

# How Much Teacher is in Teacher Rating Scales?

D. Betsy McCoach & Del Siegle  
University of Connecticut

The following individuals contributed to this work: D. Betsy McCoach, Scott J. Peters, Anthony J. Gambino, Daniel A. Long, and Del Siegle

**FUNDED BY THE INSTITUTE OF EDUCATION SCIENCES, U.S. DEPARTMENT OF EDUCATION #R305C200012 AND #R305C140018**

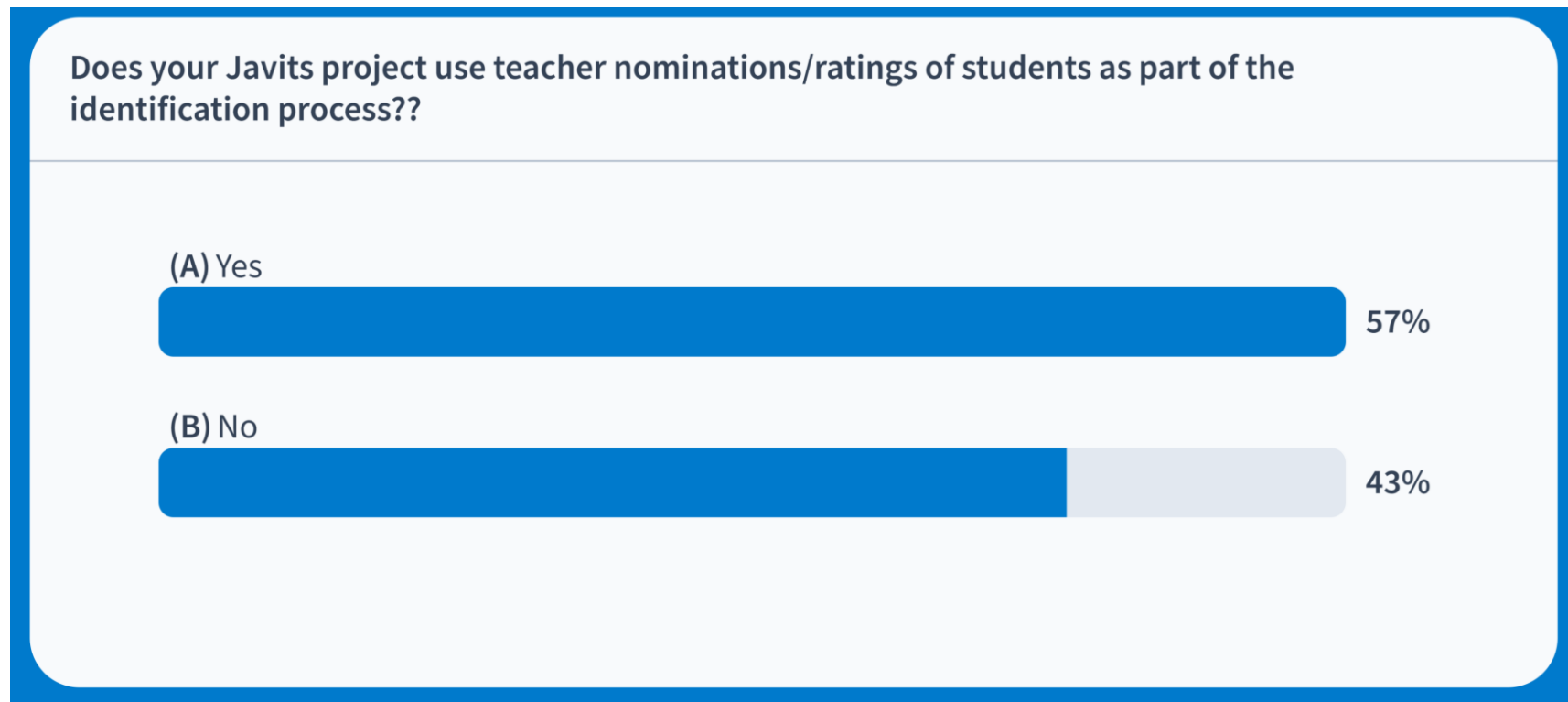
NATIONAL  
CENTER  
FOR  
RESEARCH  
ON  
GIFTED  
EDUCATION

<https://ncrge.uconn.edu>

# U.S. Department of Education Disclaimer

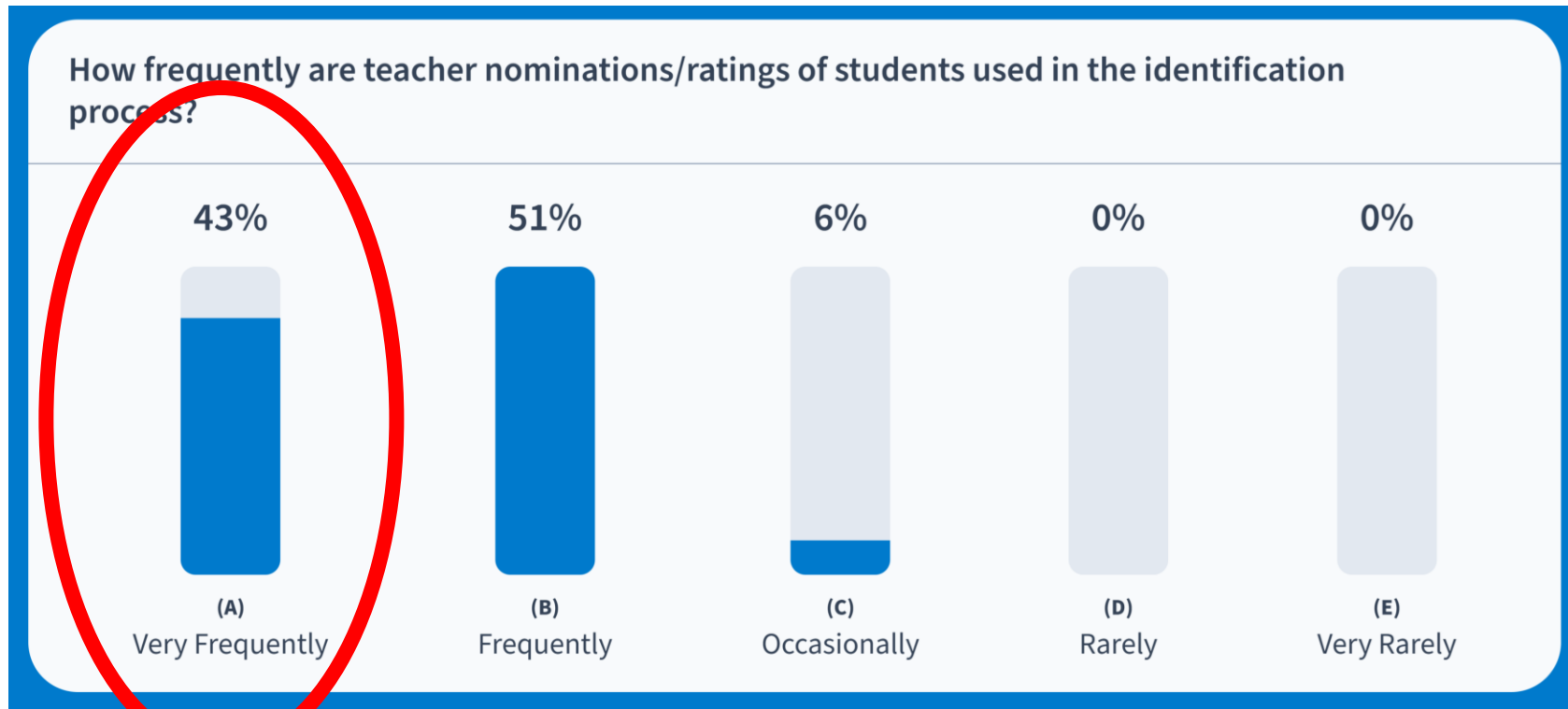
This presentation contains examples and resource materials that are provided for the user's convenience. The inclusion of any material is not intended to reflect its importance, nor is it intended to endorse any views expressed, or products or services. These materials may contain the views and recommendations of the presenter as well as hypertext links, contact addresses and websites to information created and maintained by other public and private organizations. The opinions expressed in any of these materials do not necessarily reflect the positions or policies of the U.S. Department of Education. The U.S. Department of Education does not control or guarantee the accuracy, relevance, timeliness, or completeness of any outside information included in these materials. Mentions of specific programs or products in these examples are designed to provide clearer understanding and are not meant as endorsements.

# Does your Javits project use teacher nominations/ratings of students?



# How frequently are teacher nominations/ratings of students used in the identification process?

**Note: Graphs reflect Javits meeting participants' responses to poll questions during presentation. Item circled in red reflect findings from our study.**



# Very Frequently

- Callahan et al. (2014) reported that **over 86% of school districts used teacher nominations.**
- National Center for Research on Gifted Education (NCRGE) found that **over 90% of school districts used teacher nominations and/or TRS** to identify students for gifted services (Siegle et al., 2018).

# The Role of TRS in GT Identification

**GT Identification:** Concerns about students being overlooked (Gentry et al., 2022; McBee et al., 2016; Morgan et al., 2015) and long-documented inequities within the populations of students served (Hosp & Reschly, 2003; Peters et al., 2019).

**Multiple Measures:** Multiple criteria identification systems often include teacher nominations and/or teacher rating scales (TRS).

# The Role of TRS in GT Identification

- TRS feature prominently in many **multi-criteria identification** systems
- There is a **general belief** that teacher ratings add **unique information** to the identification process, **improve multi-criteria systems**, and provide a **more nuanced assessment** of giftedness that **should increase diversity** (Harradine et al., 2014; Peters & Gentry, 2010).

# What is the biggest concern in using teacher ratings of students?

**(one- or two-word response for word cloud)**

What is the biggest concern in using teacher ratings of students? (enter one or two words for a word cloud)



A word cloud visualization of responses to the question 'What is the biggest concern in using teacher ratings of students?'. The words are arranged in a roughly circular shape, with 'bias' being the largest and most central word. Other prominent words include 'behavior', 'misconceptions', 'reliability', 'consistency', 'precision', 'calibration', 'beyond', 'bias', 'understanding', 'variability', 'professional', 'training', 'critical', 'personality', 'clash', 'misunderstanding', 'under-the-radar', 'use', 'creative', 'prep', 'scale', 'gifted', 'implicit', 'lack', 'validity', 'impact', 'see', 'student', 'teacher', 'thinking', 'traits', 'time', 'opportunity', 'stereotypes', 'knowledge', 'inconsistent', 'ratings', 'teachers', and 'stereotypes'.

teachers reliability ratings stereotypes  
precision consistency inconsistent knowledge  
calibration beyond bias understanding variability  
opportunity student misconceptions  
traits time bias behavior impact professional  
thinking teacher behavior training  
scale gifted implicit lack validity critical  
prep creative use clash personality  
misunderstanding  
under-the-radar



# **Little research documents the between-teacher variance in teacher ratings or the consequences of such rater dependence.**

TRS feature prominently in many multi-criteria identification systems and are often advocated as one way to diversify the pool of students that are identified as gifted

- **PROS**

- Additional Information
- Equity?

- **CONS**

- Rater dependence
- Potential Bias?

Systems that rely on TRS must consider the ways in which **between-teacher inconsistency** can decrease precision and the degree to which teacher bias can **compromise the equity of identification decisions.**



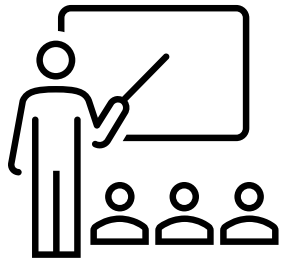
**Therefore, it is important to know...**

# How Much Teacher is in Teacher Rating Scales?

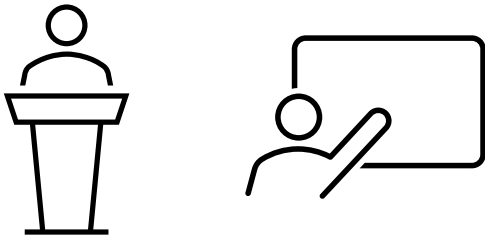
This research is part of an identification study being conducted by NCRGE. The identification study is one of four studies currently being conducted by the NCRGE .

NATIONAL  
CENTER  
FOR  
RESEARCH  
ON  
GIFTED  
EDUCATION

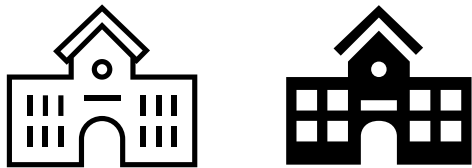
# Most Probable Sources of TRS Variance at Each of the Three Levels?



- Between-Student (Within-Teacher) Variance
  - Unconditional
  - Residual



- Between-Teacher (Within School) Variance
  - Unconditional
  - Residual



- Between-School Variance
  - Unconditional
  - Residual



# Overall Logic of the Study

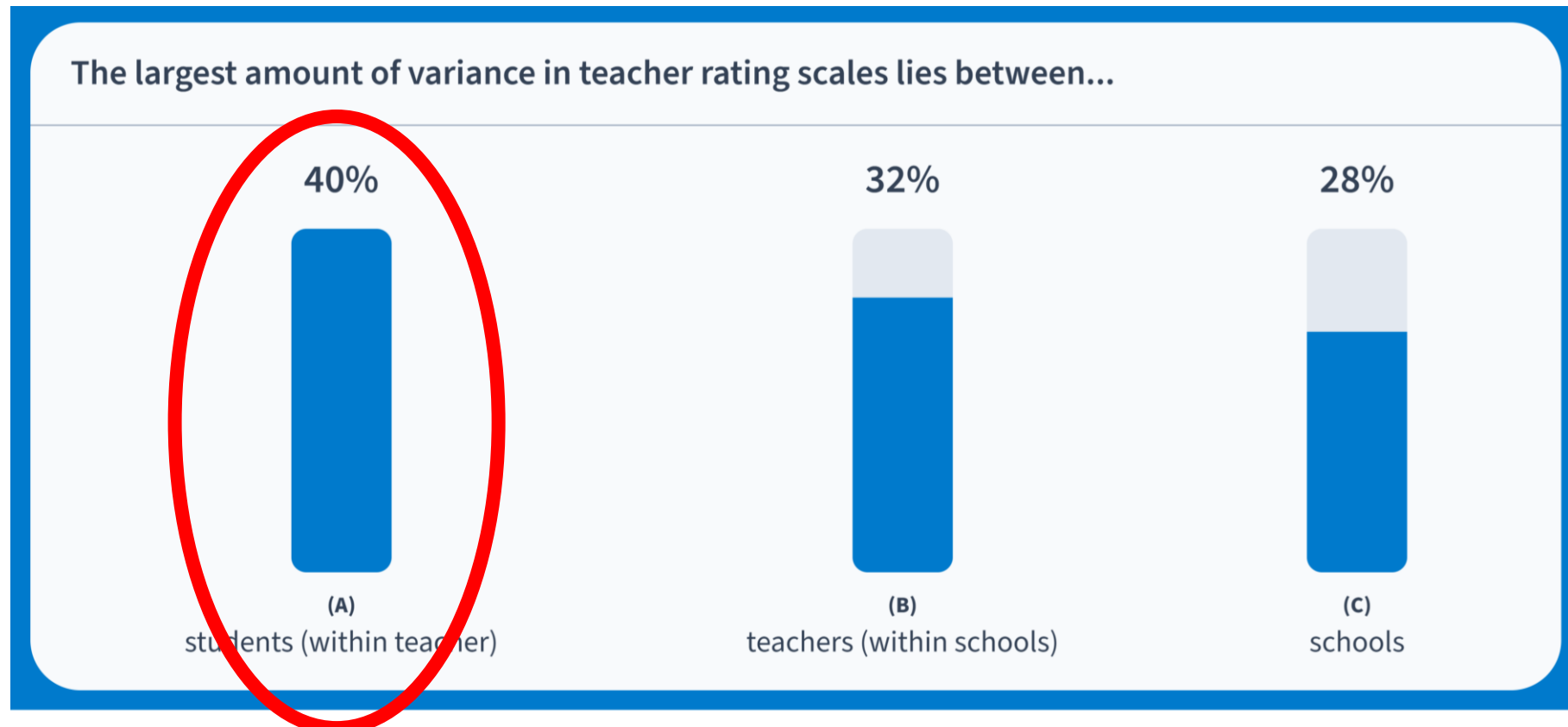
## Conditional Between-Teacher (Within-School) Variance:

If teachers use the TRS differently in ways that are unrelated to classroom-level differences in achievement and/or ability, between-teacher variance in the TRS would remain, even after controlling for ability and achievement.

**Unconditional Model:**  $\sigma_{Total}^2 = \sigma^2 + \tau_{00\pi} + \tau_{00\beta}$ .

The unconditional three-level model contains three orthogonal variance components: the **variance that lies between students within teachers** (level-1 variance,  $\sigma^2$ ), the **variance that lies between teachers within schools** (level-2 variance,  $\tau_{00\pi}$ ), and the **variance that lies between schools** (level-3 variance,  $\tau_{00\beta}$ ). The total variance in the TRS is the sum of these three variances in the unconditional model:  $\sigma_{Total}^2 = \sigma^2 + \tau_{00\pi} + \tau_{00\beta}$ .

# The largest amount of variance in teacher rating scales lies between...



**How much between-teacher variance is there in TRS? After controlling for ability, achievement, and demographics, how much between-teacher variance remains? How much is between-teacher variance likely to influence the screening and identification process?**

### **Hypotheses:**

1. ...a substantial proportion of the variance in TRS would lie between teachers (within schools)
2. ...a substantially greater proportion of the TRS variance would be between teachers (within schools), when compared to both achievement and ability scores
3. ...after controlling for ability, achievement, and student demographics, there would still be considerable between-teacher variance in TRS scores.



# Methods

- Examined the degree of **between-teacher variance** in gifted TRS, both before and after **controlling for cognitive ability and academic achievement**.
- Examined the degree to **which teachers' ratings were predicted by student demographics** such as race/ethnicity, EL and/or free-and-reduced lunch (FRL) status, **after controlling for ability and achievement**.
- Used the variance estimates from the results to estimate the degree to which teacher ratings and subsequent **identification decisions** are likely to be **influenced by which teacher completes the TRS**.

# Centering

- Group mean centered ability and achievement at level 1 (student level;  $X_{1i} = X_{ijk} - \bar{X}_{.jk}$ ), aggregated them within teachers and centered around the school mean at level 2 (teacher level;  $X_{2j} = \bar{X}_{.jk} - \bar{X}_{..k}$ ), and aggregated them within schools centered around the grand (sample) mean at level 3 (school level;  $X_{3k} = \bar{X}_{..k} - \bar{X}_{...}$ ). Demographic variables were centered and included at each level in the same way. (Hoffman & Walters, 2022; Raudenbush & Bryk, 2002, Rights, 2022; Rights & Sterba, 2023; Yaremych et al., 2022).

## Evaluating Between-Teacher Variance in the TRS

- Proportion of the total TRS variance that is residual between-teacher variance, after controlling for ability, achievement and student demographics using the following formula: 
$$\frac{\tau_{\pi 00(c)}}{\sigma_{(u)}^2 + \tau_{\pi 00(c)} + \tau_{\beta 00(u)}}$$

where  $\tau_{\pi 00(c)}$  is the between-teacher variance from the conditional model,  $\sigma_{(u)}^2$  is the between-student variance from the unconditional model,  $\tau_{\pi 00(u)}$  is the between-teacher variance from the unconditional model, and  $\tau_{\beta 00(u)}$  is the between-school variance from the unconditional model.

- Resembles the formula for the ICC. The denominator is identical to the denominator from the ICC; however, the numerator is the residual between-student variance from the conditional model.
- Residual between-teacher variance from the fully specified conditional model representing at least 10% of the total variance from the unconditional model represents a substantial amount of residual between-teacher variance.

**This slide was not included in the presentation and is provided for those who are interested in more details about the methodology.**

# 5 Models

1. Unconditional	$TRS_{ijk} = \gamma_{000} + e_{ijk} + r_{0jk} + u_{00k}$
2. Ability + Achievement	$TRS_{ijk} = \gamma_{000} + \gamma_{100}(\text{Ability}_{1i}) + \gamma_{200}(\text{Ach}_{1i}) + \gamma_{010}(\text{Ability}_{2j}) + \gamma_{020}(\text{Ach}_{2j}) + \gamma_{001}(\text{Ability}_{3k}) + \gamma_{002}(\text{Ach}_{3k}) + e_{ijk} + r_{0jk} + u_{00k}$
AA + Demographics	$TRS_{ijk} = \gamma_{000} + \sum_{p=1}^P \gamma_{p00} (\text{Var}_{p_{1i}}) + \sum_{p=1}^P \gamma_{0p0} (\text{Var}_{p_{2j}}) + \sum_{p=1}^P \gamma_{00p} (\text{Var}_{p_{3k}}) + e_{ijk} + r_{0jk} + u_{00k}$
	3. Includes Race/ethnicity
	4. Adds Low-Income status
	5. Adds EL and gender (if available)

This slide was not included in the presentation and is provided for those who are interested in more details about the methodology.

# How Much Is the Observed Teacher Variance Likely to Influence Identification Outcomes?

- Computed plausible values for TRS (Raudenbush & Bryk, 2002).
- Using  $\tau_{\pi 00}$  (the residual between-teacher variance in the rating scale) from the full MLM, estimate a range of likely values across teachers for student rating scale scores, given a particular prototypical profile.
  - $\gamma_{000} \pm \sqrt{\tau_{\pi 00}}$  provides a plausible range of TRS scores for average students with teachers who span the range from 1 SD below the average teacher on TRS (16<sup>th</sup> percentile) to 1 SD above the average teacher on TRS (84<sup>th</sup> percentile), and provides a 68% plausible interval;
  - $\gamma_{000} \pm 1.645\sqrt{\tau_{\pi 00}}$  provides a plausible range of TRS scores for average students with very low scoring (5<sup>th</sup> percentile) teachers to very high scoring teachers (95<sup>th</sup> percentile) and provides a 90% plausible interval.
- Indicates how much between-teacher differences contribute to variability in TRS and indicates how much TRS scores for a prototypical student might differ across teachers within the same school.
- Also- Cohen's  $d$  effect size - indicates the expected standard deviation unit change in students' TRS scores per between-teacher standard deviation.

**This slide was not included in the presentation and is provided for those who are interested in more details about the methodology.**

# **Are Teacher-level Variance in Ratings Likely to Substantively Change Identification Decisions?**

**To evaluate the degree to which between-teacher differences in TRS may undermine the overall identification process---**

- Used the lowest between-teacher effect size to create high and low TRS for each student.
- “Simulated” conditions in which the student’s TRS was completed by a teacher whose scores by a teacher whose scores were 1 SD lower than their current teacher.
- Computed the mean of ability, achievement, and TRS using the original and modified TRS scores.
- Determined the 90<sup>th</sup> percentile for the mean of ability, achievement, and TRS in the original dataset, and used this as the cut-off score for identification as gifted.
- Using that cut-off score, determined which students in the district were identified as gifted using the original and modified TRS scores, then calculated the proportion of overlap across two groups.

## **Overlap Between CogAT and TRS**

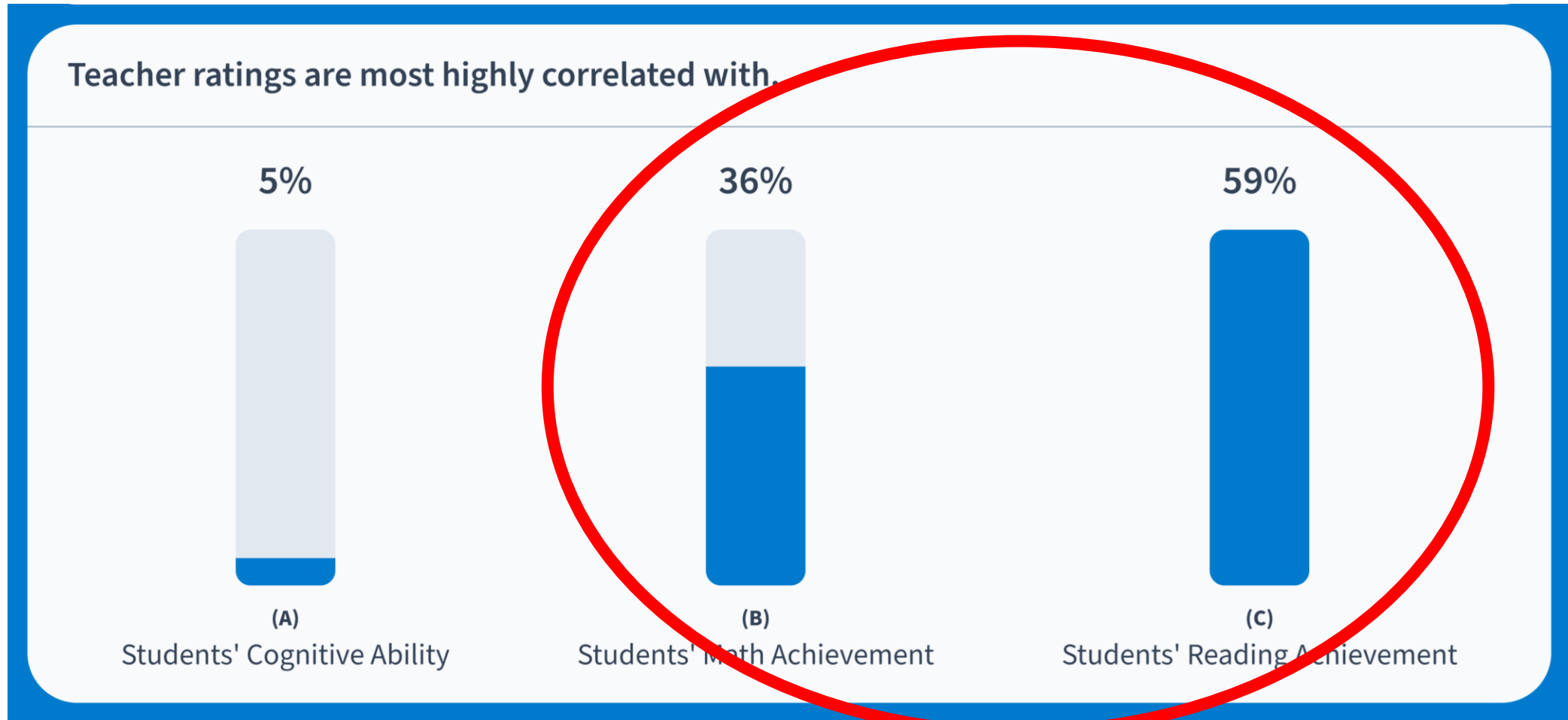
- In some districts, TRS are used as an initial screener to determine who should move to Phase 2 of the identification process.
- To estimate how many students with high cognitive ability would be missed if TRS were used as the Phase I screener, we computed the percentage of students who scored in the top 10% of the district on cognitive ability but did not score in the top 10, 20, 25, or 30% of the district on TRS.

# Assessments Used in Each of the Districts

District	Grade	School Year	Ability	Achievement	TRS
<b>C</b>	2 <sup>nd</sup>	2019-2020	CogAT	MAP Test	GRS
<b>H</b>	K	2018-2019	CogAT-Nonverbal	IOWA Test	District-made Scale
<b>M</b>	2 <sup>nd</sup>	2019-2020, 2021-2022, 2022-2023	CogAT	MAP Test	HOPE Scale
<b>O</b>	2 <sup>nd</sup> & 3 <sup>rd</sup>	2021-2022	InView	MAP Test	District-made Scale



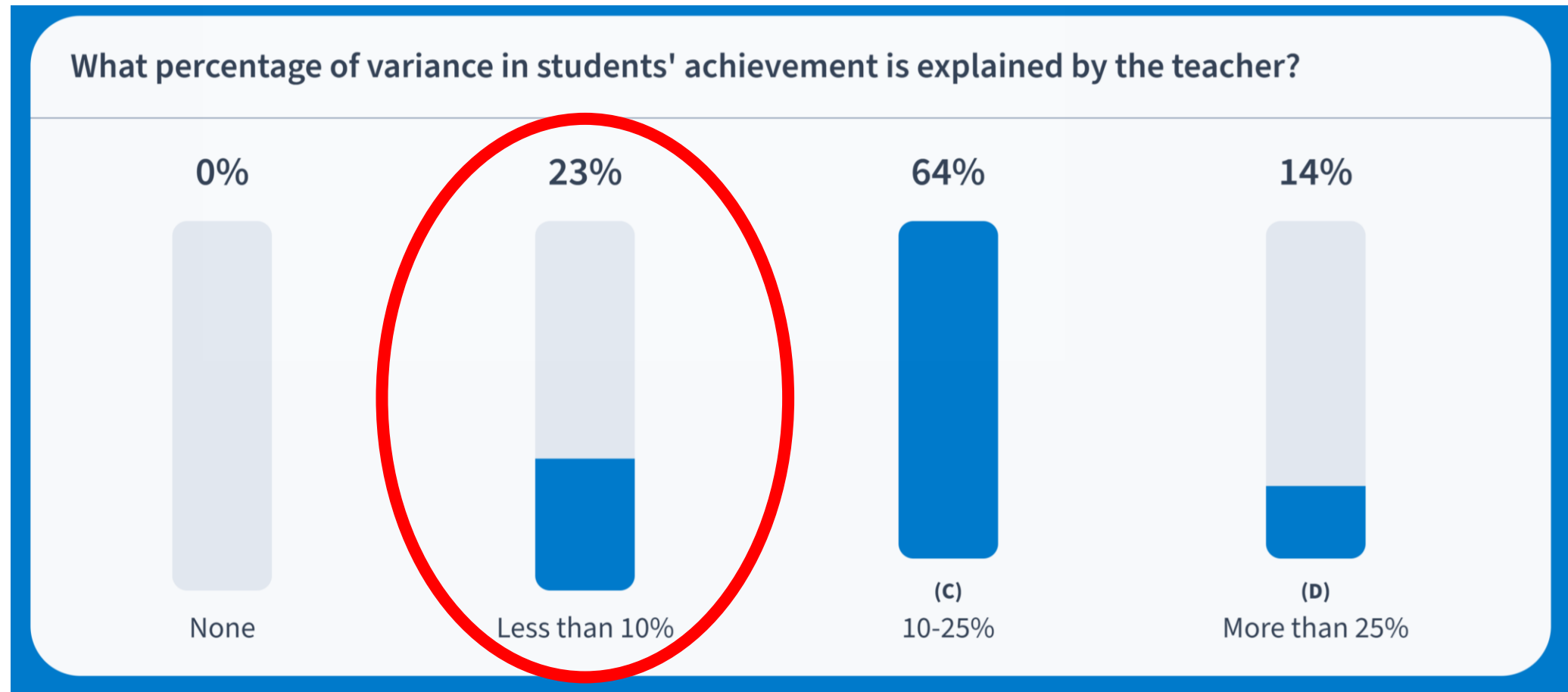
# Teacher ratings are most highly correlated with...



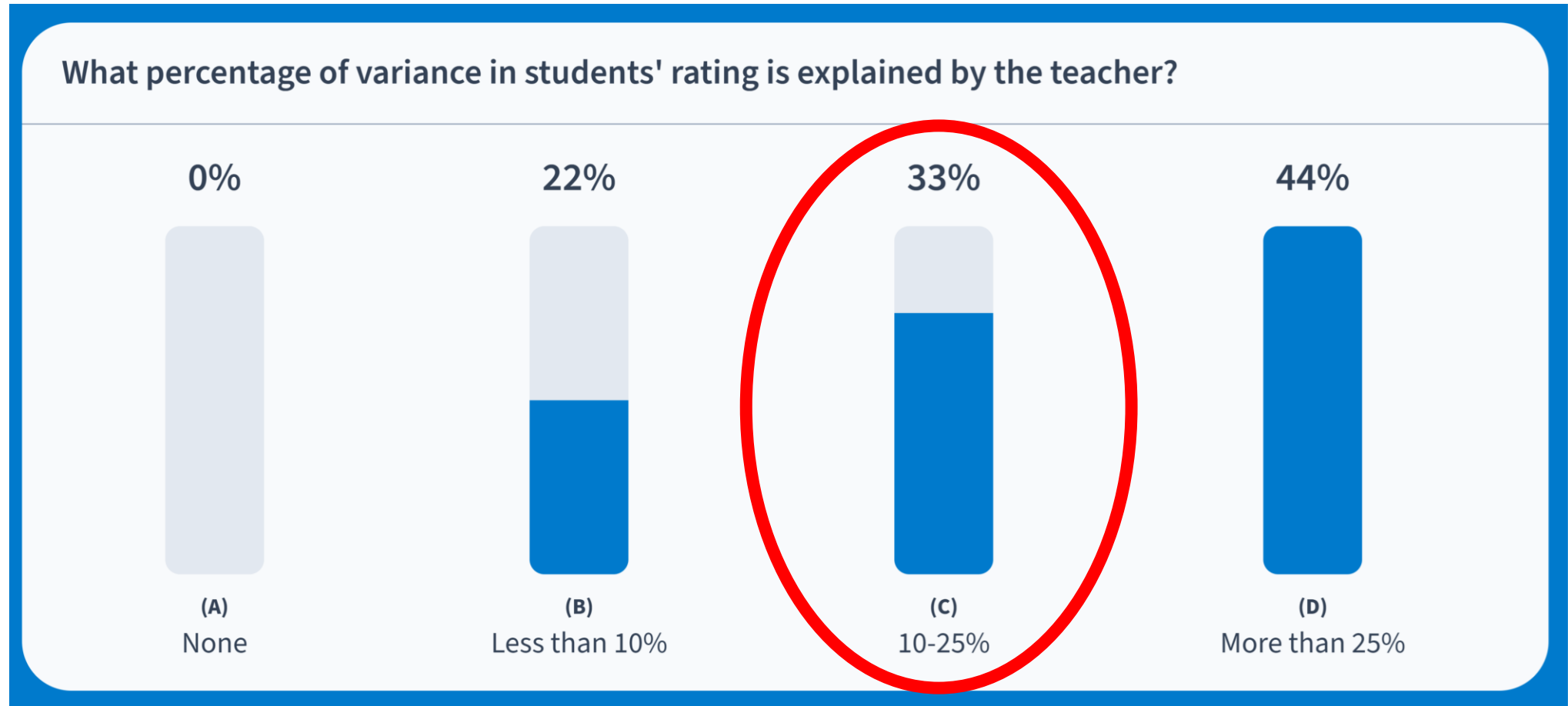
# Correlations among Assessments

District	Variable	TRS	Ability	Math
C	Ability	0.597	1	
	Math	0.697	0.747	1
	Reading	0.716	0.732	0.941
H	Ability	0.400	1	
	Math	0.500	0.550	1
	Reading	0.529	0.533	0.729
M1	Ability	0.518	1	
	Math	0.610	0.729	1
	Reading	0.581	0.647	0.742
M2	Ability	0.488	1	
	Math	0.577	0.730	1
	Reading	0.550	0.587	0.738
M3	Ability	0.526	1	
	Math	0.604	0.741	1
	Reading	0.597	0.605	0.759
O2	Ability	0.539	1	
	Math	0.558	0.855	1
	Reading	0.572	0.849	0.955
O3	Ability	0.428	1	
	Math	0.473	0.816	1
	Reading	0.479	0.815	0.949

What percentage of variance in **students' achievement** is explained by the teacher?



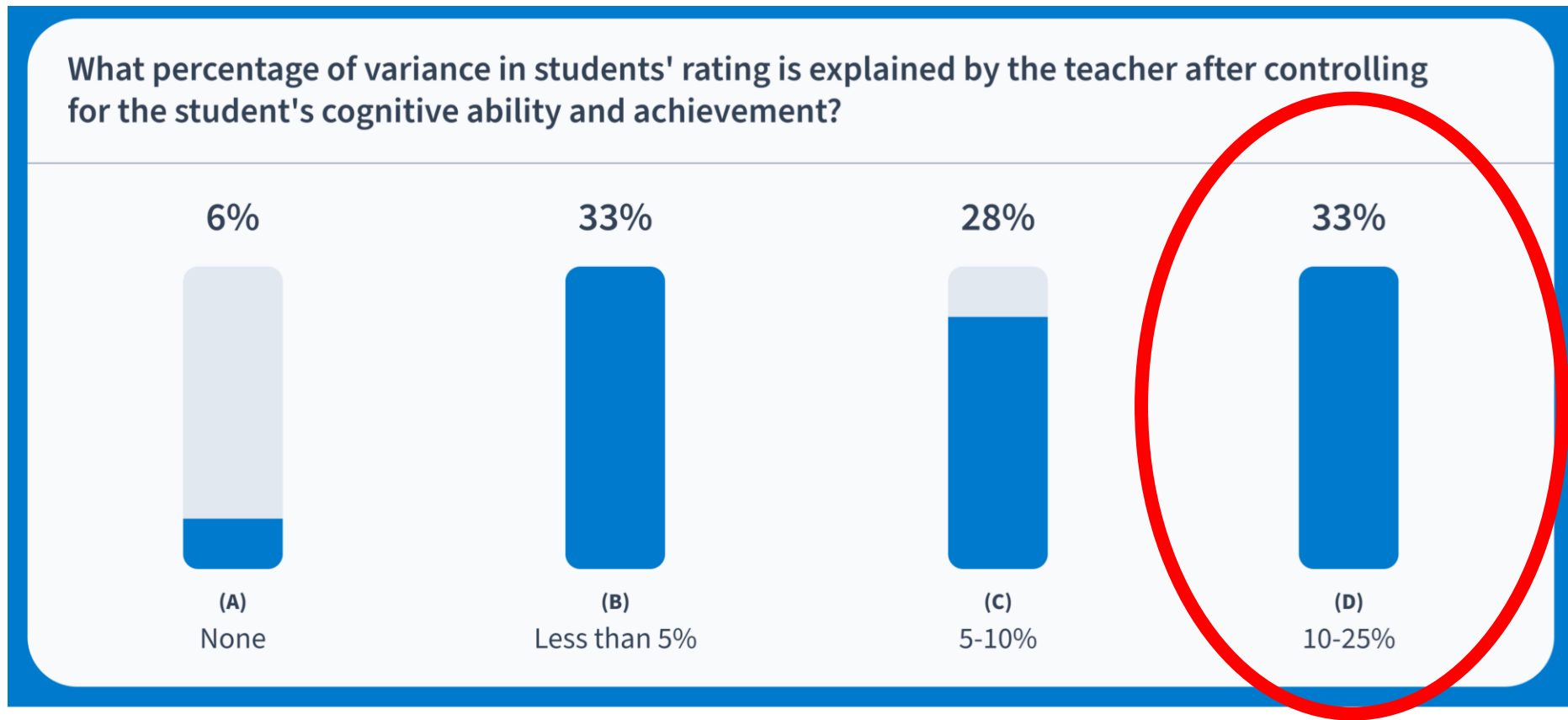
What percentage of variance in **students' rating (TRS)** is explained by the teacher?



# ICCs for Each Outcome by District

District	Level	TRS	Ability	Math	Reading
C	Student ( $N = 8,685$ )	0.809	0.780	0.729	0.723
	Teacher ( $J = 587$ )	0.104	0.023	0.042	0.038
	School ( $K = 109$ )	0.087	0.197	0.229	0.239
H	Student ( $N = 11,892$ )	0.617	0.811	0.766	0.763
	Teacher ( $J = 1,013$ )	0.246	0.070	0.120	0.117
	School ( $K = 166$ )	0.137	0.119	0.114	0.120
M1	Student ( $N = 2,036$ )	0.778	0.859	0.914	0.920
	Teacher ( $J = 92$ )	0.222	0.063	0.045	0.050
	School ( $K = 19$ )	0.000	0.078	0.041	0.030
M2	Student ( $N = 1,859$ )	0.751	0.942	0.922	0.950
	Teacher ( $J = 90$ )	0.249	0.006	0.006	0.018
	School ( $K = 19$ )	0.000	0.052	0.072	0.032
M3	Student ( $N = 1,832$ )	0.866	0.942	0.935	0.969
	Teacher ( $J = 89$ )	0.119	0.000	0.002	0.000
	School ( $K = 20$ )	0.015	0.058	0.063	0.031
O2	Student ( $N = 2,618$ )	0.803	0.846	0.836	0.838
	Teacher ( $J = 171$ )	0.137	0.018	0.020	0.024
	School ( $K = 60$ )	0.060	0.136	0.144	0.138
O3	Student ( $N = 2,176$ )	0.797	0.856	0.818	0.817
	Teacher ( $J = 153$ )	0.160	0.004	0.003	0.004
	School ( $K = 56$ )	0.043	0.140	0.179	0.179

What percentage of variance in **students' rating** is explained by the teacher **after** controlling for the student's cognitive ability and achievement?

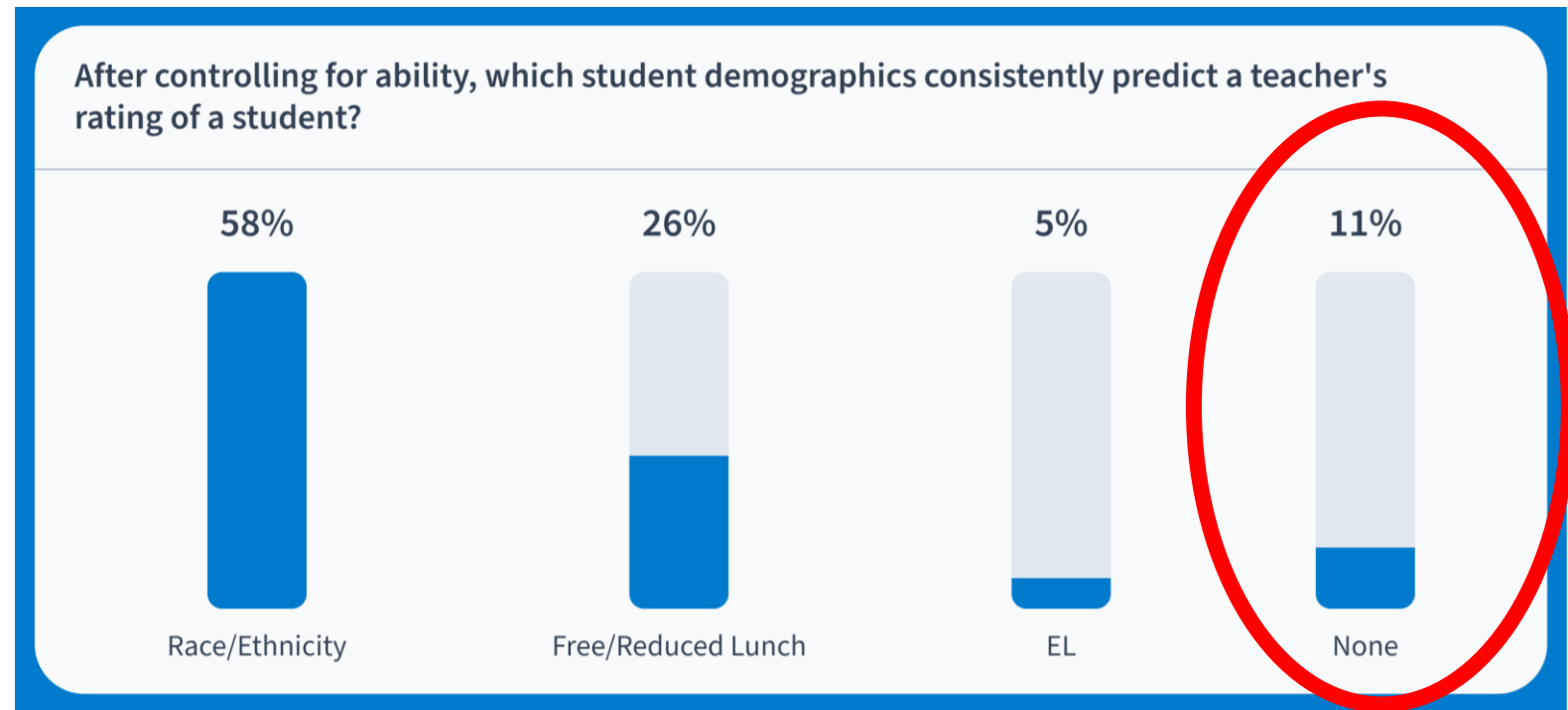


## Proportion of Teacher Rating Scale Variance that was Unexplained Between Teacher Variance across Models Using Group Mean Centering Strategy and 6-Category Race

District	Model 1*	Model 2**	Model 3a***	Model 3b****	Model 3c*****
C	0.104	0.106	0.106	0.106	0.106
H	0.246	0.242	0.239	0.239	0.239
M1	0.222	0.196	0.186	0.186	0.192
M2	0.249	0.231	0.235	0.235	0.224
M3	0.119	0.133	0.137	0.137	0.132
O2	0.137	0.156	0.142	0.142	0.143
O3	0.160	0.167	0.144	0.144	0.140

\*unconditional model; \*\*controlling for ability and achievement; \*\*\* additionally controlling for race/ethnicity;  
 \*\*\*\* additionally controlling for free/reduced lunch; \*\*\*\*\* additionally controlling for gender and EL

After controlling for ability, which student demographics consistently predict a teacher's rating of a student?





**Cohen’s d Effect Sizes for Statistically Significant Demographic Predictors and the Proportion of Level-1 Total (Residual) Variance Explained by All Included Demographics**

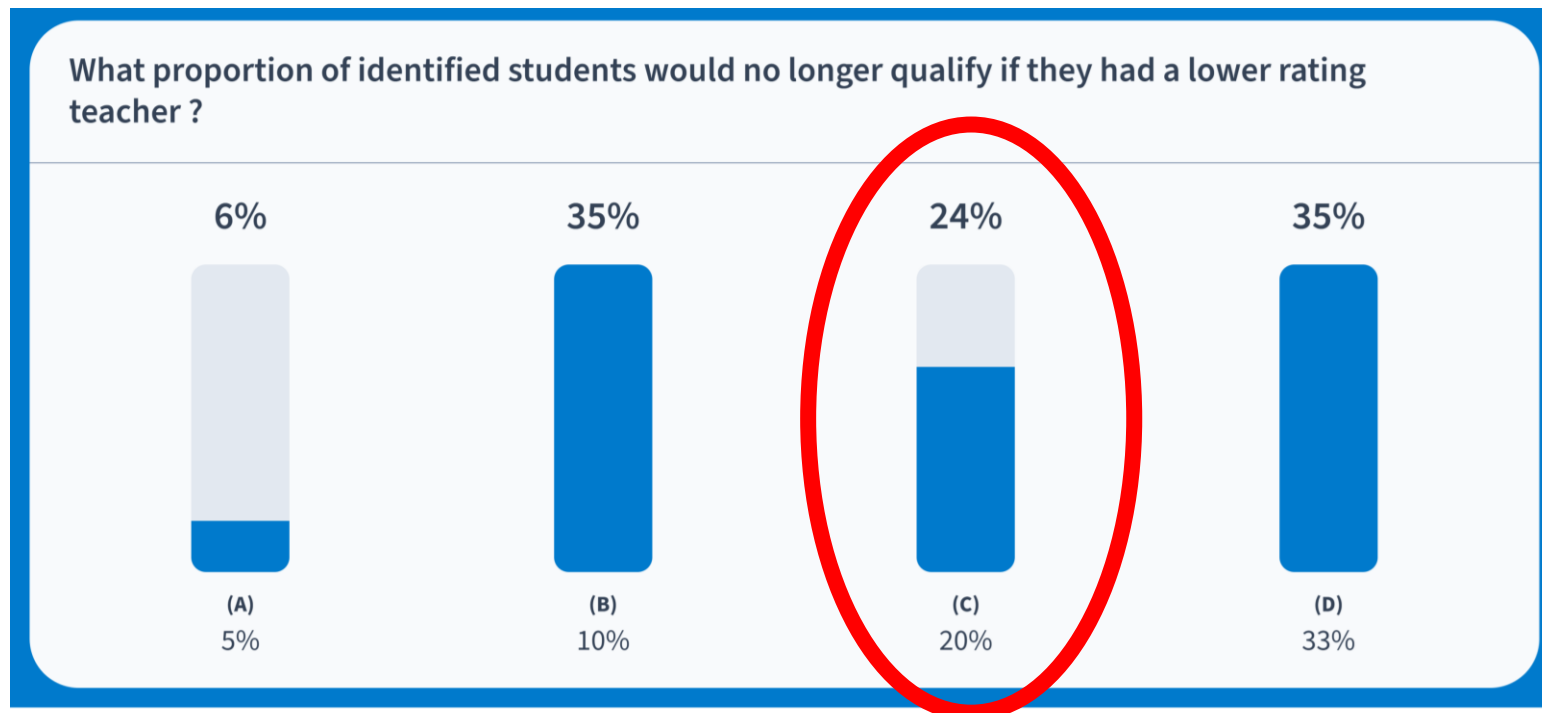
District	Effect Size (Demographics)	Percentage of Residual (Total) Level-1 Variance Explained
C	-0.11 (Black)	0.52% (0.21%)
H	0.08 (Black) <sup>a</sup> , -0.11 (FRL)	0.35% (0.21%)
M1	-0.17 (Asian)	1.90% (0.80%)
M2	-0.14 (EL) <sup>a</sup>	0.75% (0.37%)
M3	N/A	1.06% (0.54%)
O2	-0.12 (FRL), -0.18 (EL), 0.13 (Gender)	1.16% (0.60%)
O3	0.15 (Latinx), -0.14 (FRL), 0.11 (Gender)	1.20% (0.80%)

## TRS Mean and SD, Between-Teacher SD, Teacher Effect Size, and 68% Plausible Values for an Average Student as a Function of Teacher

District	TRS Mean	TRS SD	Between-Teacher SD	Effect Size (Teacher)	68% Plausible Values
C	64.65	20.76	6.75	0.33	[57.90, 71.45]
H	57.52	25.09	12.39	0.49	[45.13, 69.91]
M1	35.33	10.94	4.77	0.44	[30.59, 40.33]
M2	35.41	11.41	5.63	0.49	[30.09, 41.36]
M3	25.34	11.05	5.54	0.50	[30.14, 41.23]
O2	3.28	2.80	1.07	0.38	[2.21, 4.34]
O3	2.87	2.76	1.06	0.38	[1.85, 3.97]

# How much does this actually matter?

Imagine a school uses the mean of achievement, ability, and teacher rating scale to identify students and they identify the top 10%. What proportion of identified students would no longer qualify if they had a lower rating teacher (i.e.,- teacher is one SD lower than their current teacher, meaning their score is .33 SD lower than their current score--- but going into a system where we take the mean of TRS, achievement, and ability)?

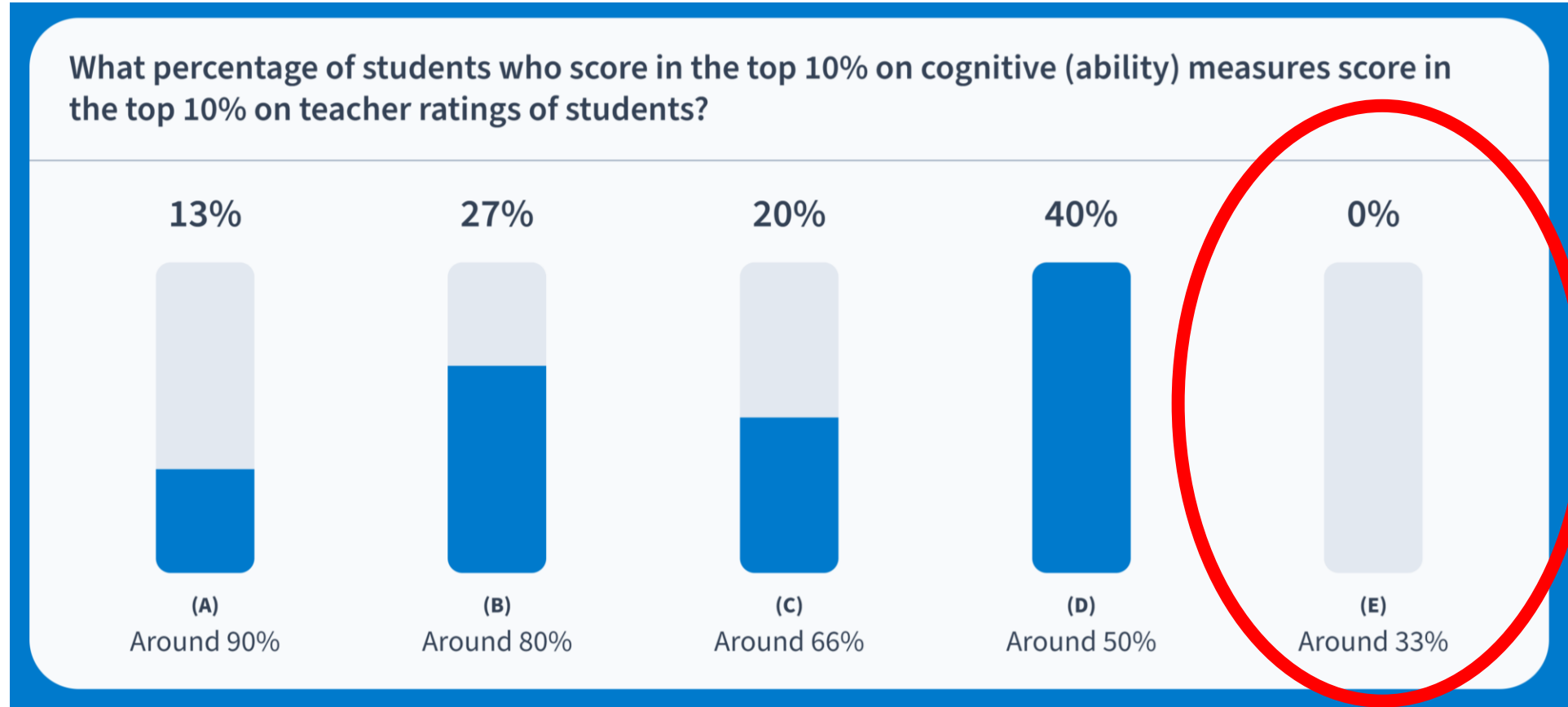


## Comparison of Identified Students when Students' TRS is Decreased by 0.33 SD Units

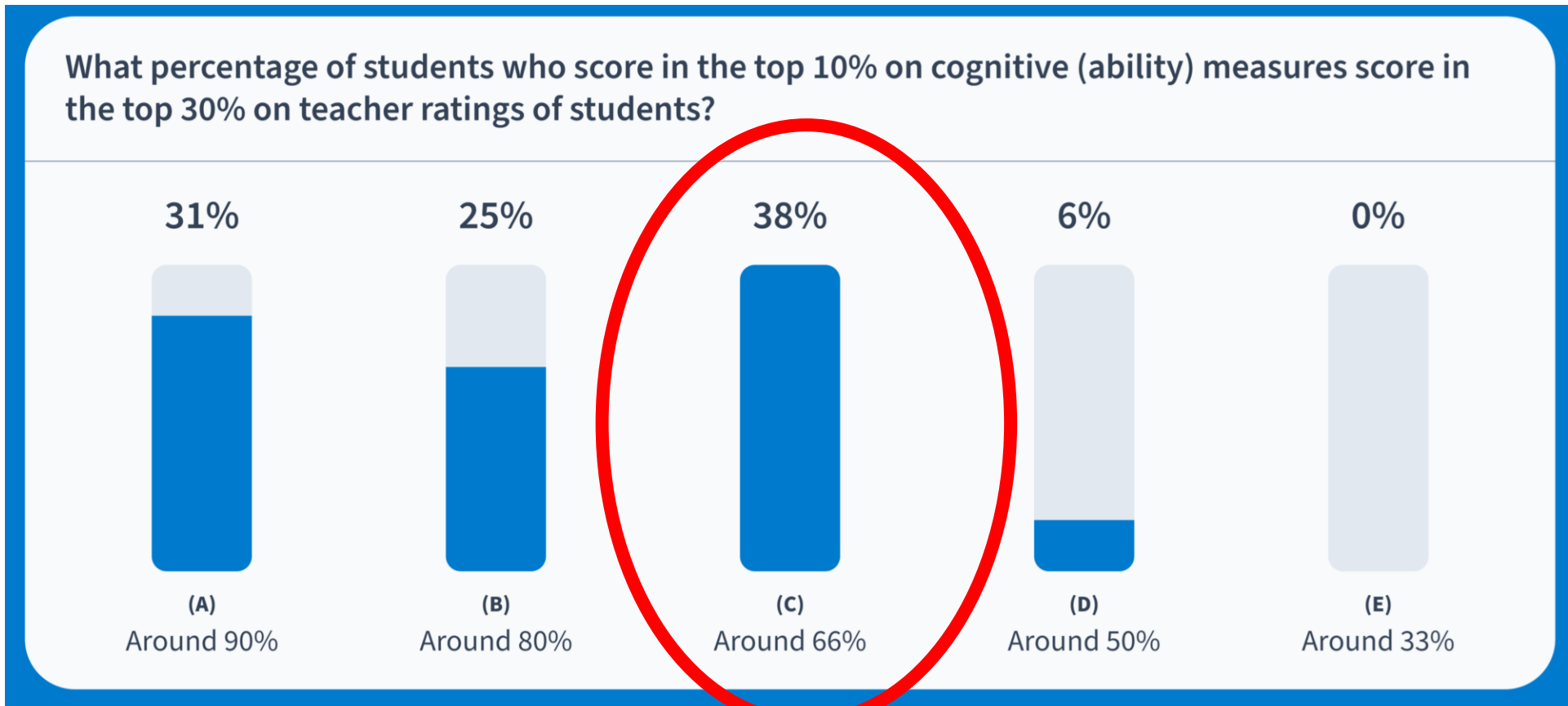
District	Still ID	Not ID	Current	% Not ID
C	708	163	871	18.71%
H	950	272	1,222	22.26%
M1	164	43	207	20.77%
M2	157	31	188	16.49%
M3	145	43	188	22.87%
O2	209	53	262	20.23%
O3	178	40	218	18.35%
Total	2,511	645	3,156	20.44%

*Note.* The Current column contains the number of students who would currently be identified if the district were to identify the top 10% of students on the mean of ability, achievement, and TRS. The Still ID column is the number of students who would still be identified if their TRS were decreased by 0.33 SD units. The Not ID column contains the number of students who would no longer be identified if their TRS were decreased by 0.33 SD units.

What percentage of students who score in the top 10% on cognitive (ability) measures score in the top 10% on teacher ratings of students?



What percentage of students who score in the top 10% on cognitive (ability) measures score in the top 30% on teacher ratings of students?



**Percentage of Students Who Are in the Top 10% of Their Districts on Cognitive Ability Who Score in the Top 10, 20, 25, and 30% of Their Districts on the TRS**

District	Top 10% TRS	Top 20% TRS	Top 25% TRS	Top 30% TRS
<b>C</b>	39.11%	64.56%	72.56%	78.33%
<b>H</b>	26.50%	51.54%	57.94%	64.34%
<b>M1</b>	35.41%	58.37%	67.46%	74.64%
<b>M2</b>	35.00%	51.67%	58.89%	71.67%
<b>M3</b>	36.22%	55.14%	62.70%	71.89%
<b>O2</b>	36.26%	62.60%	72.52%	78.63%
<b>O3</b>	24.77%	42.66%	49.08%	54.13%
<b>Overall</b>	32.37%	56.18%	63.61%	70.29%

**Teacher Variance in the TRS represents both a reliability and a validity issue.**

- Reliability
- Validity
- Accuracy





## Recommendations for Districts

1

**Never use TRS as the sole universal screening instrument to determine which students move forward to a second stage gifted identification process.**

- Less than 1/3 of students who scored in the top 10% on the Ability measure also scored in the top 10% on the TRS.
- Even with a lenient TRS cut score, almost 30% of students who were in the top 10% on ability did not score in the top 30% on TRS. (And in some datasets, almost half of students who scored in the top 10% on ability were not in the top 30% on TRS.)

# Recommendations for Districts

## 2

Consider the proportion of between-teacher variance as an additional source of error.

- **Reported reliability estimates do not take between-teacher variance into account.** In fact, between-teacher variance in TRS is likely to artificially increase the reported reliability coefficient for TRS.
- In our study, the GRS had the lowest between-teacher variance.
- However, future researchers should examine whether this result is replicated in other datasets.
- In addition, **future research on TRS should always report the proportion of between-teacher variance on the TRS**, as this is an important, but generally overlooked aspect of its psychometric adequacy.

## Recommendations for Districts

3

**Provide frequent professional learning** for teachers to try to standardize TRS usage as much as possible. This should help to decrease the proportion of between-teacher variance.

- Talk explicitly about how you would like teachers to interpret the response scale.
- Provide a handout that details these response scale interpretations.
- Frequent probably means yearly! The training can be short.
- Think of this as “tuning”

## Recommendations for Districts

# 4

Reflect on why are you including the TRS.

- Be **clear about the purpose** of including a TRS in the identification process
- Examine the TRS to **ensure that it is designed to elicit the kind of information you seek.**

# **Can we just use teacher norms / center within teacher?**

## **Three problems with teacher norms:**

1. The very small number of students per teacher makes such a process unstable/unreliable.
2. Using teacher centered TRS does not control for ability and achievement, and there are real between-teacher differences in class composition. The combination of the within-class sample size and the between-class differences in classroom composition could make classroom-normed TRS equally problematic.
3. Teacher centering the TRS does nothing to address the issue of between-teacher variability in between-student variance on the TRS.

## **Limitations of this Research**

- 1. Lack of item-level data**
- 2. Lack of detailed school-level data on such as the assignment of students to classes**
- 3. Lack of teacher-level covariates that could help to explain between-teacher variability in the use of the TRS.**
  - a) Future research should explore:**
  - b) whether certain teacher characteristics help to explain the between-teacher variance in TRS**
  - c) the degree to which professional development on the TRS can decrease the between-teacher variance on TRS.**

# Take home message...

Teacher Ratings of Students (TRS) should NOT be used as a sole, first stage identification screen to determine who should be tested for gifted services. They are not necessarily bad. However, 1) teachers must be frequently trained on how to explicitly interpret the scales and 2) the scales must clearly align with the purpose of identifying students for gifted services.



<https://ncrge.uconn.edu>