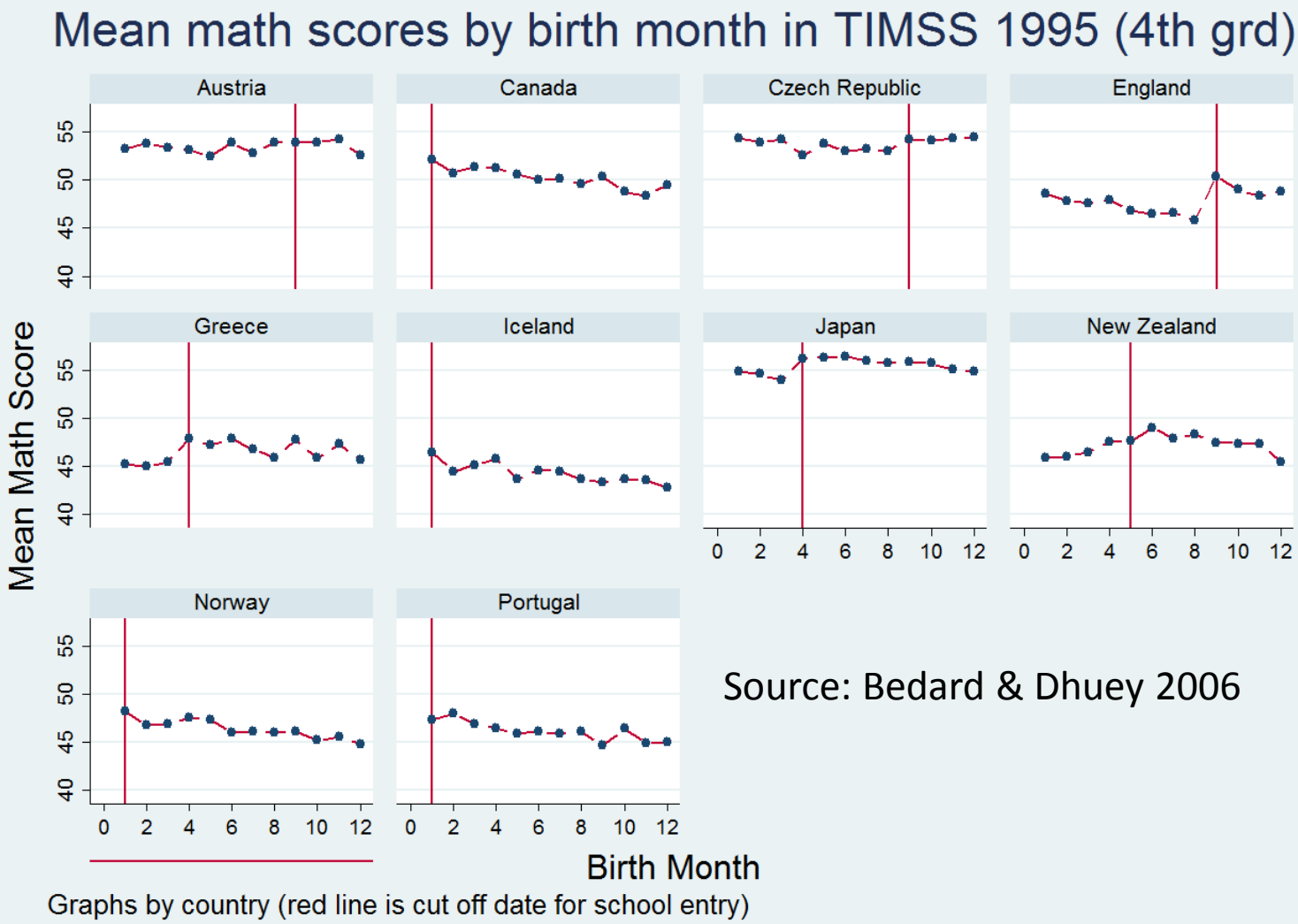
1(b). Random cross-national variation

- Exploit (close to) random variation in national policies
- Examples:
  - **Variation in share of Catholicism in 1900 as IV for size of private school sector:** West & Woessmann 2010 (PISA 2003)
  - **Extent of traditional voice-telephony network as IV for broadband availability:** Falck et al. 2016 (PIAAC 2011)

1(b). Random cross-national variation

- Exploit (close to) random variation in national policies
- Examples:
  - **Variation in share of Catholicism in 1900 as IV for size of private school sector:** West & Woessmann 2010 (PISA 2003)
  - **Extent of traditional voice-telephony network as IV for broadband availability:** Falck et al. 2016 (PIAAC 2011)

2. Within-country variation



Source: Bedard & Dhuey 2006

2. Within-country variation

- Exploit (close to) random variation in student treatments within countries
- Other examples:
  - **School fixed effects to estimate classroom peer effects (assume random class assignment within primary schools in Europe):** Ammermueller & Pischke 2009 (PIRLS 2001)
  - **Subject-specific class size (student fixed effects):** Altinok & Kingdon 2012 (TIMSS 2003)
  - **Between-grade variation in cohort size as IV for class size, school fixed effects (TIMSS 1995 7th & 8th grades):** Ammermueller et al. 2005; Woessmann 2005; Woessmann & West 2006
  - **Propensity-score matching – likelihood of participating in shadow education:** Park 2013 (PISA 2006)
  - **Variation in tracking policies across age cohorts in cross-sectional IALS 1994 (ages 16-65):** Ariga & Brunello 2007; Brunello & Checchi 2007

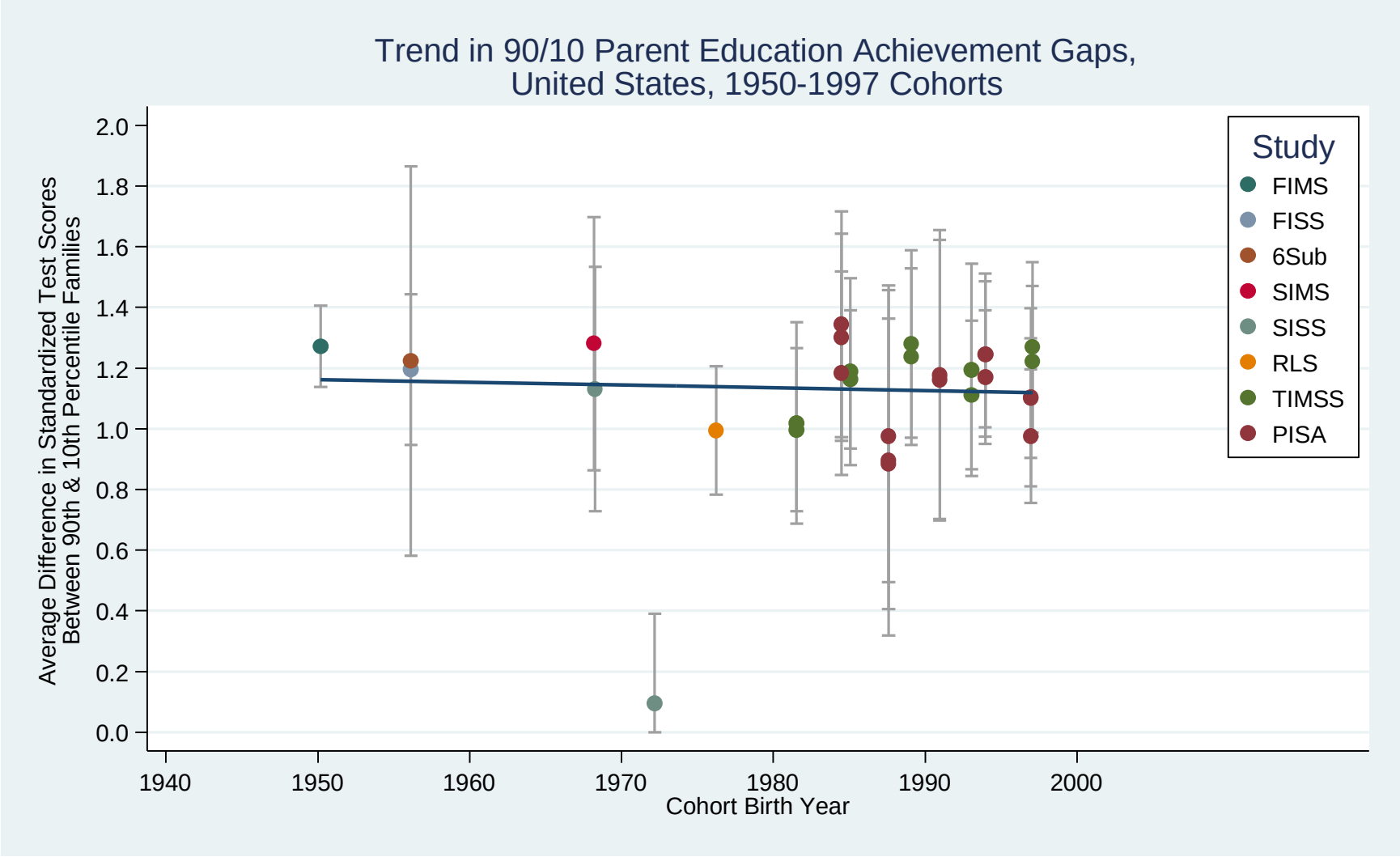
2. Within-country variation

- Exploit (close to) random variation in student treatments within countries
- Other examples:
  - **School fixed effects to estimate classroom peer effects (assume random class assignment within primary schools in Europe):** Ammermueller & Pischke 2009 (PIRLS 2001)
  - **Subject-specific class size (student fixed effects):** Altinok & Kingdon 2012 (TIMSS 2003)
  - **Between-grade variation in cohort size as IV for class size, school fixed effects (TIMSS 1995 7th & 8th grades):** Ammermueller et al. 2005; Woessmann 2005; Woessmann & West 2006
  - **Propensity-score matching – likelihood of participating in shadow education:** Park 2013 (PISA 2006)
  - **Variation in tracking policies across age cohorts in cross-sectional IALS 1994 (ages 16-65):** Ariga & Brunello 2007; Brunello & Checchi 2007

2. Within-country variation

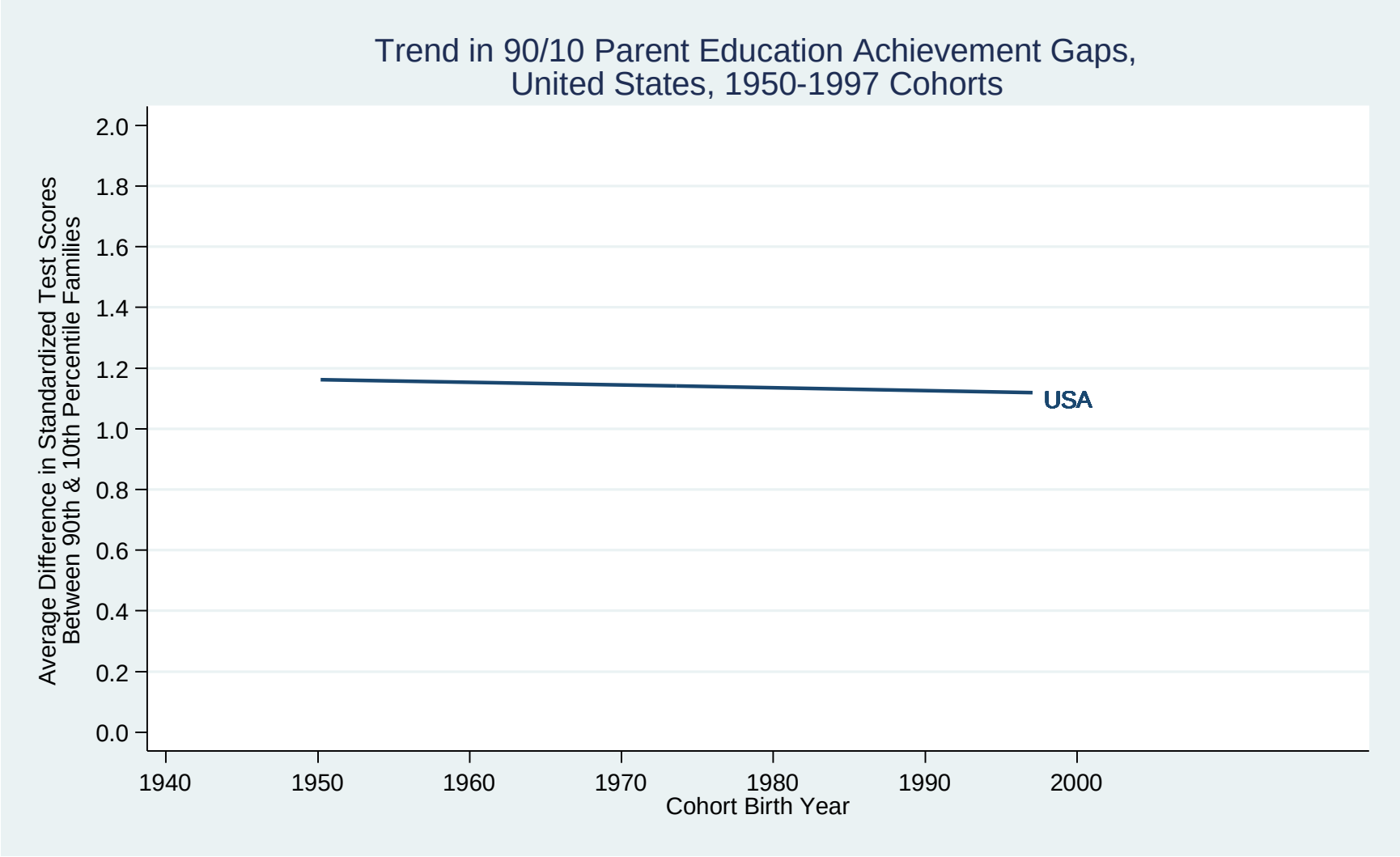
- Exploit (close to) random variation in student treatments within countries
- Other examples:
  - **School fixed effects to estimate classroom peer effects (assume random class assignment within primary schools in Europe):** Ammermueller & Pischke 2009 (PIRLS 2001)
  - **Subject-specific class size (student fixed effects):** Altinok & Kingdon 2012 (TIMSS 2003)
  - **Between-grade variation in cohort size as IV for class size, school fixed effects (TIMSS 1995 7th & 8th grades):** Ammermueller et al. 2005; Woessmann 2005; Woessmann & West 2006
  - **Propensity-score matching – likelihood of participating in shadow education:** Park 2013 (PISA 2006)
  - **Variation in tracking policies across age cohorts in cross-sectional IALS 1994 (ages 16-65):** Ariga & Brunello 2007; Brunello & Checchi 2007

3. Variation across birth cohorts within countries



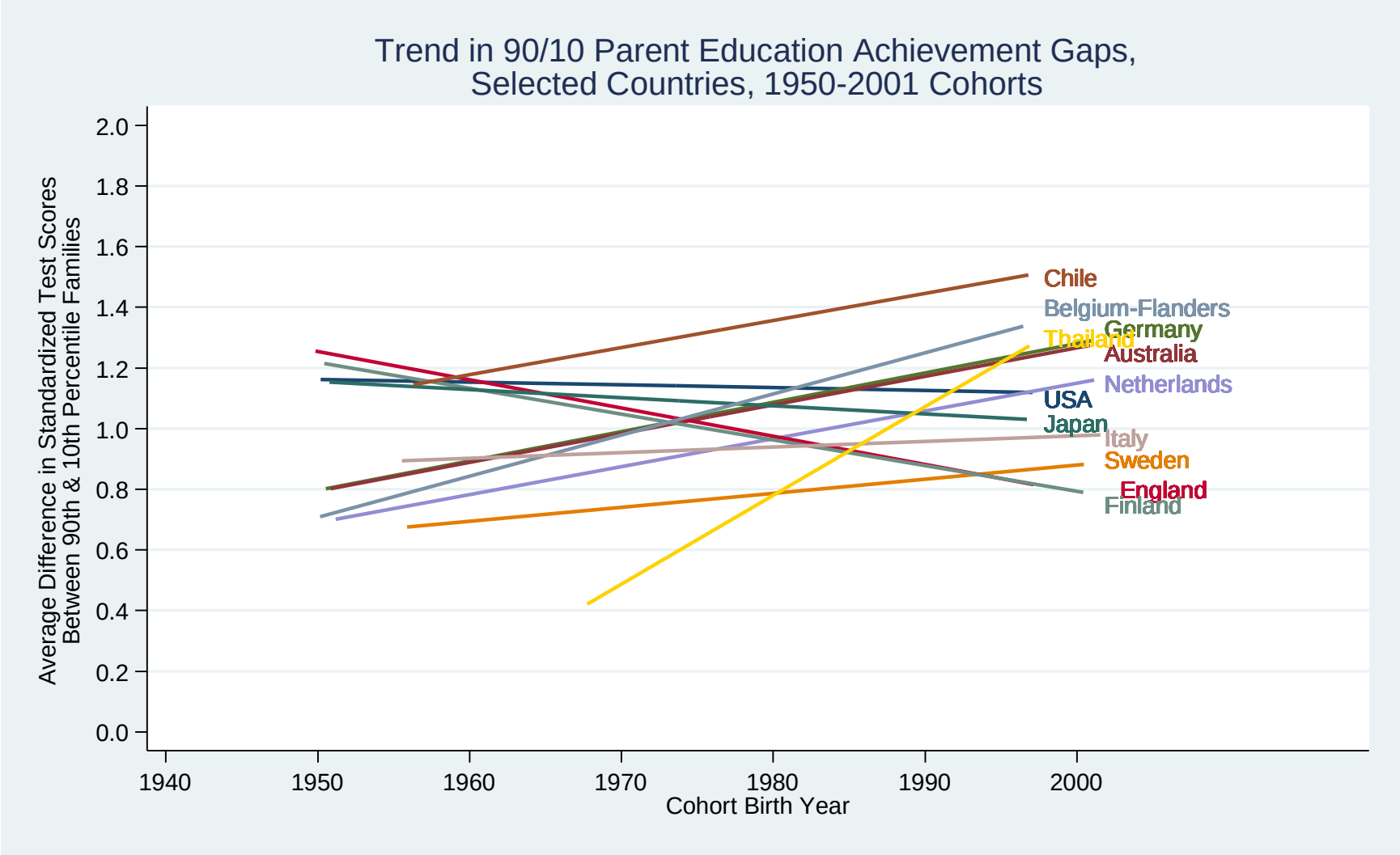
Source: Chmielewski, under review

3. Variation across birth cohorts within countries



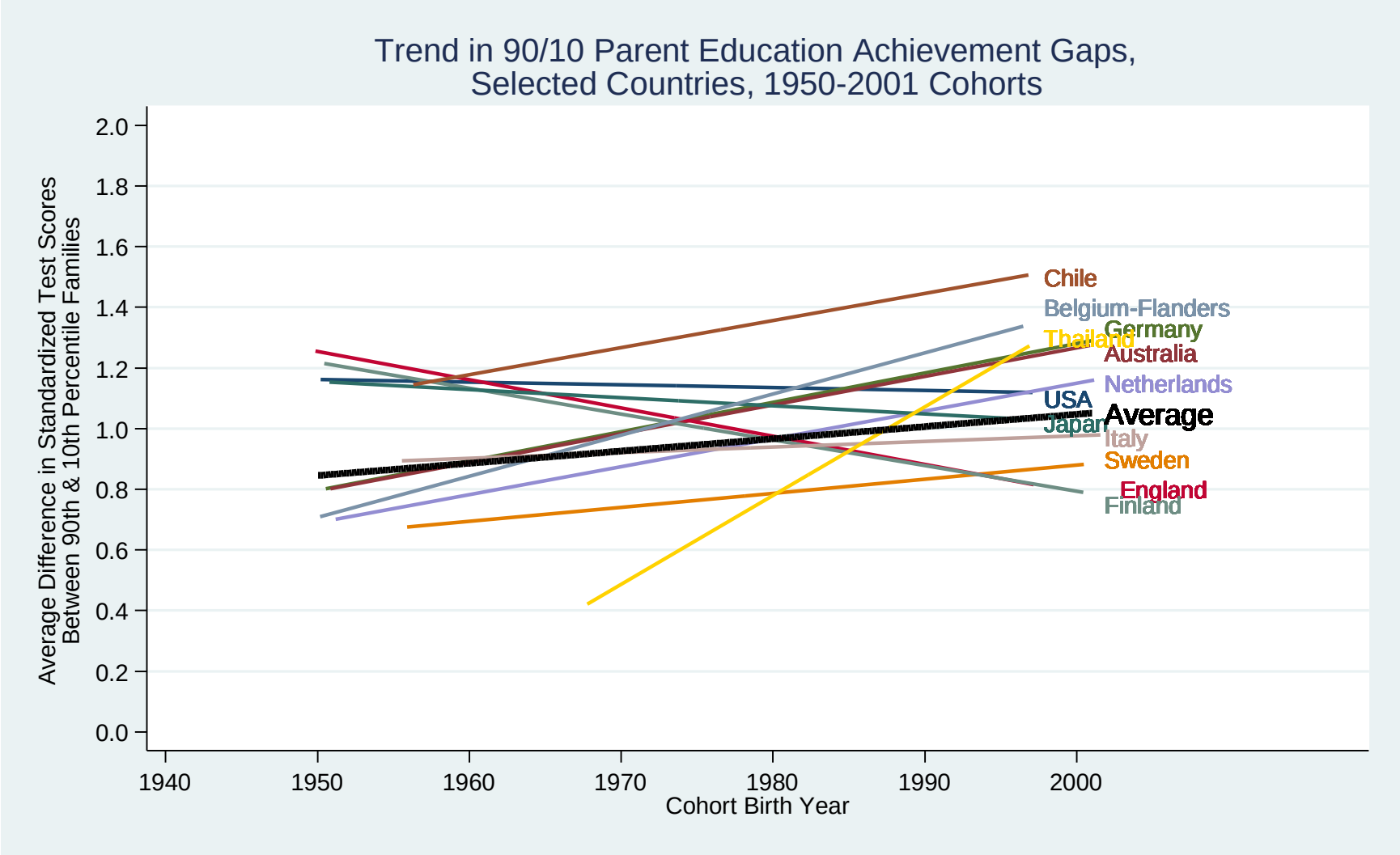
Source: Chmielewski, under review

3. Variation across birth cohorts within countries



Source: Chmielewski, under review

3. Variation across birth cohorts within countries



Source: Chmielewski, under review

3. Variation across birth cohorts within countries

- Association between changes in country policies and changes in educational outcomes (hold constant unobservable country characteristics)
- Designs such as diff-in-diff, country fixed and/or random effects/HLM, growth regressions
- Other examples:
  - **Multiple cycles of TIMSS, PIRLS, or PISA:** Chmielewski & Savage 2015 (PISA 2000-2012); Gustafsson 2013 (TIMSS 2003 & 2007); Gustafsson & Nilsen 2016 (TIMSS 2007 & 2011); Hanushek et al. 2013 (PISA 2000-2009); Rosen & Gustafsson 2016 (RLS 1991, PIRLS 2001 & 2006); Rutkowski et al. 2012 (TIMSS 1995-2007)
    - Same instrument – can compare levels of performance
  - **Combining older and recent tests (FIMS, FISS, SIMS, SISS, TIMSS, PISA, etc.):** Falch & Fischer 2012; Gundlach et al. 2001; Hanushek & Woessmann 2012; Wiseman et al. 2009
    - Different instruments – must use external anchor, standardize, etc.

3. Variation across birth cohorts within countries

- Association between changes in country policies and changes in educational outcomes (hold constant unobservable country characteristics)
- Designs such as diff-in-diff, country fixed and/or random effects/HLM, growth regressions
- Other examples:
  - **Multiple cycles of TIMSS, PIRLS, or PISA:** Chmielewski & Savage 2015 (PISA 2000-2012); Gustafsson 2013 (TIMSS 2003 & 2007); Gustafsson & Nilsen 2016 (TIMSS 2007 & 2011); Hanushek et al. 2013 (PISA 2000-2009); Rosen & Gustafsson 2016 (RLS 1991, PIRLS 2001 & 2006); Rutkowski et al. 2012 (TIMSS 1995-2007)
    - Same instrument – can compare levels of performance
  - **Combining older and recent tests (FIMS, FISS, SIMS, SISS, TIMSS, PISA, etc.):** Falch & Fischer 2012; Gundlach et al. 2001; Hanushek & Woessmann 2012; Wiseman et al. 2009
    - Different instruments – must use external anchor, standardize, etc.

3. Variation across birth cohorts within countries

- Association between changes in country policies and changes in educational outcomes (hold constant unobservable country characteristics)
- Designs such as diff-in-diff, country fixed and/or random effects/HLM, growth regressions
- Other examples:
  - **Multiple cycles of TIMSS, PIRLS, or PISA:** Chmielewski & Savage 2015 (PISA 2000-2012); Gustafsson 2013 (TIMSS 2003 & 2007); Gustafsson & Nilsen 2016 (TIMSS 2007 & 2011); Hanushek et al. 2013 (PISA 2000-2009); Rosen & Gustafsson 2016 (RLS 1991, PIRLS 2001 & 2006); Rutkowski et al. 2012 (TIMSS 1995-2007)
    - Same instrument – can compare levels of performance
  - **Combining older and recent tests (FIMS, FISS, SIMS, SISS, TIMSS, PISA, etc.):** Falch & Fischer 2012; Gundlach et al. 2001; Hanushek & Woessmann 2012; Wiseman et al. 2009
    - Different instruments – must use external anchor, standardize, etc.

4. Variation in age (synthetic cohorts)

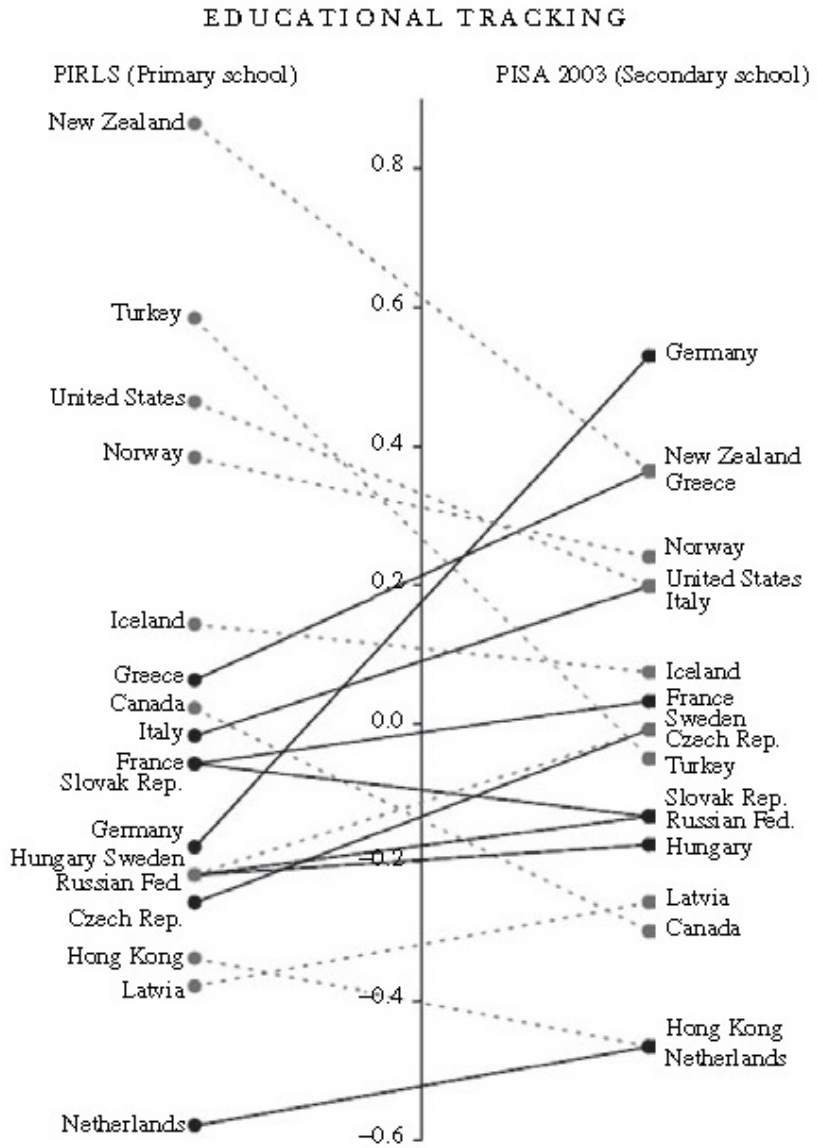


Fig. 1. *Inequality in Primary and Secondary School*
Notes. Standard deviation of test scores in the national population (difference from international average of national standard deviations in each test). Countries with a tracked school system before the age of 16 have solid lines, countries without tracking before age 16 have dashed lines.

Source: Hanushek & Woessmann 2006

4. Variation in age (synthetic cohorts)

- Match by cohort birth year, examine changes in educational outcomes by age (hold constant unobservable country *and* cohort characteristics); often use diff-in-diff designs
- Other examples:
  - **Effects of tracking (matching PIRLS, PISA, TIMSS 4th & 8th grades):** Ammermueller 2013; Jakubowski 2010; Ruhose & Schwerdt 2016
  - **Comparing achievement of TIMSS 1995 3rd & 4th grades, 7th & 8th grades:** Schmidt et al. 2001
  - **Matching early childhood & PISA data (children of NLSY79 in US, NLSCY in Canada):** Merry 2013
  - Looking at variation across cohorts *and* age:
    - **Matching cohorts in adult data (IALS, ALL, PIAAC):** Green & Riddell 2013 (variation in compulsory schooling laws as IV for educational attainment); Paccagnella 2016
    - **Matching childhood & adult data:** Thorn & Montt 2014 (PISA & PIAAC); Chmielewski in progress (FIMS, FISS, SIMS, SISS, TIMSS, PISA, IALS, ALL, PIAAC)

4. Variation in age (synthetic cohorts)

- Match by cohort birth year, examine changes in educational outcomes by age (hold constant unobservable country *and* cohort characteristics); often use diff-in-diff designs
- Other examples:
  - **Effects of tracking (matching PIRLS, PISA, TIMSS 4th & 8th grades):** Ammermueller 2013; Jakubowski 2010; Ruhose & Schwerdt 2016
  - **Comparing achievement of TIMSS 1995 3rd & 4th grades, 7th & 8th grades:** Schmidt et al. 2001
  - **Matching early childhood & PISA data (children of NLSY79 in US, NLSCY in Canada):** Merry 2013
  - Looking at variation across cohorts *and* age:
    - **Matching cohorts in adult data (IALS, ALL, PIAAC):** Green & Riddell 2013 (variation in compulsory schooling laws as IV for educational attainment); Paccagnella 2016
    - **Matching childhood & adult data:** Thorn & Montt 2014 (PISA & PIAAC); Chmielewski in progress (FIMS, FISS, SIMS, SISS, TIMSS, PISA, IALS, ALL, PIAAC)

4. Variation in age (synthetic cohorts)

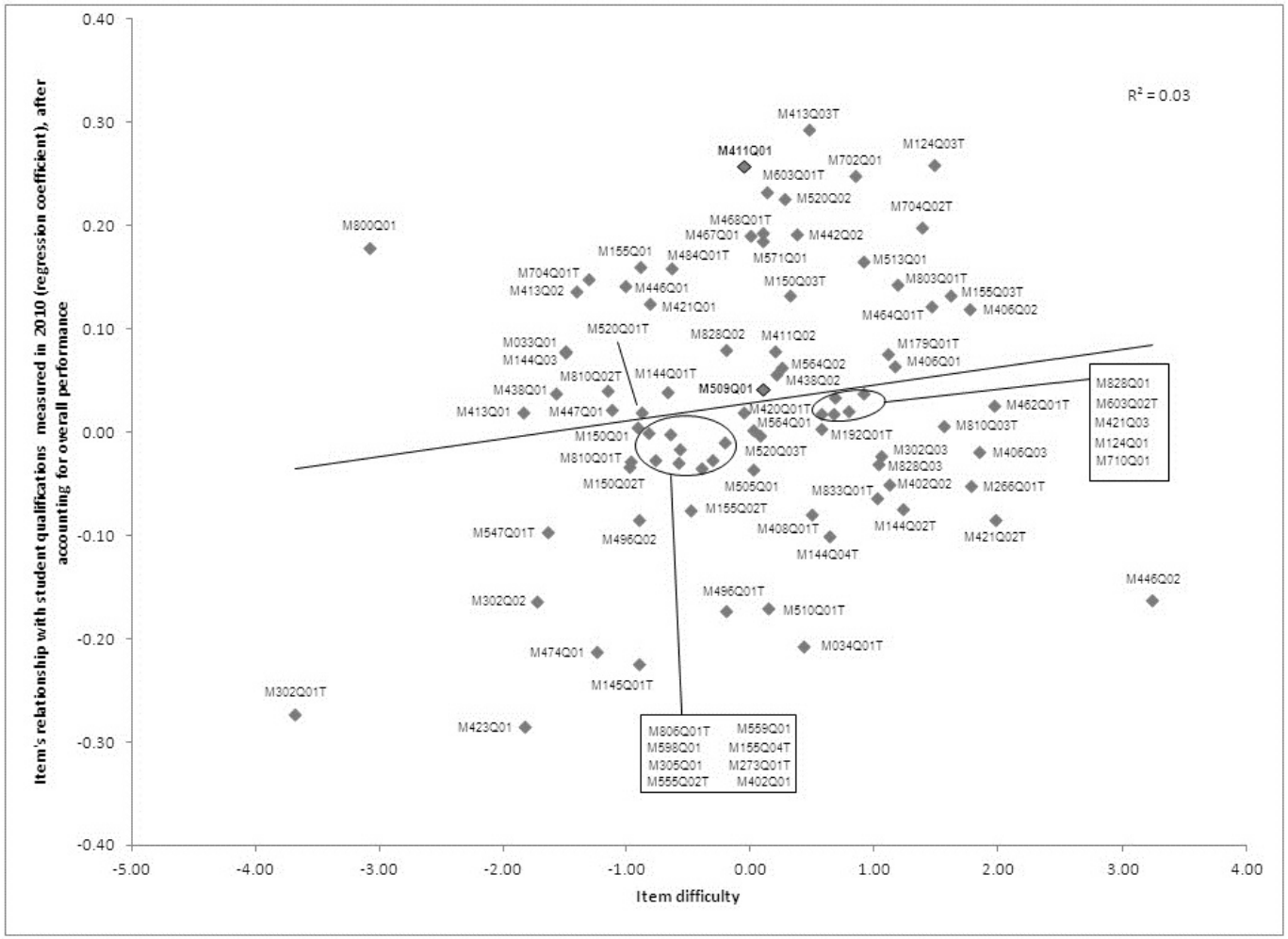
- Match by cohort birth year, examine changes in educational outcomes by age (hold constant unobservable country *and* cohort characteristics); often use diff-in-diff designs
- Other examples:
  - **Effects of tracking (matching PIRLS, PISA, TIMSS 4th & 8th grades):** Ammermueller 2013; Jakubowski 2010; Ruhose & Schwerdt 2016
  - **Comparing achievement of TIMSS 1995 3rd & 4th grades, 7th & 8th grades:** Schmidt et al. 2001
  - **Matching early childhood & PISA data (children of NLSY79 in US, NLSCY in Canada):** Merry 2013
  - Looking at variation across cohorts *and* age:
    - **Matching cohorts in adult data (IALS, ALL, PIAAC):** Green & Riddell 2013 (variation in compulsory schooling laws as IV for educational attainment); Paccagnella 2016
    - **Matching childhood & adult data:** Thorn & Montt 2014 (PISA & PIAAC); Chmielewski in progress (FIMS, FISS, SIMS, SISS, TIMSS, PISA, IALS, ALL, PIAAC)

4. Variation in age (synthetic cohorts)

- Match by cohort birth year, examine changes in educational outcomes by age (hold constant unobservable country *and* cohort characteristics); often use diff-in-diff designs
- Other examples:
  - **Effects of tracking (matching PIRLS, PISA, TIMSS 4th & 8th grades):** Ammermueller 2013; Jakubowski 2010; Ruhose & Schwerdt 2016
  - **Comparing achievement of TIMSS 1995 3rd & 4th grades, 7th & 8th grades:** Schmidt et al. 2001
  - **Matching early childhood & PISA data (children of NLSY79 in US, NLSCY in Canada):** Merry 2013
  - Looking at variation across cohorts *and* age:
    - **Matching cohorts in adult data (IALS, ALL, PIAAC):** Green & Riddell 2013 (variation in compulsory schooling laws as IV for educational attainment); Paccagnella 2016
    - **Matching childhood & adult data:** Thorn & Montt 2014 (PISA & PIAAC); Chmielewski in progress (FIMS, FISS, SIMS, SISS, TIMSS, PISA, IALS, ALL, PIAAC)

5. Variation in age (longitudinal)

Figure 2. Correlation between the level of difficulty of a mathematics item and the item's predictive power



Source: Jakubowski 2013

5. Variation in age (longitudinal)

- True longitudinal follow-ups to international assessments
- Other examples:
  - **Follow-ups to PISA (LSAY in Australia, YITS in Canada):** Jerrim, Chmielewski & Parker 2015; Jerrim & Vignoles 2015; Jerrim, Parker, Chmielewski & Anders 2015
  - **Longitudinal component of SIMS 1980:** Zimmer & Toma 2000

5. Variation in age (longitudinal)

- True longitudinal follow-ups to international assessments
- Other examples:
  - **Follow-ups to PISA (LSAY in Australia, YITS in Canada):** Jerrim, Chmielewski & Parker 2015; Jerrim & Vignoles 2015; Jerrim, Parker, Chmielewski & Anders 2015
  - **Longitudinal component of SIMS 1980:** Zimmer & Toma 2000

5. Variation in age (longitudinal)

- True longitudinal follow-ups to international assessments
- Other examples:
  - **Follow-ups to PISA (LSAY in Australia, YITS in Canada):** Jerrim, Chmielewski & Parker 2015; Jerrim & Vignoles 2015; Jerrim, Parker, Chmielewski & Anders 2015
  - **Longitudinal component of SIMS 1980:** Zimmer & Toma 2000

RECOMMENDATIONS

Is it acceptable to combine different tests?

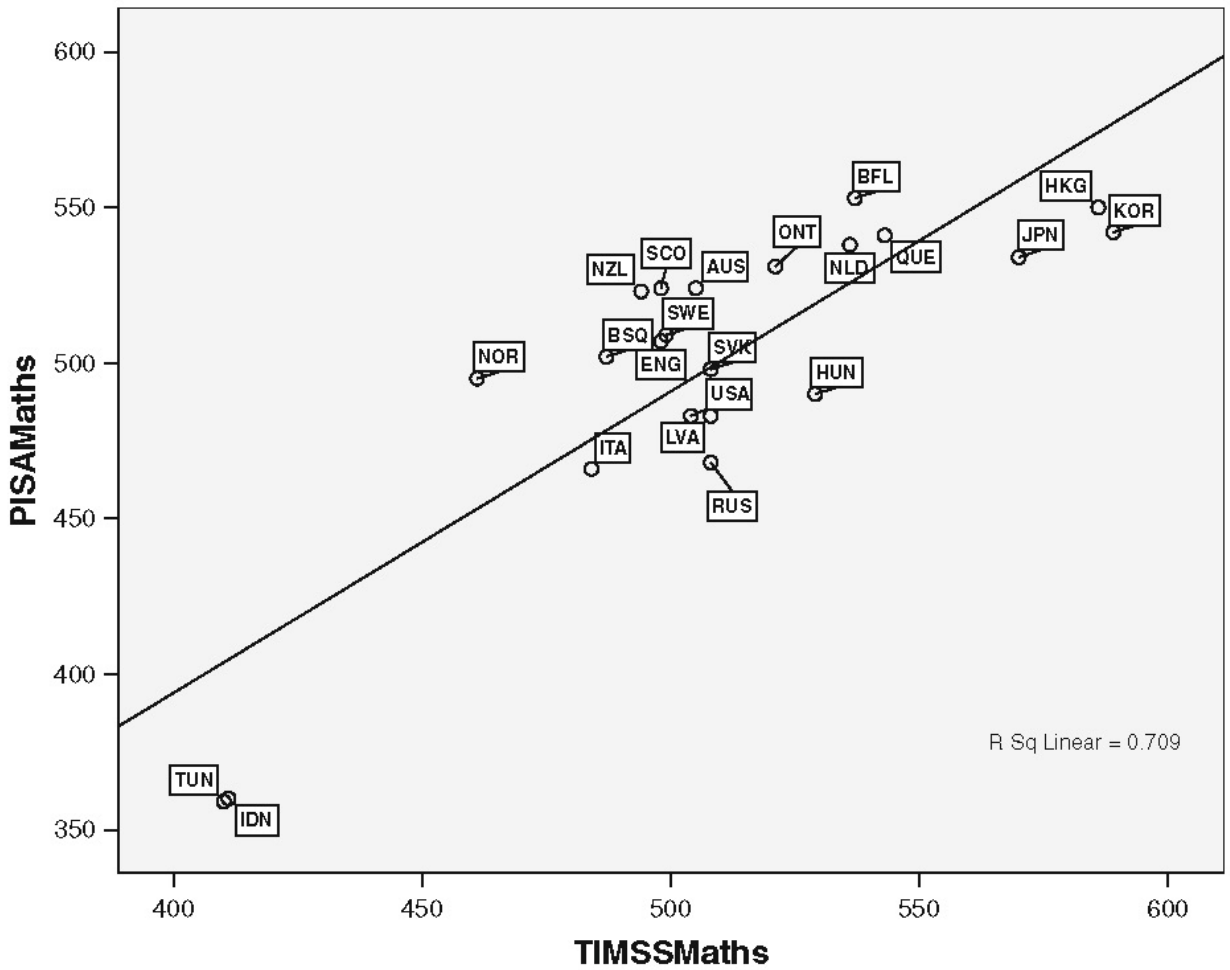


Fig. 1 Mathematics country mean scores: PISA 2003 versus TIMSS 2003

Source: Wu 2009

Years of schooling

Table 3 Age at time of TIMSS testing (proxy for number of years of schooling at time of PISA testing)

Country (sorted by age)	Age at time of TIMSS testing	Country (sorted by age)	Age at time of TIMSS testing
LVA	15.0	RUS	14.2
SWE	14.9	USA	14.2
TUN	14.8	QUE	14.2
KOR	14.6	BFL	14.1
HUN	14.5	BSQ	14.1
IDN	14.5	NZL	14.1
HKG	14.4	ITA	13.9
JPN	14.4	AUS	13.9
SVK	14.3	ONT	13.8
NLD	14.3	NOR	13.8
ENG	14.3	SCO	13.7

Source: Wu 2009

Content balance

Table 4 Proportions of PISA items by TIMSS/NAEP content domains by various researchers

	PISA items by Neidorf classification (by NAEP content domains)[a] (%)	PISA items by Gronmo & Olsen classification[b] (%)	PISA items by our classification (%)	TIMSS items by IEA classification[c] (%)	Difference between PISA and TIMSS, (using our classification) (%)
Number	21	25	38	29	9
Algebra	11	8	8	24	−16
Measurement	17	10	9	16	−7
Geometry	12	18	14	16	−2
Data	39	35	31	14	17

[a] 2% of the items were unclassified. Percentages in the Neidorf document add up to 103%, due to rounding or a minor error. The percentages are scaled so they add up to 100%. It made a 1% difference in some categories

[b] Note that four items were deemed unclassified; they did not fit the TIMSS framework

[c] See IEA (2004)

Source: Wu 2009

Is it acceptable to combine different tests?

- Wu (2009) explains 93% of difference
  - Content balance and years of schooling
- What are sources of 7%?
  - Differences in samples – target population exclusions (but usually similar), response rates
  - Curriculum-based vs. literacy-based test
  - Reading comprehension matters more for PISA than TIMSS
  - Other?
- Implications when combining different tests:
  - Carefully consider modeling age
  - Consider content domains separately

Is it acceptable to combine different tests?

- Approaches taken in the literature:
  - Want to compare **levels of performance**:
    - **Assume pooled mean of "core" countries is constant over time:** Falch & Fischer 2012
    - **Anchor US ILSA scores to NAEP scores:** Hanushek & Woessmann 2012; Gundlach et al. 2001
  - Want to compare **dispersion**:
    - **Assume pooled SD of "core" countries is constant over time:** Ammermueller 2013; Falch & Fischer 2012; Gundlach et al. 2001; Hanushek & Woessmann 2012; Jakubowski 2010; Ruhose & Schwerdt 2016
    - **SD compared to average national SD on each test:** Hanushek & Woessmann 2006
    - **Meta-analytic techniques – effect sizes (standardized scores within each country-year) and precision weighting:** Chmielewski & Reardon 2016; Chmielewski under review

Changes in background questionnaire wording

- Across waves:

PISA International version: **Q6 How old were you when you started <ISCED 1>?**

_____ *Years*

US adaptation 2003: **Q21 How old were you when you started elementary school?**

Average response: 5.38 _____ *Years*

US adaptation 2009: **Q9 How old were you when you started first grade?**

Average response: 5.9 _____ *years*

- Across countries:

TIMSS 2011 8th grade: **How far in your education do you expect to go?**

USA

Nationally defined options:
1 = Junior high/middle school
2 = High school
3 = Vocational/technical certificate after high school **1.9%**
4 = Associate's degree (AA) in a vocational/technical program **1.9%**
5 = 2-year or 4-year college or university degree (i.e., Associate's or Bachelor's degree) **42.6%**
6 = Beyond a bachelor's degree
7 = I don't know

Ontario, Canada

National options recoded for international comparability:
1 = Elementary school
2 = High school
3 = College or CÉGEP **19.1%**
4 = Business, trade or vocational school **1.2%**
5 = University - Bachelor's degree **22.9%**
6 = University - Master's or PhD or equivalent
7 = I don't know

Cross-cultural equivalence

- Differences in response patterns (Likert scales, etc.)
- Meaning of objective categories like college/university, private school, etc.
- Different explanatory variables might be important in different countries (Rutkowski & Delandshere 2016)

Weighting

Two issues:

1. Sample weights (e.g. student, classroom, school probability weights)
   - Affect **point estimates**
   - Sample weights are easy at student level
   - Sample weights are harder decisions with multilevel data (deriving conditional within-school weights, choosing house vs. senate weights at country level)
2. Estimation method (e.g., JRR, BRR)
   - Affect **standard errors**
   - IEA/OECD data user guides should provide recommendations for complex models

Conclusions

- This is a relatively small literature, but with a strikingly diverse set of approaches to data harmonization, combining tests, etc.
  - We have described many of these different approaches
- More methodological research is needed to weigh the relative advantages of the different approaches
- IEA/OECD/governments need to be involved in this discussion and dissemination