Florida's Value-Added Model

Teacher Leader Preparation Implementation Committee Meeting

Orlando, Florida

November 9 and 10, 2011

D_{ata} U_{se}

Using Teacher Value-Added (VAM) Data

- Data available
    - VAM scores for teachers overall and by subject
    - Institution and program type (DACP, EPI, ITP)
    - Completer scores on certification exams
    - Identity and characteristics of destination school (e.g., FRL, percent minority, Title I status)
- Analysis conducted
    - Aggregate completers at state, institution, program levels
        - Choice of completers in terms of time after completion
- Results by institution and program type for tested teachers
        - Estimated average VAM scores
        - Precision of estimated scores as confidence intervals
- Decisions to be made
    - Whether and how to use multiple years of data
    - Confidence thresholds to use in evaluating programs
    - How much "weight" to give VAM data, what other information to use

AIR®

VAM Scores Available for Analysis

Annual VAM scores for completers

- by program, institution, institution type
- from 2007-08, 2008-09, and 2009-10
- for reading and math teachers in grades 4-10 and in tested subjects (by reading, math and overall)
- for the year following program completion

Cohort	Program Completion Year	VAM score year
Cohort 1	2007-08	2008-09
Cohort 2	2008-09	2009-10
Cohort 3	2009-10	2010-11

AIR®

Selection of One-Year Cohorts

- Used VAM data for completers one year out from program completion
  - For Cohort 1, could potentially have up to three years of teacher VAM data, and two years for Cohort 2
- Option of using VAM one year out helps to make cohorts more comparable; using more years out for cohorts introduces experience and other effects related to time
- Combine one-year completer cohorts across three years to have more observations

AIR®

Data
Use

AIR Analysis of VAM Data by Preparation Program

- Computed average VAM scores for each program with completers in tested subjects
  - Compared averages for completers by program to overall performance of these completers across state
- Applied different thresholds of statistical confidence to classify programs in terms of performance and assessed potential accuracy of classification
- Examined implications of different classification approaches

AIR®

Program Analysis Summary

Number of Programs with Completers with VAM Scores*

Year	DACP	EPI	ITP	Total
2007-08	30	26	28	84
2008-09	32	29	26	87
2009-10	28	30	32	90
Total	39	32	36	107

AIR®

*PTO is excluded from the analysis due to the small number of completers

ata
U_{se}

Total Number of Completers

Number of Completers by Program Type

Year	DACP	EPI	ITP	Total
2007-08	1,684	1,345	6,976	10,005
2008-09	1,626	1,944	7,249	10,819
2009-10	1,211	1,768	6,392	9,371
Total	4,521	5,057	20,617	30,195

AIR®

Completers with VAM Scores

Number of Completers by Program Type with VAM Scores

Year	DACP	EPI	ITP	Overall
2007-08	643	276	1,335	2,254
2008-09	605	422	1,343	2,370
2009-10	476	392	1,317	2,185
Total	1,724	1,090	3,995	6,809

■AIR®

Data Use

Confidence in Estimates is Influenced by Number of Completers with VAM Scores

Completers with VAM scores	Number of Programs		
	2007-08	2008-09	2009-10
10 or fewer	44	37	47
11 to 20	15	18	18
21 to 50	12	19	13
More than 50	13	13	12
Total	84	87	90

By combining data across years:
- 48 programs with fewer than 20 completers with VAM scores
- 16 programs with 51 to 150 completers with VAM scores
- 14 programs with more than 150 completers with VAM scores

AIR®

Data
Use

AIR Analysis of VAM Data by Preparation Program

- Compared averages for completers by program to overall performance of all tested completers in state

  - Computed by year
  - Computed across years

- Applied different thresholds of statistical confidence to classify programs in terms of performance and assessed potential accuracy of classification

- Provided summary data by program on VAM scores and confidence in estimates

AIR®

Average *Overall* VAM Scores of Completers

VAM scores are expressed as a proportion of a year's growth. On average, 2007-08 DACP completers have an overall VAM score (reading and math aggregate) that is about 2.2 percent lower than the expectation.

Year	DACP	EPI	ITP	All Programs
2007-08	-0.022	-0.046	-0.018	-0.023
2008-09	-0.006	-0.019	-0.029	-0.021
2009-10	0.013	-0.027	-0.045	-0.029
Overall	-0.007	-0.029	-0.031	-0.024

Average *Reading* VAM Scores of Completers

On average, 2007-08 DACP completers have a reading VAM score that is about 2.1 percent lower than the expectation in reading.

Year	DACP	EPI	ITP	All Programs
2007-08	-0.021	-0.065	-0.015	-0.022
2008-09	0.002	-0.02	-0.028	-0.02
2009-10	0.013	-0.066	-0.047	-0.039
Overall	-0.004	-0.048	-0.03	-0.027

AIR®

Average *Math* VAM Scores of Completers

On average, 2007-08 DACP completers have a math VAM score that is about 1.8 percent lower than the expectation in math.

Year	DACP	EPI	ITP	All Programs
2007-08	-0.018	-0.019	-0.022	-0.021
2008-09	-0.011	-0.025	-0.048	-0.035
2009-10	0.021	0.016	-0.045	-0.022
Overall	-0.005	-0.009	-0.038	-0.026

AIR

VAM Scores by Field of Study

May want to examine average reading and math VAM scores by field of study by program

- Compute reading VAM scores for those who completed reading programs; math VAM scores for those who completed math programs
- Includes ITPs only and reduces data for analysis to about 500 to 700 tested completers per year

Average *Reading* VAM Scores of *Reading* Program Completers

Reminder: VAM scores are expressed as a proportion of a year's growth. On average, 2007-08 reading completers had a reading VAM score that was about 1.2 percent lower than the expectation in reading.

Year	Reading
2007-08	-0.012
2008-09	-0.029
2009-10	-0.061
Overall	-0.034

Data **U**se

Average *Math* VAM Scores of *Math* Program Completers

On average, 2007-08 math program completers had a math VAM score that was about 2.7 percent lower than expectation in math.

Year	Math
2007-08	-0.027
2008-09	-0.076
2009-10	-0.061
Overall	-0.054

AIR®

Approach to Analysis of Programs

- For each program, we compute average single-year VAM scores and combined-year VAM scores for completers

  - *How should we compare the average for the program to the average for the state?*

- We will use statistical approach of estimating confidence intervals to compare program means to the state mean

- Use statistical results to discuss accuracy of classification of programs relative to the state mean

D

Role of Statistical Confidence

- Programs with similar outcomes will have variation in average VAM scores over time

  - Estimates for larger programs will be more precise; generally will have more confidence in whether the average VAM score for a program differs from the state mean

- Confidence intervals show the range from which the mean for a given program could have come

  - Similar to voting poll results reported with "margin of error" – e.g., program mean is 0.05 plus or minus 0.02

AIR

18

Impact of Using Different Levels of Confidence

We will have more confidence that a program mean differs from the state mean if it is estimated with a high level of precision (smaller standard error)

- The number of programs estimated to be different from the state mean will decrease as higher levels of confidence are used
- With higher levels of confidence, the less likely that programs with small numbers of completers will appear different from mean
- With higher levels of confidence, the less likely that programs with means close to the state average will be classified as different from mean

Importance of Confidence Level in Reviewing Program VAM Scores

Even if there is no systematic difference across programs, there will be program means that differ from state mean when we draw confidence intervals

- With lower levels of confidence, then more likely to say a program is different from state mean when it is not

- Can estimate the classification accuracy, which increases with higher levels of confidence

- By setting higher levels of confidence, we have more confidence we can identify programs that differ from the state mean

Describing Confidence of Program VAM Scores

- **"High" confidence**
  - At least 95 percent confident that the program mean differs from state mean
- **"Medium" confidence**
  - At least 90 percent confident that the program mean differs from state mean
- **"Low" confidence**
  - At least 68 percent confident that the program mean differs from state mean
- **"Very low" confidence**
  - program mean is not precisely estimated
  - less than 68 percent confident that the program mean is different from state mean
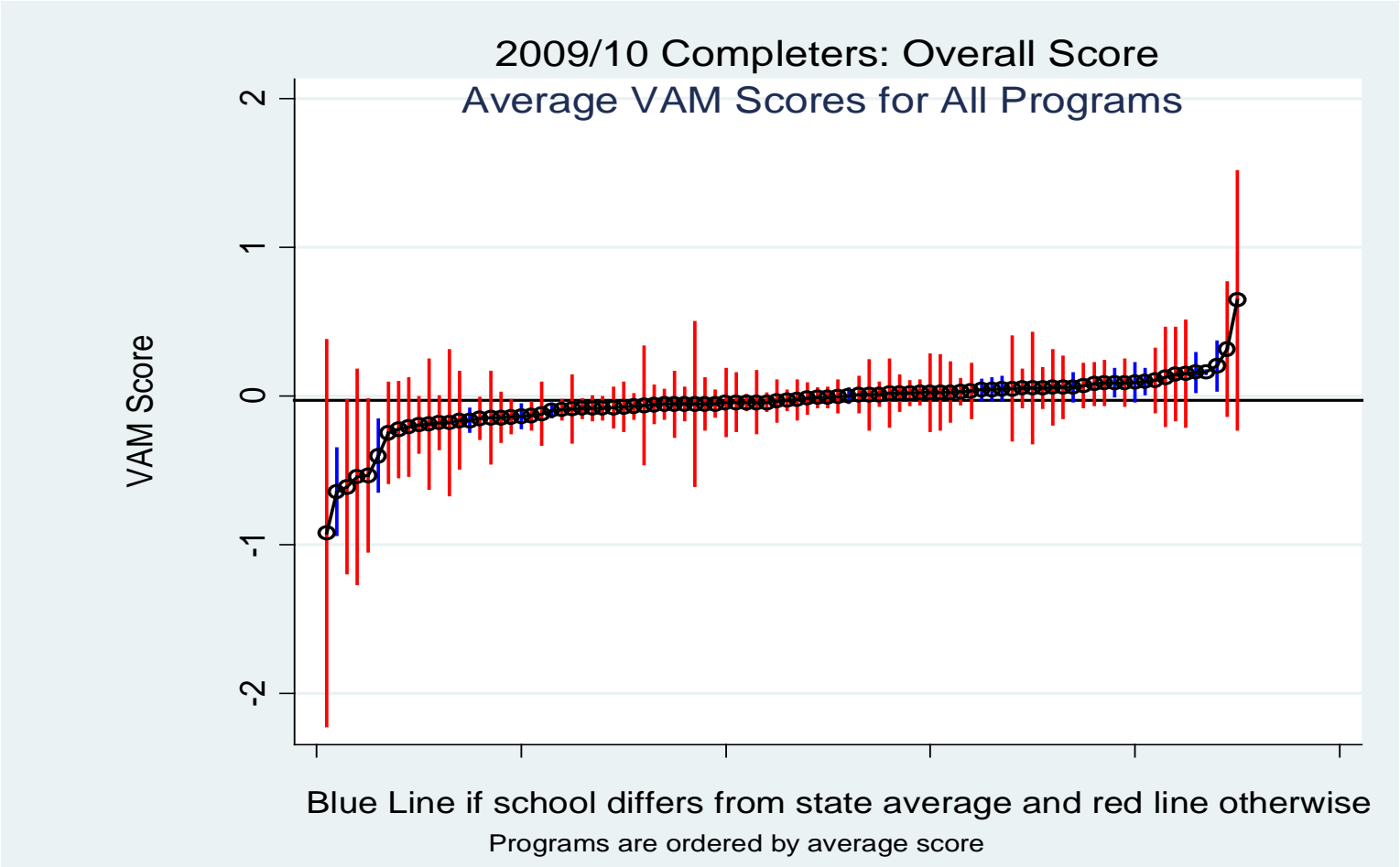
D_{ata} U_{se}

Using "Caterpillar" Charts to Show Confidence Given Level of Precision

- Next slides show average overall VAM scores in 2009-10 by program – hollow circles at center of each vertical line gives program average

  - *Programs ranked lowest to highest score*

- Vertical line around each score shows confidence interval given the level of confidence specified

- Rule: If vertical line does <u>not</u> touch horizontal line for state mean, then conclude program mean differs from state mean given level of confidence used
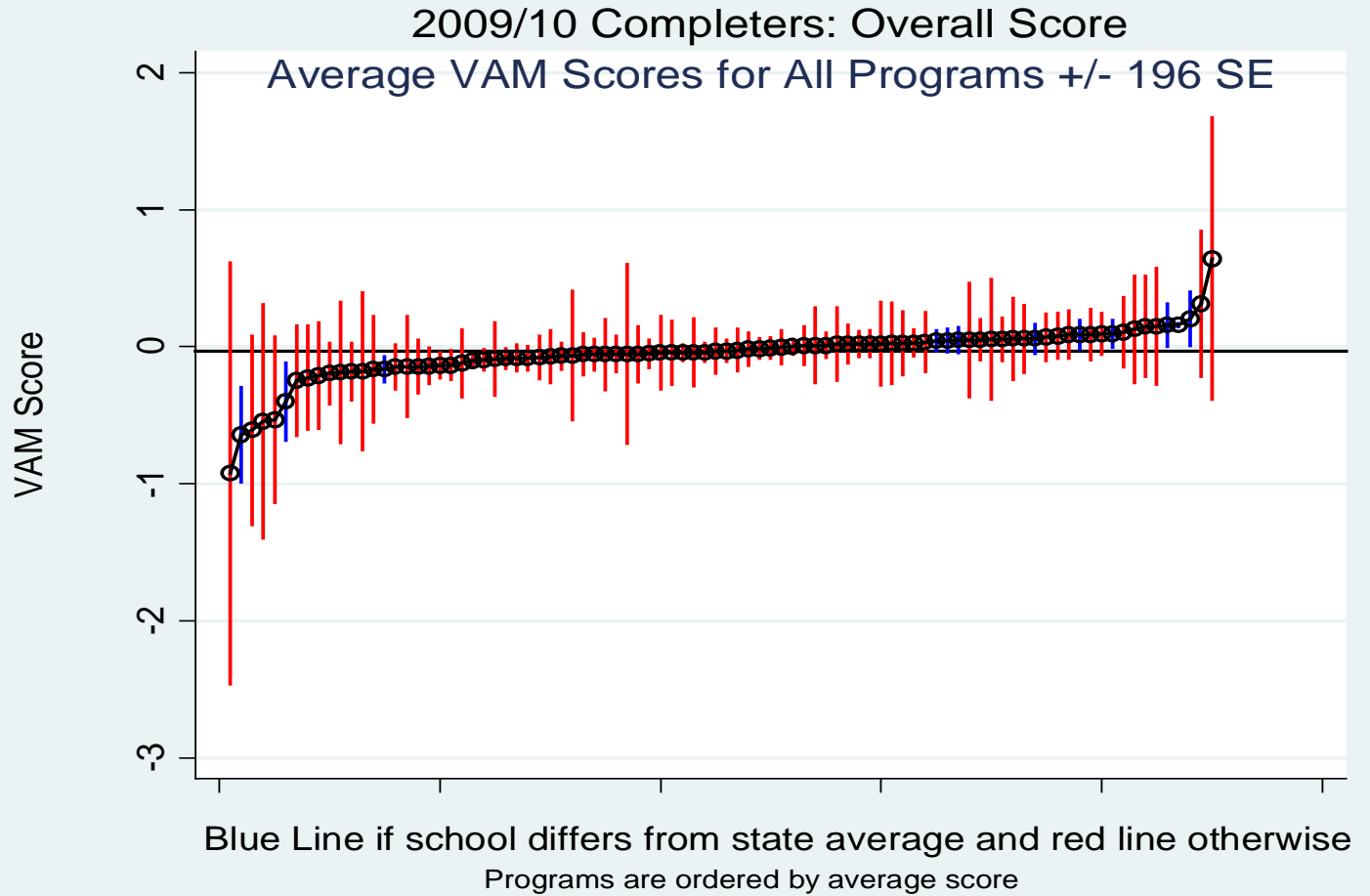
AIR

Mean Program Scores with Confidence Intervals for "Low" Confidence Level

2009/10 Completers: Overall Score
Average VAM Scores for All Programs

Blue line if school differs from state average and red line otherwise
Programs are ordered by average score

23
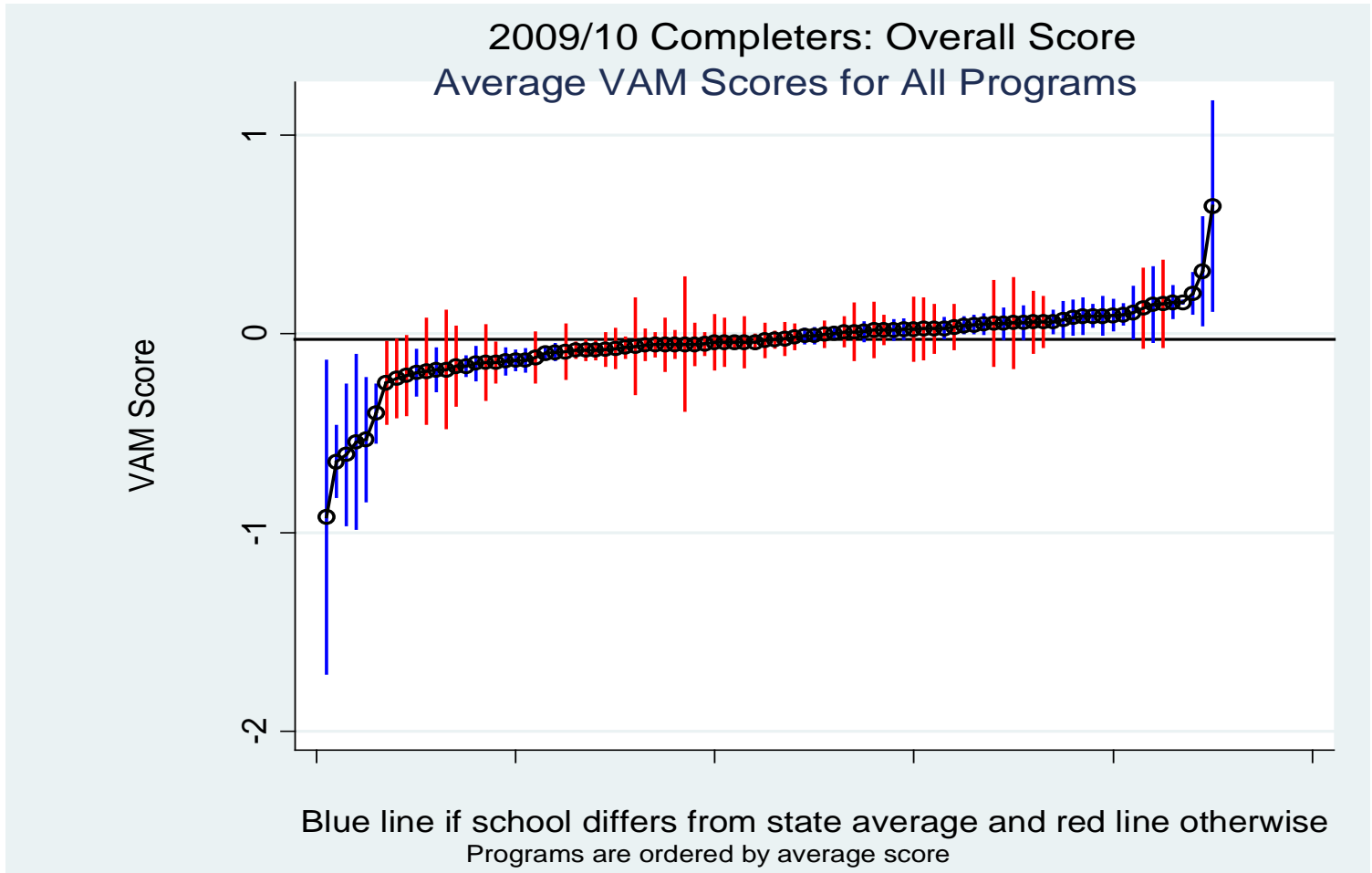
Mean Program Scores with Confidence Intervals for "Medium" Confidence Level

2009/10 Completers: Overall Score
Average VAM Scores for All Programs

Blue Line if school differs from state average and red line otherwise
Programs are ordered by average score

Mean Program Scores with Confidence Intervals for "High" Confidence Level

2009/10 Completers: Overall Score

Average VAM Scores for All Programs +/- 196 SE

VAM Score

Blue Line if school differs from state average and red line otherwise
Programs are ordered by average score

Data **U**se

Using Differences from State Average to Classify Programs

- Using information on average VAM by program and confidence level, can examine share of programs whose estimates are found to be different from mean (i.e., caterpillar "hairs" don't cross horizontal line for state mean on chart)

- *However,* some programs will likely appear different from mean when no true difference exists, especially when small number of completers

- Should consider implication of "false differences" for classification

  - Ask what share of observed differences likely reflect chance and therefore how many programs are potentially misclassified

Base Levels of Misclassification For Different Levels of Precision

There will be higher level of classification accuracy as required confidence level is set higher

- "High" level of confidence – 95%
  - 95 out of 100 programs expected to be classified accurately
  - 5 programs expected to be different from state mean when they are not
- "Medium" level of confidence – 90%
  - 90 out of 100 programs expected to be classified accurately
  - 10 programs expected to be different from state mean when they are not
- "Low" level of confidence – 68%
  - 68 out of 100 programs expected to be classified accurately
  - 32 programs found to be different from state mean when they are not

AIR®

Information on Classification of ITP Programs in Florida
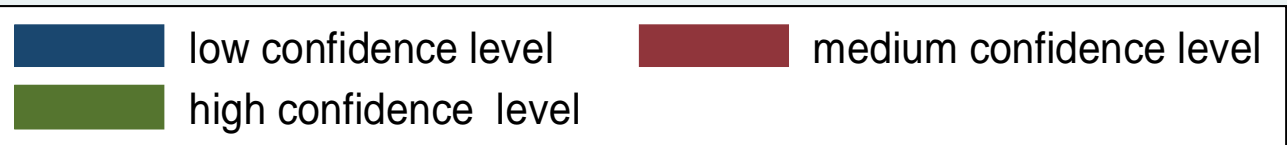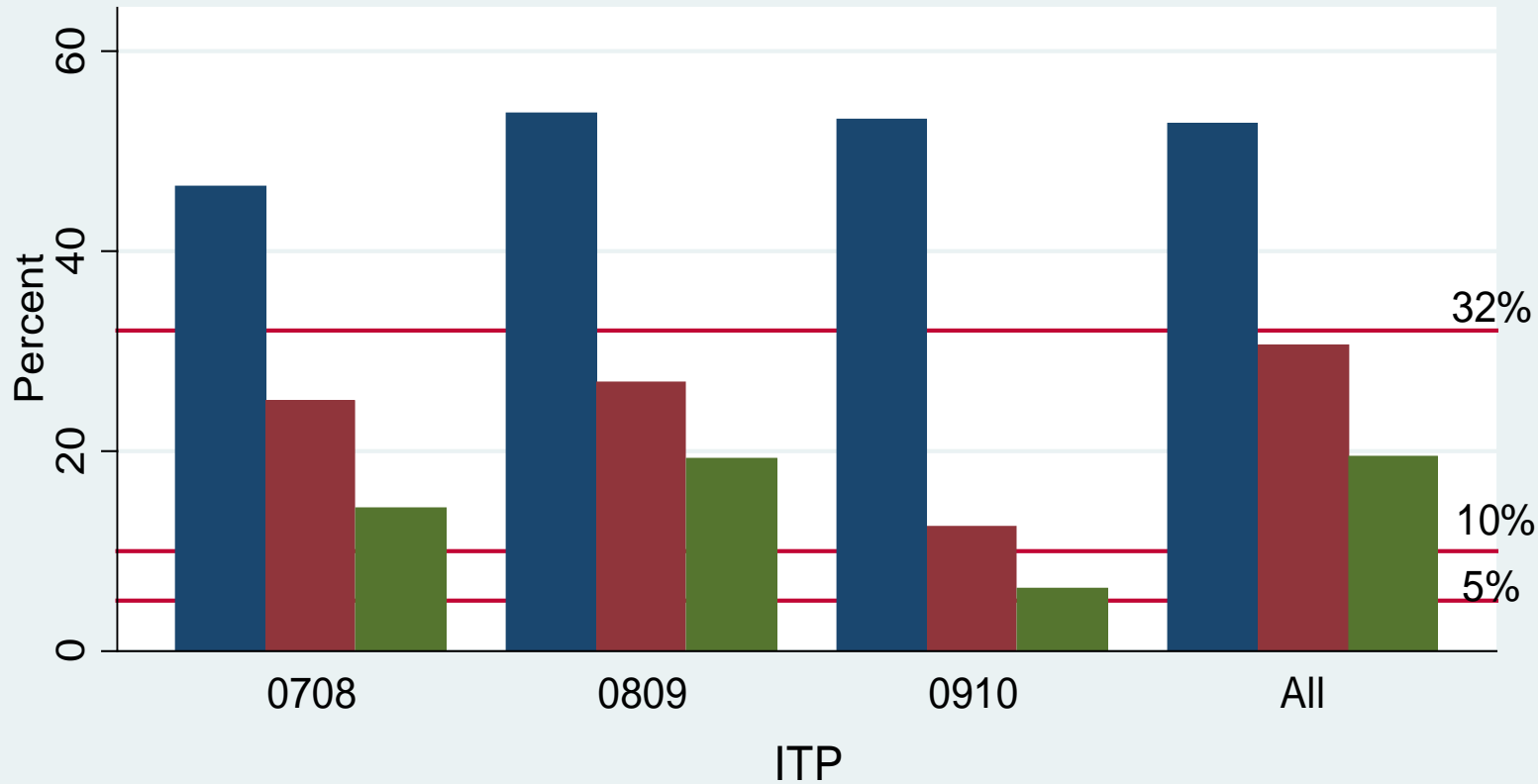
Data **U**se

- The next slide shows percent of programs classified as different from state for different confidence levels (different colored bars)

    For 2007-08 data, ITP programs are classified as follows:

    - ■ "High" level of confidence (green bar)
        - 14 percent of programs are classified as different from the mean
        - 5 percent expected by chance - see horizontal reference line

    - ■ "Medium" level of confidence (red bar)
        - 25 percent of programs are classified as different from the mean
        - 10 percent expected by chance - see horizontal reference line

    - ■ "Low" level of confidence (blue bar)
        - 46 percent of programs are classified as different from the mean
        - 32 percent expected by chance - see horizontal reference line

AIR®

Data
Use

Percent of ITP Programs Different from Mean by Year
Bars Represent Confidence Levels Used
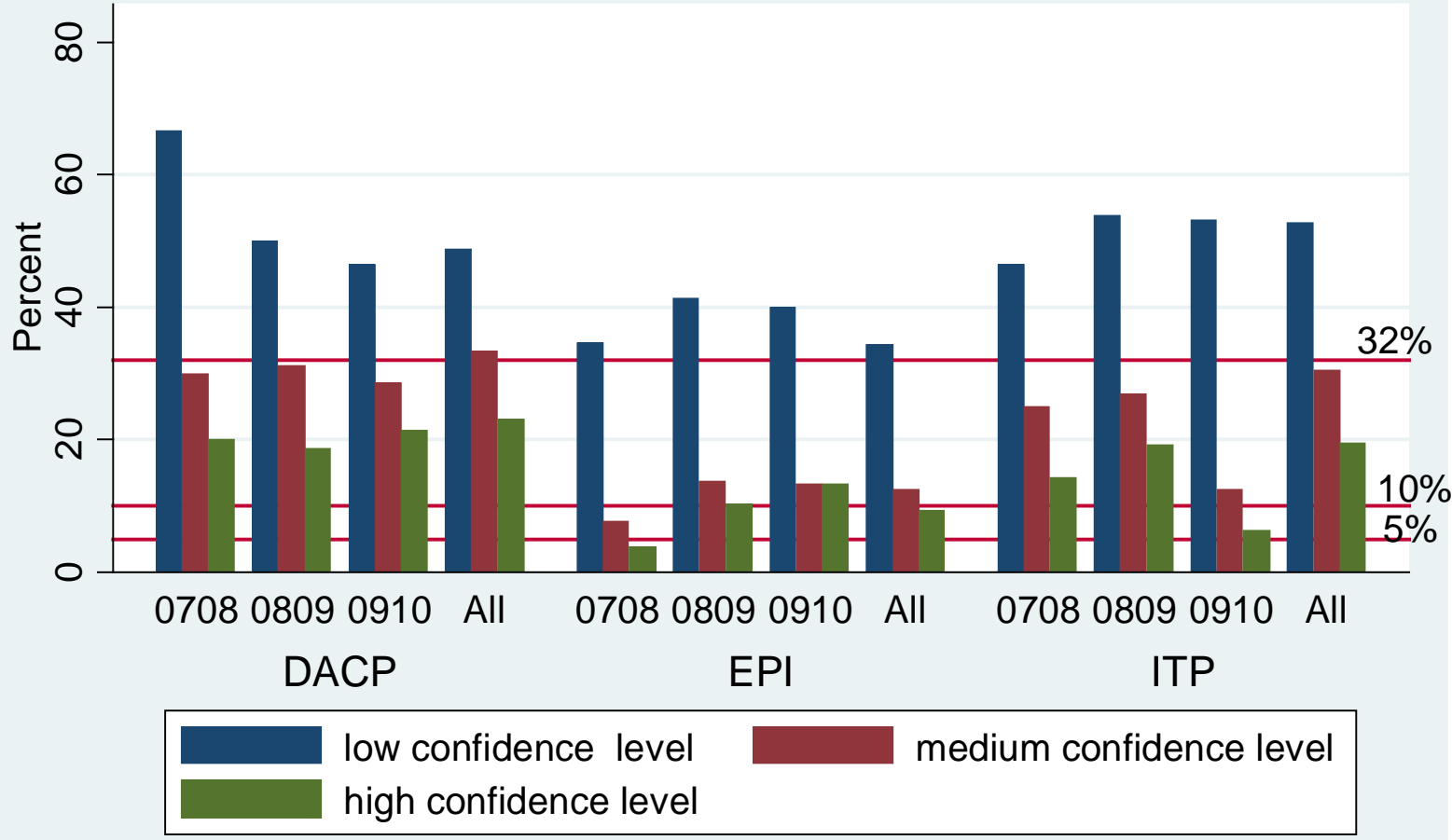

ITP

- low confidence level
- medium confidence level
- high confidence level

Horizontal lines (5, 10, 32 percent) represent potential misclassification

AIR®

Percent of Programs Different from Mean by Year
Bars Represent Confidence Levels Used

Horizontal lines (5, 10, 32 percent) represent potential misclassification

Data
Use

Classification of Other Programs

The next slides show DACP, EPI, and ITP programs in same format as previous chart

- In general, fewer EPI programs are classified as different from state mean and are less likely to be accurately classified
  - Percent of EPI programs different from mean is closer to level expected by chance for most years and precision levels - may reflect small program size

- For DACP and ITP programs, percent of programs identified as different from the mean is above expected misclassification rates (horizontal reference lines)

- Generally appears that combining several years of data strengthens estimates (see "All" results)

AIR®

Frequency of Programs Identified as Different from State Mean Across Years

- Percent of programs identified as different from the state mean *at least once in three years*
    - High confidence = 30%
    - Medium confidence = 42%
    - Low confidence = 78%

- Percentage of programs that are different from the state mean *for two of three years*
    - High confidence = 5%
    - Medium confidence   = 10%
    - Low confidence = 29%

- Only 6 percent of programs show up as different *for three years* but only with a low level of confidence

AIR®

Description of Data Handout

- Data on overall VAM scores are given by program and by year, along with VAM combined across years

  - Confidence of estimate relative to mean is noted in terms of four categories: high, medium, low, and very low

- Data sorted by number of completers and then confidence

- Data can be more precise for some individual years than in combination

- Data can be more precise in combination than in individual years as programs have more completers

AIR®

Data **U**se

- How much should VAM scores be used in program or institution approval?
  - Note: Inclusion may be dependent on number of completers with VAM scores in program and share of total program completers with scores
- Use only one-year completers or longer-term completers?
- Confidence levels to use?
- Use single-year or combined-year VAM data?
- Consider characteristics of schools where completers teach?
- What other data and analyses would inform decisions?

AIR®

C_{onclude}

Student Growth Materials

Information about the activities, membership, meeting schedule, materials, recording of conference calls, and webinars of the Student Growth Implementation Committee (SGIC), are posted at: http://www.fldoe.org/committees/sg.asp.

Contact Information

FLDOE:

Juan Copa, Director of Research and Analysis in Educator Performance

850-245-0744 (office)

Juan.Copa@fldoe.org

AIR:

Christy Hovanetz, Ph.D., Project Director

850-212-0243 (cell)

ChristyHovanetz@gmail.com