ON THE WRONG TRACK?
BIAS OF PUBLIC SECTOR WORKER PERFORMANCE MONITORING: THEORY AND EMPIRICAL EVIDENCE FROM MIDDLE SCHOOL TEACHERS

Douglas N. Harris
Associate Professor of Economics
University Endowed Chair in Public Education
Tulane University

Andrew A. Anderson, UW-Madison

MSU Value-Added Conference, Oct. 10-11,2013

Motivation

- Obvious concern about selection bias in using value-added for high-stakes decisions
- But, almost everything we known about selection bias comes from evidence of <u>elementary</u> schools
  - Rothstein
  - Kane and Staiger
  - Guarino, Reckase, and Wooldridge
- The problem is structurally different in middle and high schools where "tracking" is nearly universal

Tracking & 2 "Misalignment" Problems

- Metric Misalignment
  - Prime example: Ceiling effect (Koedel and Betts, 2009)
  - More generally, refers to anything we would normally define as measurement error for a given outcome that is systematically different across tracks (i.e., non-classical)
- Multidimensionality Misalignment
  - Well known that test scores capture only certain outcomes, i.e., educational output is "multidimensional"
  - This is a problem if only one output dimension is measured, and if the optimal level of the other output varies by track
- If the metric and multidimensionality problems may vary by track, then we have a new problem
- Both of these problems are much more likely in middle and high school compared with elementary

Optimal Monitoring and Tracking

- Decomposing the track effects is important to understanding how policymakers should use value-added
  - If metric misalignment, then get a <u>better</u> metric
  - If multidimensionality, then get <u>more</u> metrics
  - If neither, and it's sorting, then we may have a problem that simulations don't account for

Questions

- How much does it matter if we ignore tracks (no track indicators)?
- What share of the “track effect” is due to selection and the two forms of misalignment?

Preview

- Ignoring tracks can lead to considerable bias: 30-70% are in the wrong performance category
- The bias is composed mainly of selection bias, but there are two forms of “misalignment” that largely cancel out
- Jackson (2011) says we can address this by simply adding track indicators, but given the complexity of the potential role of tracks, it’s not clear this works

Model (a)

- Suppose we have the following measurement error:

$$A_{it} = A_{it}^{*} + \beta_0 r_{0it} + \beta_2 r_{2it} + \varepsilon_{it}$$

Metric Misalign.

- Suppose we have the following simple EPF:

$$A_{it}^{*} = \lambda A_{i,t-1}^{*} + \nu_{j(i,t)} + \gamma_i + \eta_{it}$$

NO Multid. Misalign.

- By substitution:

$$A_{it} = \lambda A_{i,t-1} + \beta_0 (r_{0it} - \lambda r_{0i,t-1}) + \beta_2 (r_{2it} - \lambda r_{2i,t-1}) + \nu_{j(i,t)} + \gamma_i + \varepsilon_{it} - \lambda \varepsilon_{i,t-1} + \eta_{it}$$

Model (b)

- Suppose we have the following measurement error:

$$A_{it} = A^*_{it} + \beta_0 r_{0it} + \beta_2 r_{2it} + \varepsilon_{it}$$

Metric Misalign.

- Suppose we have the following simple EPF:

$$A^*_{it} = \lambda A^*_{i,t-1} + \nu_{j(i,t)} + \gamma_i + \alpha_0 r_{0it} + \alpha_2 r_{2it} + \eta_{it}$$

YES Multid. Misalign.

- By substitution:

$$A_{it} = \lambda A_{i,t-1} + (\alpha_0 + \beta_0) r_{0it} - \lambda \beta_0 r_{0i,t-1} +$$
$$(\alpha_2 + \beta_2) r_{2it} - \lambda \beta_2 r_{2i,t-1} + \nu_{j(i,t)} + \gamma_i + \varepsilon_{it} - \lambda \varepsilon_{i,t-1} + \eta_{it}$$

Identification Issues

- Student ability (γ) typically unobserved
  - Here, we try FD and FE
- But probably imperfect and:

$$\mathrm{cov}(r_{it}, \gamma_i) \neq 0$$

$$\mathrm{cov}(v_{jt}, \gamma_i) \neq 0$$

- Identify from within-teacher variation in some cases (doesn't seem to matter)
- Instrument for tracks (track availability)

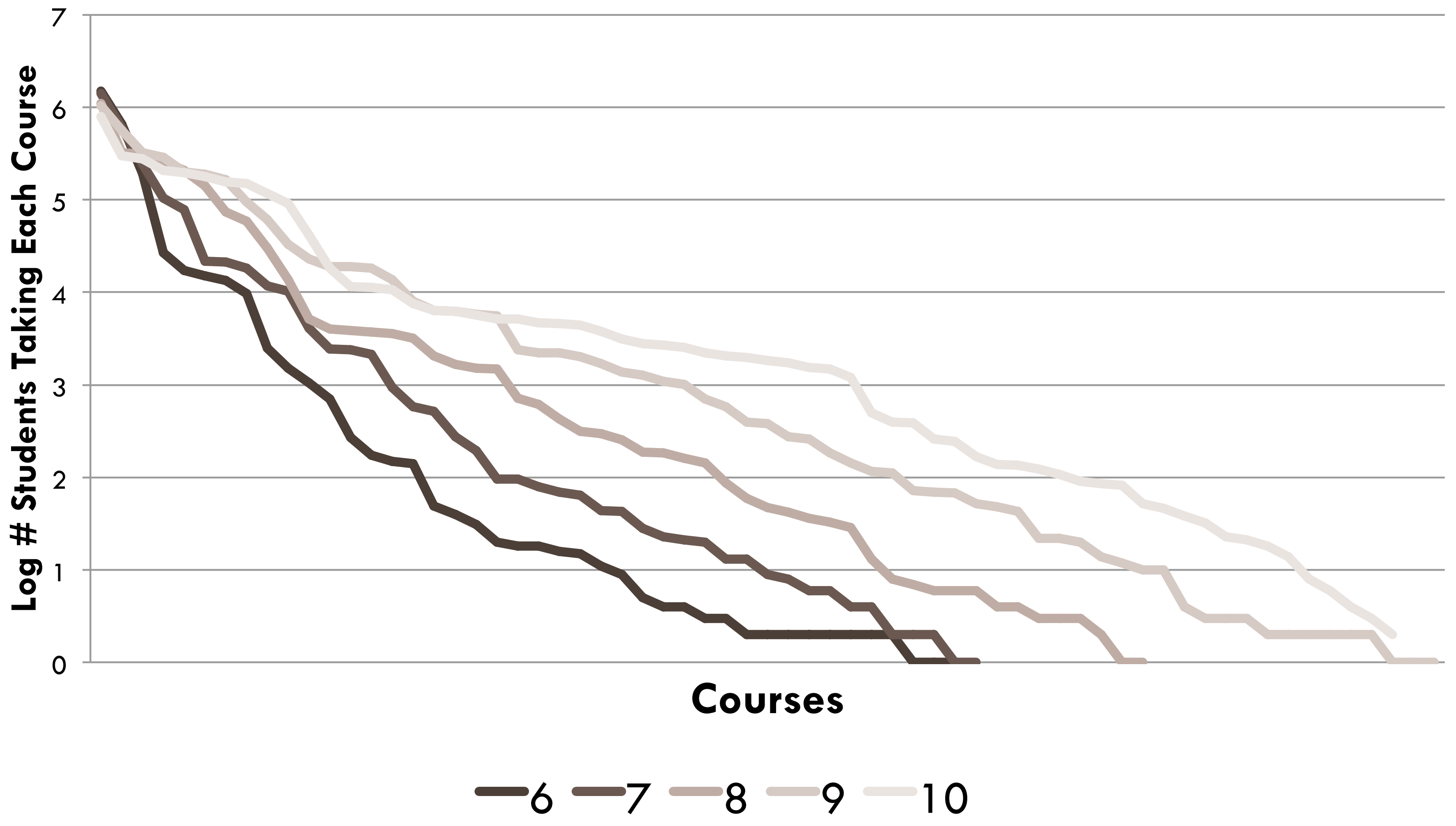
Florida Data

- Good matching of students to teachers
- Data on two tests: high-stakes (SSS) and low-stakes norm-referenced (NRT)
- Track data
  - Lower, middle, high (all by grade)
- Sample restrictions
  - Middle schools with grades 6-8
  - Only traditional public schools (no charters)
  - Classrooms with 10-40 students
  - Students with no more than 2 math teachers per year

Most Common Courses (& Tracks)

	Grade 6	Grade 7	Grade 8	Grade 9	Grade 10
Most Common	MATH 1	MATH 2	MATH 3	ALGEBRA 1	GEOMETRY
N	1,500,697	1,420,752	1,107,434	1,033,988	798,907
2nd Most	MATH 1, ADVANCED	MATH 2, ADVANCED	ALGEBRA 1	ALGEBRA 1-A	ALGEBRA 1-B
N	651,925	599,285	320,710	565,827	297,175
3rd Most	INTENSIVE MATH	INTENSIVE MATH	ALGEBRA 1 HONORS	GEOMETRY HONORS	ALGEBRA 2 HONORS
N	192,772	244,072	308,151	324,246	277,588
4th Most	GREAT EXPLOR. PRE-ALG.	PRE-ALGEBRA	MATH 3, ADVANCED	INTENSIVE MATH	GEOMETRY HONORS
N	26,799	103,879	263,034	290,304	209,644

Different Courses by Grade (log)



Results: Proof of Concept (MS Only)

Table 3: Track Effect Estimates from Standard Value-Added Specification (DOLS)

Variable	*SSS Test*	*NRT Test*
Lagged achievement	0.841**	0.795**
	(0.004)	(0.004)
Lower track indicator	-0.138**	-0.108**
	(0.013)	(0.013)
Advanced track indicator	0.241**	0.357**
	(0.003)	(0.006)
Student / year observations:	253,730	209,563

s.d. in teacher effects is 0.25

What if we omitted the track indicators? (MS Only)

Quartile of Teacher Value-Added with Track Coefficients ***Omitted***	*Quartile of Teacher Value-Added with Track Coefficients* ***Included***			
	1 (Low)	*2*	*3*	*4 (High)*
Lower Track Teachers				
1	0.706	0.221	0.074	0.000
2	0.056	0.250	0.389	0.306
3	0.000	0.000	0.275	0.725
4	0.000	0.012	0.000	0.988
Upper Track Teachers				
1	0.917	0.083	0.000	0.000
2	0.422	0.477	0.099	0.002
3	0.092	0.383	0.429	0.096
4	0.000	0.090	0.342	0.568

Track Coefficients with FD-IV(a) (MS)

Table 4: Track Effects from First-Differenced Model: All Teachers in All Schools

Model	*Sunshine State Standards Test (SSS)*			*Norm-Referenced Test (NRT)*		
	$\widehat{\beta_0}$	$\widehat{\beta_2}$	$\hat{\lambda}$	$\widehat{\beta_0}$	$\widehat{\beta_2}$	$\hat{\lambda}$
FD-Most Exogenous IVs:						
λ unrestricted	0.121***	0.019**	0.037	0.044*	0.116***	0.195***
	(0.022)	(0.009)	(0.078)	(0.025)	(0.012)	(0.060)
λ=0.4	0.129***	0.018*	0.400	0.045*	0.119***	0.400
	(0.023)	(0.010)	--	(0.025)	(0.013)	--
FD-Stronger IVs (less exogenous):						
λ unrestricted	0.117***	0.019**	-0.021	0.040*	0.111***	-0.038**
	(0.021)	(0.009)	(0.017)	(0.023)	(0.011)	(0.015)
$\lambda \geq 0$	0.119***	0.019**	0.000	0.042*	0.112***	0.000
	(0.021)	(0.009)	(0.001)	(0.023)	(0.011)	(0.001)
$\lambda = 0.4$	0.138***	0.016	0.400	0.056**	0.115***	0.400
	(0.023)	(0.010)	--	(0.026)	(0.013)	--

We also report results for FD-only (no IV) and the results are essentially the same

Track Coefficients with FE(b)

	7th		8th		9th		10th	
	SSS	NRT	SSS	NRT	SSS	NRT	SSS	NRT
α_0	-0.2829***	-0.0970***	-0.1639***	-0.0661***	-0.1142***	-0.0559***	-0.2456***	-0.1474***
	(0.0099)	(0.0129)	(0.0057)	(0.0067)	(0.0049)	(0.0063)	(0.0084)	(0.0095)
α_2	0.0977***	0.0421***	0.0573***	0.0440***	0.0385***	-0.0121*	0.0169*	-0.1364***
	(0.0074)	(0.0098)	(0.0045)	(0.0052)	(0.0038)	(0.0049)	(0.0072)	(0.0083)
β_0	0.3228***	0.1290***	0.2158***	0.1016***	0.1521***	0.0738***	0.2278***	0.1585***
	(0.0102)	(0.0135)	(0.0060)	(0.0071)	(0.0053)	(0.0067)	(0.0085)	(0.0097)
β_2	-0.0718***	-0.0174	-0.0282***	0.0029	-0.0052	0.0724***	0.0600***	0.2569***
	(0.0073)	(0.0098)	(0.0048)	(0.0055)	(0.0040)	(0.0052)	(0.0070)	(0.0082)

Track Coefficients with FE(b)

	7th		8th		9th		10th	
	SSS	NRT	SSS	NRT	SSS	NRT	SSS	NRT
α_0	-0.2829***	-0.0970***	-0.1639***	-0.0661***	-0.1142***	-0.0559***	-0.2456***	-0.1474***
	(0.0099)	(0.0129)	(0.0057)	(0.0067)	(0.0049)	(0.0063)	(0.0084)	(0.0095)
α_2	0.0977***	0.0421***	0.0573***	0.0440***	0.0385***	-0.0121*	0.0169*	-0.1364***
	(0.0074)	(0.0098)	(0.0045)	(0.0052)	(0.0038)	(0.0049)	(0.0072)	(0.0083)
β_0	0.3228***	0.1290***	0.2158***	0.1016***	0.1521***	0.0738***	0.2278***	0.1585***
	(0.0102)	(0.0135)	(0.0060)	(0.0071)	(0.0053)	(0.0067)	(0.0085)	(0.0097)
β_2	-0.0718***	-0.0174	-0.0282***	0.0029	-0.0052	0.0724***	0.0600***	0.2569***
	(0.0073)	(0.0098)	(0.0048)	(0.0055)	(0.0040)	(0.0052)	(0.0070)	(0.0082)
$\alpha_0+\beta_0$	0.040	0.032	0.052	0.036	0.038	0.018	-0.018	0.011
$\alpha_2+\beta_2$	0.026	0.025	0.029	0.047	0.033	0.060	0.077	0.121
Diff	-0.014	-0.007	-0.023	0.011	-0.005	0.042	0.095	0.109

A Problem in High School

- At the very least, the role of tracking is different at the high school level
- Also, many reasons not to trust the high school results
  - Challenge of placing courses into tracks
  - Other coefficients in model are surprising
    - Teacher experience is insignificant and wrong sign
    - Avg. class score is significant and wrong sign
- This may be why Jackson finds that current VA is a worse predictor of future VA in high school compared with elementary
- Two other related reasons . . .

More on High School Problem

- What if you're the teacher who teaches "Geometry, Honors" in both 9^{th} and 10^{th} grade
  - The test aligns poorly with Geometry in 9^{th} grade, but well in 10^{th}
  - Possible solution: Estimate separately by grade
    - And can just use course indicators rather than grouping into tracks
    - Jackson uses 9^{th} grade (not 10^{th})
- As our model shows, the lagged track also matters (through it's effect on lagged achievement)

Simulation Evidence

- Preliminary simulation evidence suggests that using only contemporaneous track is sufficient when students don’t switch tracks
- Much more bias for teachers of students who switch tracks
- Another case of the “exception is the rule problem”
  - Other examples: inexperienced teachers, team teaching, teaching highly advanced students in elementary grades, etc.

Conclusions

- General finding is the same in models (a) and (b): track coefficients largely reflect sorting, not misalignment
- The complex structure of the tracking problem has not apparently been addressed in prior simulation evidence
- This may explain the Jackson result that VA, as currently estimated, works less well in high school