Are Novice Teacher and Novice Teacher Supervisor Survey Performance Ratings Comparable? Implications for Teacher Preparation Program Accountability Policies.

Matthew Finster

Westat®

Introduction

Despite calls for increased accountability on teacher preparation institutions or programs (TPPs), there is little consensus about which measures should be used to evaluate them. Teacher and supervisor surveys can be used for program improvement and for program accountability (Feur, Floden, Chudowsky, & Ahn, 2013). The regulations to implement requirements for the TPP accountability system under Title II of the Higher Education Act of 1965 (HEA), as amended, and which were rescinded in 2017, would have required state education agencies (SEAs) to report on performance indicators. In part, these reports would be based on novice teacher and employer surveys designed to capture perceptions of whether novice teachers have the academic content knowledge and teaching skills needed for success. To date, there are no studies that I am aware of that have compared survey data about novice teachers' perceptions of their preparation for teaching to supervisors' perceptions of their preparation for teaching and examined the implications for accountability policies, such as ranking TPPs.

Questions and answers

Q1 How closely related are novice teachers' perceptions of their preparation and their supervisors' perceptions of their preparation?

Overall, there is low agreement between novice teachers' and their supervisors' survey responses.

- Inter-rater statistics indicate low agreement: absolute agreement percentage is 41%, Cohen's Kappa values range from -.02 to .134, intra-class correlation coefficient is .102, and the correlation is .107.[1]
- The multi-group SEM indicates item intercepts are noninvariant (i.e., nonequivalent); items Q9, Q11, Q12, Q14, and Q16 are particularly problematic.

Q2 How do rankings of teacher preparation institutions change if supervisors' responses are used instead of teachers' responses?

There is low agreement (25% agreement) between TPP quartile rankings based on teachers' vs. supervisors' ratings.

Q3 How much of the variation in novice teacher' and their supervisors' ratings lies within versus between institutions and schools?

For teachers' rating of themselves, there is more variation within cases rather than across institutions. However, for supervisors' ratings of novice teachers, a large percentage (53%) of variation is explained by school (i.e., supervisor) grouping.

[1] Cohen's Kappa and intra-class correlation estimates likely are influenced by unequal distribution of ratings. Intra-class correlation coefficient and correlation values are based on average teacher and supervisor summative ratings.

Data

- Original survey data from novice teachers and their supervisors in Iowa state (Response rate: Teachers 26.7%, supervisors 31.0%. Total N = 558, matched responses n = 279).
- Survey based on InTASC standards and uses a 4-point Likert scale—*not very well* to *very well* (for more information, see Finster, 2017).

Exhibit 1. Novice Teacher Supervisor Survey

	Survey item	Response scale			
	How well does the novice teacher perform each of the following:	**Not very well**	**Somewhat well**	**Well**	**Very well**
Domain 1: The Learner and Learning	**1.1** Design and implement developmentally appropriate learning experiences for all learners. (Q1)				
	1.2 Ensure an inclusive learning environment for all learners. (Q2)				
	1.3 Develop and maintain a positive learning environment that engages all learners. (Q3)				
Domain 2: Content	**2.1** Demonstrate understanding of content area by using central concepts, tools of inquiry, and structures of your discipline. (Q4)				
	2.2 Make your discipline accessible and meaningful for learners. (Q5)				
	2.3 Integrate cross-disciplinary skills (e.g., critical thinking, problem solving, creativity, communication) to help learners use content. (Q6)				
Domain 3: Instructional Practice	**3.1** Develop and use multiple methods of assessment. (Q7)				
	3.2 Plan for instruction aligned to content standards. (Q8)				
	3.3 Use a variety of instructional strategies appropriately. (Q9)				
	3.4 Differentiate instruction for all learners. (Q10)				
	a. For students with disabilities. (Q11)				
	b. For English language learners. (Q12)				
	3.5 Use technology in the classroom appropriately to support instruction. (Q13)				
Domain 4: Professional Responsibility	**4.1** Engage in ongoing professional learning to provide all learners with engaging learning experiences. (Q14)				
	4.2 Evaluate outcomes of teaching using a variety of data (e.g., systematic observation, information about learners, research) to adapt planning and practice. (Q15)				
	4.3 Reflect on teaching practice to improve instruction. (Q16)				
	4.4 Work collaboratively with colleagues to meet the needs of all learners. (Q17)				
Comments and recommendations	In what areas, if any, do you think the preparation program should have prepared the novice teacher more effectively?	Open ended			
	What recommendations do you have for the novice teacher's preparation program?				
	Additional comments?				

Methods

- Descriptive and inter-rater statistics comparing novice teachers' and their supervisors' ratings.
- Multigroup structural equation modeling (SEM) to test for equivalence (or invariance) between novice teachers' and their supervisors' ratings.
- Multilevel modeling to examine variance within vs. between institutions and schools/supervisors.

Discussion

- This data indicates that novice teachers' and their supervisors' ratings of a teacher's preparation have low levels of agreement and are not equivalent or interchangeable measures. On average, supervisors provide slightly higher ratings than teachers themselves.
- Using teachers' or supervisors' ratings has substantial implications for which institutions are identified as high and low performing, with only some (25%) agreement across institution rankings based on quartiles.
- Results of the multilevel null models identity several problematic issues:
  1. **There is insufficient variance in teachers' and supervisors' ratings to model random intercepts by institution**, and teachers' and supervisors' ratings are not significantly nested by institutions. This finding is consistent with other research that demonstrates there is more variation in teacher effectiveness within TPPs than across TPPs (Koedel, Parsons, Podgursky, & Ehlert, 2012)[1], which is problematic for policies that require making meaningful distinctions between institutions or TPPs based on performance indicators, although others have detected significant differences between TPPs (Ronfeldt & Campbell, 2016).[2]
  2. **Supervisors' ratings of novice teachers are significantly nested by school (i.e., supervisor), indicating there is either a rating or hiring effect**, which may be problematic if the ratings are used as a measure of institutional performance.
  3. **The variance in teachers' ratings of their preparation is explained more so by their current school than by their institution of preparation.** This indicates that teachers' ratings may be influenced by their school environment (evaluation context). While this source of influence is well known in evaluation research (e.g., DeCotiis & Petit, 1978; Landy & Farr, 1980), it raises new concerns regarding potential sources of bias for survey ratings used to assess institutional performance.

[1] Based on value-added measures.
[2] Using teachers' observational ratings of program completers.

Results

Table 1. Descriptive Statistics of Novice Teachers' and their Supervisors' Survey Ratings on Preparation

	N	Minimum	Maximum	Mean	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Q1_S	279	1	4	3.23	0.72	0.52	-0.72	0.15	0.45	0.29
Q1_T	279	2	4	3.02	0.57	0.33	0.00	0.15	0.09	0.29
Q2_S	279	1	4	3.35	0.75	0.56	-0.98	0.15	0.51	0.29
Q2_T	279	2	4	3.10	0.61	0.38	-0.06	0.15	-0.36	0.29
Q3_S	279	1	4	3.35	0.73	0.54	-0.88	0.15	0.16	0.29
Q3_T	279	1	4	3.25	0.64	0.40	-0.35	0.15	-0.26	0.29
Q4_S	279	1	4	3.23	0.75	0.56	-0.60	0.15	-0.29	0.29
Q4_T	279	2	4	3.12	0.64	0.41	-0.11	0.15	-0.61	0.29
Q5_S	279	1	4	3.22	0.74	0.55	-0.70	0.15	0.19	0.29
Q5_T	279	2	4	3.05	0.63	0.40	-0.04	0.15	-0.47	0.29
Q6_S	279	1	4	3.05	0.80	0.63	-0.44	0.15	-0.44	0.29
Q6_T	279	1	4	2.95	0.69	0.48	0.01	0.15	-0.75	0.29
Q7_S	279	1	4	3.01	0.74	0.55	-0.24	0.15	-0.56	0.29
Q7_T	278	1	4	2.87	0.69	0.47	0.11	0.15	-0.72	0.29
Q8_S	278	1	4	3.28	0.72	0.51	-0.76	0.15	0.37	0.29
Q8_T	279	1	4	3.05	0.69	0.48	-0.20	0.15	-0.50	0.29
Q9_S	275	1	4	3.14	0.74	0.54	-0.44	0.15	-0.35	0.29
Q9_T	279	1	4	3.16	0.64	0.41	-0.23	0.15	-0.25	0.29
Q10_S	271	1	4	3.12	0.76	0.58	-0.36	0.15	-0.75	0.29
Q10_T	271	1	4	2.92	0.67	0.45	0.02	0.15	-0.58	0.29
Q11_S	276	1	4	3.14	0.79	0.62	-0.53	0.15	-0.44	0.29
Q11_T	277	1	4	2.73	0.76	0.58	0.25	0.15	-0.82	0.29
Q12_S	261	1	4	2.97	0.80	0.65	-0.39	0.15	-0.38	0.30
Q12_T	275	1	4	2.28	0.84	0.71	0.40	0.15	-0.34	0.29
Q13_S	279	1	4	3.18	0.77	0.59	-0.60	0.15	-0.23	0.29
Q13_T	278	2	4	3.07	0.73	0.53	-0.11	0.15	-1.11	0.29
Q14_S	278	1	4	3.40	0.67	0.45	-0.90	0.15	0.56	0.29
Q14_T	279	2	4	3.16	0.64	0.40	-0.15	0.15	-0.59	0.29
Q15_S	279	1	4	3.05	0.74	0.55	-0.35	0.15	-0.38	0.29
Q15_T	279	1	4	2.94	0.73	0.53	-0.13	0.15	-0.54	0.29
Q16_S	279	1	4	3.25	0.76	0.58	-0.85	0.15	0.45	0.29
Q16_T	279	1	4	3.36	0.61	0.37	-0.49	0.15	-0.12	0.29
Q17_S	278	1	4	3.50	0.72	0.52	-1.30	0.15	1.04	0.29
Q17_T	278	2	4	3.39	0.64	0.41	-0.57	0.15	-0.63	0.29

Note. N = 558, Supervisors n = 279, Teachers = 279, listwise N = 231.

Table 2. Fit Statistics of Multigroup Models (Configural and Invariant Models) for Novice Teachers and their Supervisors

Tests of Model fit	Configural Model	Invariant Model_v1.	Invariant Model_v2.
Rationale of test	Configural model is a multigroup representation of baseline models that tests invariance across the groups simultaneously and provides baseline values to compare further models	Assesses whether factor loadings are equivalent across groups	Assesses whether the intercepts are equivalent across groups
Number of free parameters	114	98	81
Chi-square test of model fit			
Value	535.98	577.31	701.24
Degrees of freedom	226	242	259
p-value	<.001	<.001	<.001
scaling correction factor	1.20	1.19	1.18
Δ Degrees of freedom	—	16	17
Difference Test Scaling Correction (CD)	—	0.92	1.10
Sattora-Bentler Scaled Chi-square Difference (TRd)	—	42.29	129.97
Δχ2 significance	—	<.001	<.001
Akaike (AIC)	16359.89	16366.89	16475.95
Bayesian (BIC)	16852.87	16790.68	16826.23
Bentler Comparative Fit Index (CFI)[1]	0.92	0.92	0.89
Tucker-Lewis Index (TLI)	0.91	0.91	0.88
Root Mean Square Error of Approximation (RMSEA)[2]	0.07	0.07	0.08
Standardized Root Mean Square Residual (SRMR)[3]	0.05	0.15	0.13

Notes. Chi-square Test of Model Fit for the Baseline Model, value = 4235.80, DF = 272, p < .001.* p < .05.
[1] Values 0.90 to 0.95 indicative of acceptable fit (Bentler, 1990). [2] Values < 0.08 indicative of adequate fit (Brown & Cudeck, 1993) and values 0.80 to 0.10 indicative of mediocre fit (MacCallcum et al., 1996). [3] Values < 0.05 indicative of well-fit model (Byrne, 2012) and values < 0.08 indicative of acceptable fit (Hu & Bentler, 1999).

Table 3. Cross Tabulation of TPP Rankings by Supervisors and Teachers

		Ranking by Teacher Ratings				
		1st Quartile	2nd Quartile	3rd Quartile	4th Quartile	Total
Ranking by Supervisor Ratings	1st Quartile	3	0	2	2	7
	2nd Quartile	2	1	2	2	7
	3rd Quartile	1	4	1	1	7
	4th Quartile	1	2	2	2	7
Total		7	7	7	7	28

Note. Overall agreement is 25%. Includes cases with n ≥ 2.

Table 4. Multilevel Null Model Estimates of Supervisors' Ratings of Novice Teachers by School (Supervisor)

	Coef.	Std. Err.	z	P>z	[95% Conf. Interval]	
Intercept (_con)	3.20	0.04	82.24	0.00	3.12	3.28
Random-effects Parameters						
var(Intercept_con)	0.17	0.03	—	—	0.12	0.25
var(Residual)	0.16	0.02	—	—	0.12	0.21

Note. Dependent variable Avg_S. Grouping variable = School. N = 279. Number of groups = 198 (min =1, avg = 1.4, max = 5). Log likelihood = -222.75. Chibar2(01) = 39.30, Prob > chibar2 = 0.000.

Null model results indicate:

- For teachers' ratings by institution, there is insufficient variation among institutions to model random intercepts.[1]
- For teachers' ratings by school, there is more variation within cases (0.16) than among different schools (0.02), but the grouping explains 11% of the variance (ICC = .11)
- For supervisors' ratings of novice teachers by school (i.e., supervisor), there is less variation within cases (0.16) than among the different schools/supervisors (0.17), with an ICC of .53.

[1] The final Hessian matrix is not positive definite although all convergence criteria are satisfied. Intercept covariance parameter is redundant.