

Observer-Rated Perspective-Taking in Medical Students: A Longitudinal Measurement Study of Standardized Patient and Faculty Ratings

Richard V. Thompson

`richard.thompson@osumc.edu`

The Ohio State University

Kavitha Norton

The Ohio State University

Matthew J. Flanigan

The Ohio State University

Megan Conroy

The Ohio State University

Sheryl Pfeil

The Ohio State University

Research Article

Keywords: Perspective-taking, Medical education, Standardized patients, Faculty evaluation, Simulation-based assessment

Posted Date: October 4th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10807829/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Empathy-related communication may follow dimension-specific trajectories during medical training; however, longitudinal research rarely compares external ratings across evaluator roles. We examined observer-rated perspective-taking and differences in standardized patient (SP) and faculty ratings across three preclinical years.

This longitudinal measurement study included 203 medical students at a Midwestern academic medical center. SPs rated live encounters, whereas faculty rated recorded encounters retrospectively. For each student and evaluator, we averaged ratings within instructional blocks and assigned each block-level observation to its curricular year. The primary student fixed-effects model treated curricular year as categorical and used student-clustered robust standard errors. Sensitivity analyses assessed stability across alternative specifications and curricular contexts.

SPs rated perspective-taking lower than faculty in all three years, although the difference narrowed from approximately 0.6 points in Year 1 to 0.2 points in Year 3 on the 1–9 scale. This pattern held across alternative student-level models, but its magnitude and trajectory changed once we accounted for the number and composition of assessments contributing to each year.

Observer-rated perspective-taking reflected both evaluator role and assessment context. SP and faculty ratings should be retained as separate, complementary sources of evidence, and longitudinal changes should be interpreted only after examining the scenarios and blocks contributing to each time point.

Introduction

Perspective-taking is an important component of patient-centered communication because it reflects whether learners convey an understanding of patients' concerns and viewpoints [1, 2]. Longitudinal research indicates that self-reported empathy outcomes among medical students vary across training by dimension and measurement approach, and cross-sectional comparisons across training stages have likewise produced mixed evidence of decline [3]. Carrard et al. [4] identified early increases followed by plateaus in cognitive and affective empathy, with distinct trajectories for behavioral empathy and emotion recognition. The present study extends this dimension-specific account by examining whether longitudinal patterns in observable perspective-taking also depend on external evaluator role and assessment context.

A critical gap in this literature concerns how observer role and assessment context shape longitudinal judgments of a specific empathy-related communication behavior. Prior work in this simulation program compared broader communication scores across standardized patient (SP) and faculty evaluators. It did not isolate observable perspective-taking, estimate within-student evaluator differences by curricular year, or test the stability of those differences across alternative specifications and block compositions. These distinctions matter because a total communication score can conceal dimension-specific

patterns, and longitudinal score changes can reflect both learner performance and the conditions under which ratings were generated.

Perspective-taking is commonly treated as a cognitive component of empathy and is distinguishable from affective sharing [1, 5, 6]. In clinical communication, perspective-taking becomes educationally relevant when learners convey an understanding of patients' concerns and viewpoints. The Kalamazoo framework identifies understanding the patient's perspective as one of seven essential communication tasks [7]. We therefore used an adapted Kalamazoo item to examine whether SPs and faculty rated this observable communication behavior differently.

Two evaluator groups rated students across all encounters: SPs, who experienced encounters in real time, and faculty, who reviewed video recordings retrospectively. Although these perspectives are commonly combined or compared in simulation-based assessment, limited evidence addresses whether they yield equivalent judgments of perspective-taking over time [8]. Observer-based assessment is also susceptible to rater and contextual influences that can shape scores independently of learner performance [9]. Establishing whether the evaluator difference changes across training is therefore necessary before longitudinal ratings can be interpreted as evidence of learner change.

Accordingly, this study isolated a single item, understanding the patient's perspective, from the Kalamazoo Communication Skills Assessment (KCSA) and estimated within-student differences between SP and faculty ratings across three preclinical years. We asked whether the evaluator difference changed over time and whether the result remained stable when we represented time, missing data, scenario contribution, and block composition differently. Social Cognitive Theory served as an interpretive lens for understanding rating patterns as context-dependent observations rather than direct measures of internal empathic capacity; the study did not test the theory's mechanisms.

Materials and Methods

This longitudinal measurement study followed 203 medical students from a single cohort across three academic years at a large Midwestern academic medical center. Students participated in structured, scripted simulated clinical encounters embedded within 11 instructional blocks distributed across the preclinical curriculum. Year 1 comprised Blocks 1–4, Year 2 comprised Blocks 5–8, and Year 3 comprised Blocks 9–11. An SP rated each encounter immediately after the live interaction, and a faculty evaluator later reviewed the video recording. We collected data between October 2021 and April 2024.

For each student and evaluator group, we averaged all available scenario-level perspective-taking ratings within each instructional block. This produced one block-level mean for every student–block–evaluator combination and yielded 3,857 observations, which were the units entered into the statistical models. Curricular year was assigned from each observation's instructional block; therefore, the year coefficients summarize the block-level observations contributing to each year rather than a second, separately calculated annual average. SP ratings were available for all 11 blocks, whereas faculty ratings were available for 10 blocks because faculty ratings were structurally absent in Block 3, within Year 1. Later

blocks involved more rigorous clinical interactions and higher expectations for student performance, as Thompson et al. [8] describe in detail.

Evaluators

SPs rated scripted clinical encounters immediately after each live interaction, whereas faculty independently rated video-recorded encounters retrospectively. SP onboarding followed the Association of Standardized Patient Educators Standards of Best Practice [10] and included observation of experienced peers, scenario-specific portrayal and feedback training, supervised formative assignments, readiness-based advancement, and implicit bias education. Faculty completed calibration using benchmarked video cases and participated in semiannual quality-improvement sessions based on scoring distributions. Because the two groups held different roles and rated under different conditions, we analyzed their scores as distinct observer perspectives rather than interchangeable measures.

Across the three-year study period, 173 faculty members contributed ratings. Faculty completed a median of 22 assessments each, with participation ranging from 1 to 101 assessments; 25% completed 12 or fewer assessments, and 25% completed more than 41. A separate SP interrater ICC could not be estimated because SP identifiers were not retained consistently in the analytic dataset. Room-based assignments and data-capture procedures did not permit ratings to be linked reliably to individual SPs across encounters.

Measures

We assessed observer-rated perspective-taking using Item 4 of the locally adapted KCSA, which asks whether the learner demonstrated understanding of the patient's perspective. Ratings were made on a 1–9 scale. The underlying Kalamazoo framework identifies understanding the patient's perspective as an essential communication task [7], whereas the 1–9 response scale reflects the local adaptation used in this study. Because the item captures one observable component of perspective-taking, it should not be interpreted as a full empathy scale.

In the present analytic dataset, the perspective-taking item showed corrected item-total correlations of $r = 0.607$ for faculty ratings and $r = 0.708$ for SP ratings when correlated with the remaining KCSA domains using scenario-level observations before aggregation. These associations indicate that the item was related to the broader communication construct but do not establish its reliability or validity as a stand-alone measure of empathy.

Statistical Analysis

The analysis addressed three questions: whether SP and faculty ratings differed for the same students, whether that difference changed across curricular years, and whether the findings depended on specific analytic choices or assessment contexts. Of the block-averaged perspective-taking scores, 279 (7.2%)

were missing. We generated 20 imputed datasets using two-level predictive mean matching, with repeated observations nested within students and evaluator type, curricular year, instructional block, and number of contributing scenarios included as predictors [11, 12]. Structurally absent Block 3 faculty ratings were excluded before aggregation and were not imputed.

To compare evaluator ratings within the same students, we used a student fixed-effects panel model in R. This approach controlled for student characteristics that remained stable during the study and estimated how ratings varied by evaluator role and curricular year. The primary model treated each year as a separate category, with faculty ratings in Year 1 as the reference, and directly estimated whether the SP–faculty difference changed in Years 2 and 3. A random-intercept mixed-effects model provided an alternative student-level specification [13].

Student-clustered Arellano-robust standard errors accounted for unequal error variance and correlation among repeated ratings from the same students. Coefficients and uncertainty estimates from the imputed datasets were pooled using Rubin’s Rules [14]. All analyses were conducted in R version 4.4.1. Additional model-selection and diagnostic results are reported with the primary results.

Sensitivity and Robustness Checks

Sensitivity analyses examined whether the central findings depended on how we represented time, missing data, the student-level model, scenario contribution, or block context. These analyses used quadratic and linear time specifications, a random-intercept mixed-effects model, block adjustment, complete cases, adjustment for the number of scenarios contributing to each block average, and sequential omission of individual blocks. Full coefficient sets and model-fit statistics are provided in Supplemental Table S1.

Results

Participant Demographics

The cohort comprised 203 medical students. The cohort was 55% female ($n = 111$) and 45% male ($n = 92$), and 53 students (26.1%) were classified as underrepresented in medicine (Table 1).

Table 1
Demographic summary by race/ethnicity, gender, and underrepresented-in-medicine status

Category	Female	Male	Total
Asian	18	22	40
Black	16	9	25
Latino	7	6	13
Native American	1	0	1
Other	9	5	14
White	60	50	110
Not URM	79	71	150
URM	32	21	53
Total	111	92	203
<i>Note.</i> URM = underrepresented in medicine. Race/ethnicity and URM status are separate classifications of the same 203 students; each set of rows sums to the column totals.			

Primary Analysis

The primary model estimated the SP–faculty difference within each curricular year and whether that difference changed relative to Year 1. Faculty and Year 1 served as the reference categories. Accordingly, the evaluator coefficient represents the SP–faculty difference in Year 1, and the interaction terms represent changes in that difference in Years 2 and 3 relative to Year 1. Student-level effects were jointly significant ($F = 1.542$, $df = 202/3,649$, $p < .001$), supporting within-student analysis. Diagnostic tests identified unequal error variance and correlation among repeated ratings; we therefore used student-clustered robust standard errors and pooled estimates across 20 imputed datasets using Rubin’s Rules. Table 2 presents the principal estimates.

Table 2
Student fixed-effects model with categorical curricular year

Term	β	SE	95% CI	t	p
Evaluator (SP vs. Faculty) in Year 1	-0.612	0.044	[- 0.697, - 0.526]	-14.03	< .001
Year 2 vs. Year 1	0.216	0.042	[0.134, 0.298]	5.18	< .001
Year 3 vs. Year 1	0.043	0.048	[- 0.052, 0.138]	0.90	.370
SP $\times$ Year 2	0.207	0.051	[0.108, 0.307]	4.08	< .001
SP $\times$ Year 3	0.393	0.059	[0.278, 0.508]	6.72	< .001
Note. CI = confidence interval; SE = standard error; SP = standardized patient.					

Main Effects and Interactions

SPs assigned lower observer-rated perspective-taking scores than faculty in each curricular year. The estimated SP–faculty difference was -0.612 in Year 1 (95% CI [- 0.697, - 0.526], $p < .001$), -0.405 in Year 2 (95% CI [- 0.456, - 0.354], $p < .001$), and -0.219 in Year 3 (95% CI [- 0.295, - 0.143], $p < .001$). Relative to Year 1, the evaluator difference narrowed by 0.207 points in Year 2 (95% CI [0.108, 0.307], $p < .001$) and by 0.393 points in Year 3 (95% CI [0.278, 0.508], $p < .001$). Thus, SP ratings remained lower than faculty ratings, but the difference declined from approximately 0.6 points in Year 1 to approximately 0.2 points in Year 3 on the 1–9 scale (Fig. 1).

Note

Points represent model-based SP minus faculty contrasts with 95% confidence intervals. Values below the dashed line at zero indicate that SPs rated perspective-taking lower than faculty.

Faculty ratings were 0.216 points higher in Year 2 than Year 1 (95% CI [0.134, 0.298], $p < .001$), whereas the Year 3–Year 1 difference was not significant ($\beta = 0.043$, 95% CI [- 0.052, 0.138], $p = .370$). These categorical-year estimates describe differences among curricular years without imposing a quadratic trajectory.

Sensitivity Analyses

In the quadratic-year fixed-effects model, the evaluator coefficient was -0.841 (95% CI [- 1.143, - 0.539]), with significant linear and quadratic year terms; evaluator-by-year terms were individually nonsignificant. The random-intercept mixed-effects model produced closely matching point estimates (evaluator $\beta = -0.841$, 95% CI [- 1.145, - 0.536]); marginal and conditional R^2 values were 0.153 and 0.206, respectively. After adjustment for block, the evaluator coefficient was -1.31 (95% CI [- 1.63, - 0.99]), the SP $\times$ Year term was significant ($\beta = 0.631$, $p < .001$), and the SP $\times$ Year² term was also significant ($\beta = -0.089$, $p = .019$).

The linear-time specification also supported lower SP ratings and a reduction in the evaluator difference over time; full estimates are reported in Supplemental Table S1.

In the scenario-count-adjusted model, the number of contributing scenarios was associated with ratings ($\beta = 0.250$, $p < .001$), and both evaluator-by-time terms were significant (SP $\times$ Year: $\beta = 0.514$, $p = .001$; SP $\times$ Year²: $\beta = -0.080$, $p = .037$). Across leave-one-block-out models, evaluator coefficients ranged from -1.31 to -0.18 , and evaluator-by-time estimates varied in magnitude, direction, and statistical significance.

The complete-case analysis produced the same overall conclusion as the multiply imputed analysis: SP ratings remained lower than faculty ratings, and the evaluator difference was smaller in later curricular years. Full coefficients and uncertainty estimates are provided in Supplemental Table S1.

Discussion

This study addresses a gap in how observer role and assessment context shape longitudinal judgments of empathy-related communication. It extends the prior program-level analysis by isolating perspective-taking rather than a multidomain communication total, estimating evaluator differences within students across curricular years, and testing whether the apparent longitudinal pattern persisted across assessment compositions. The central contribution is not a second demonstration that SPs score differently from faculty [8], but evidence that the size and longitudinal interpretation of that difference depend on the communication behavior selected and the curricular contexts represented. This finding extends dimension-specific accounts of empathy-related communication [1, 4, 5, 6] to differences among observers and assessment contexts and is consistent with frameworks that treat assessment judgments as context-dependent and strengthened through multiple sources of evidence [15, 16].

This context-dependent pattern parallels Carrard et al.'s longitudinal finding that medical students' self-reported cognitive, affective, and behavioral empathy followed distinct trajectories rather than a uniform decline [4]. The present findings extend that observation beyond self-report: external evaluations of an observable empathy-related behavior also varied across the curriculum, although the apparent trajectory depended partly on evaluator role and assessment composition. Together, the studies provide convergent evidence that longitudinal changes in empathy-related outcomes are dimension- and measurement-specific. However, the present findings should not be interpreted as a direct validation of students' self-reported empathy because the studies used different cohorts, measures, and assessment conditions.

The consistent SP–faculty difference directly addresses whether distinct evaluator roles yield equivalent judgments of perspective-taking over time. It should not be interpreted as evidence that either evaluator group provided more accurate ratings. SPs rated students immediately after live encounters, whereas faculty evaluated recorded encounters retrospectively; the resulting scores were therefore generated under distinct experiential and observational conditions and should not be treated as interchangeable. Prior simulation-assessment literature suggests that SP and faculty ratings may contribute

complementary information about clinical communication [8, 10, 17]. However, we did not examine evaluator cognition or the specific communication features informing each score, so the mechanisms underlying the difference remain a subject of further research.

The smaller SP–faculty difference in later years should be interpreted in relation to substantial block dependence rather than as an isolated developmental trend. Clinical interactions became more rigorous and performance expectations increased in later blocks, and prior research in the same simulation-assessment context identified variation in evaluator scoring across blocks [8]. Viewed through the Social Cognitive Theory lens introduced in the study aims, this pattern is consistent with communication ratings arising from interactions among learner behavior, repeated practice and feedback, evaluator position, and an increasingly demanding environment [18]. Because the study did not directly test these mechanisms, comparable cross-year studies are needed to distinguish learner change from curricular-context effects.

The within-student fixed-effects model, paired with student-clustered Arellano robust standard errors, supported inference about longitudinal evaluator differences by controlling for stable student characteristics and addressing detected heteroskedasticity and serial correlation [19, 20]. These choices strengthen the group-level comparison but do not remove heterogeneity among individual raters or contextual differences among assessments.

In the prior program-level analysis, agreement between SP and faculty ratings was moderate to good using a two-way mixed-effects ICC for absolute agreement ($ICC = 0.71$) [8]. That aggregate agreement does not preclude the systematic mean differences identified in the present longitudinal models. The previously reported poor-to-moderate agreement among individual faculty raters ($ICC = 0.26$) suggests that faculty estimates reflect averages across heterogeneous rating tendencies [8]. Aggregation across raters and scenarios, within-student comparisons, and clustered robust standard errors limit the influence of this heterogeneity on the group-level estimates; however, the findings do not establish the reliability of an individual faculty judgment. We could not evaluate a corresponding SP interrater ICC because room assignments and data-capture procedures did not preserve reliable individual SP identifiers across encounters, preventing direct comparison of within-group rater reliability between SPs and faculty.

The corrected item-total correlations linked the perspective-taking item to the broader KCSA communication construct but did not establish it as a stand-alone empathy measure. The findings therefore concern one observable communication behavior rather than empathy as a comprehensive construct, a distinction that should guide how the scores are labeled and communicated.

These findings suggest three practical steps for programs that use simulated encounters to assess communication. First, because the corrected item-total correlations did not establish the perspective-taking item as a stand-alone empathy measure, programs should specify whether an assessment represents overall empathy or one observable behavior [1, 5, 6]; programs should not report a single perspective-taking item to learners or committees as an empathy score. Second, because SPs assigned

lower ratings than faculty in every curricular year, programs that collect both ratings should report them as separate indicators rather than averaging them into a single composite, and should examine the direction and size of the evaluator difference before combining or trending them. Because the two groups rate under different conditions, a discrepancy may be informative rather than an error to be reconciled. Third, because scenario-count adjustment and leave-one-block-out analyses changed the magnitude and trajectory of the evaluator difference, year-over-year comparisons are most defensible when restricted to comparable scenarios; a narrowing SP–faculty gap should prompt a review of case mix and assessment composition before attributing it to learner improvement. Together, these steps treat SP and faculty ratings as complementary evidence within a programmatic approach to assessment [8, 16].

Several features of the design constrain interpretation. The single locally adapted Kalamazoo item did not provide a comprehensive measure of empathy, and the asymmetric absence of faculty ratings in Block 3 limited the Year 1 comparison. In addition, block averages represented different numbers and combinations of scenarios; the sensitivity analyses confirmed that assessment composition affected both the magnitude and trajectory of the estimates, limiting separation of curricular-time patterns from case-content effects. Although only 7.2% of block-averaged scores were imputed using a two-level model that accounted for repeated observations within students, inference still depended on a missing-at-random assumption and could not address unmeasured predictors of missingness. Finally, one cohort at one institution was observed under evaluator-specific rating conditions, precluding broad generalization and causal conclusions about the source of evaluator differences.

Conclusions

Observer-rated perspective-taking reflected both who evaluated the encounter and the assessment context in which the rating occurred. SPs consistently assigned lower scores than faculty, but the size and apparent narrowing of that difference depended partly on scenario contribution and block composition. These findings extend prior evaluator-comparison research by showing that differences between SP and faculty ratings are specific to the communication behavior and curricular contexts being assessed. Programs should therefore retain SP and faculty ratings as separate, complementary sources of evidence and examine assessment composition before interpreting longitudinal changes as learner development.

Declarations

Ethical Approval

The Ohio State University Institutional Review Board approved this study and waived the informed consent requirement because the study involved secondary analysis of educational assessment data.

Competing Interests

The authors declare no competing interests.

Generative Artificial Intelligence Disclosure

The authors used Microsoft Copilot during manuscript preparation to support journal-specific organization, language revision, concision, and clarification of author-written text. The tool did not generate or analyze study data, calculate statistical results, select analytic methods, or independently interpret the findings. The authors reviewed and revised all AI-assisted text, verified factual claims, statistical reporting, and references against original sources and analytic outputs, and take full responsibility for the manuscript's integrity and final content.

Funding

No external funding was received for this study.

Author Contribution

Richard V. Thompson: Conceptualization, Methodology, Software, Formal analysis, Data curation, Visualization, Writing – original draft, Project administration. Kavitha Norton: Conceptualization, Investigation, Writing – review and editing. Matthew J. Flanigan: Investigation, Writing – review and editing. Megan Conroy: Investigation, Writing – review and editing. Sheryl Pfeil: Conceptualization, Resources, Supervision, Writing – review and editing. All authors reviewed and approved the final manuscript and accept responsibility for the work.

Acknowledgement

The authors thank the standardized patients, faculty evaluators, curriculum staff, and medical students whose participation and assessment data made this study possible.

Data Availability

The datasets analyzed during the current study are not publicly available because they contain educational assessment data that could compromise student confidentiality. Subject to institutional approval and applicable data-use restrictions, de-identified aggregate materials and analytic code may be available from the corresponding author upon reasonable request.

References

1. Davis MH. Measuring individual differences in empathy: evidence for a multidimensional approach. *J Pers Soc Psychol.* 1983;44(1):113–26. <https://doi.org/10.1037/0022-3514.44.1.113>.
2. Hojat M, Gonnella JS. Eleven years of data on the Jefferson Scale of Empathy–Medical Student Version (JSE-S): proxy norm data and tentative cutoff scores. *Med Princ Pract.* 2015;24(4):344–50. <https://doi.org/10.1159/000381954>.
3. Quince TA, Kinnersley P, Hales J, da Silva A, Moriarty H, Thiemann P, Hyde S, Brimicombe J, Wood D, Barclay M, Benson J. Empathy among undergraduate medical students: a multi-centre cross-sectional comparison of students beginning and approaching the end of their course. *BMC Med Educ.* 2016;16:92. <https://doi.org/10.1186/s12909-016-0603-7>.
4. Carrard V, Bourquin C, Berney S, Ranjbar S, Schlegel K, Gaume J, Bart P-A, Schmid Mast M, Preisig M, Berney A. Trajectories of medical students' empathy nowadays: a longitudinal study using a comprehensive framework of empathy. *BMC Med Educ.* 2025;25:534. <https://doi.org/10.1186/s12909-025-07051-8>.
5. Reniers RLEP, Abu-Akel A, Seara-Cardoso A. Editorial: cognitive empathy and perspective taking: understanding the mechanisms of normal and abnormal experiences and abilities. *Front Psychiatry.* 2022;13:945258. <https://doi.org/10.3389/fpsyt.2022.945258>.
6. Stietz J, Jauk E, Krach S, Kanske P. Dissociating empathy from perspective-taking: evidence from intra- and inter-individual differences research. *Front Psychiatry.* 2019;10:126. <https://doi.org/10.3389/fpsyt.2019.00126>.
7. Bayer-Fetzer Conference Participants. Essential elements of communication in medical encounters: the Kalamazoo Consensus Statement. *Acad Med.* 2001;76(4):390–3. <https://doi.org/10.1097/00001888-200104000-00021>
8. Thompson RV, Conroy M, Flanigan M, Norton K, Pfeil S. A longitudinal evaluation of faculty and simulated participant communication. *Clin Simul Nurs.* 2026;113:101929. <https://doi.org/10.1016/j.ecns.2026.101929>.
9. Ismaeel R, Pusic L, Gottlieb M, Chan TM, Oyedokun TO, Thoma B. Bias in observed assessments in medical education: a scoping review. *Acad Med.* 2024;99(12):1438–50. <https://doi.org/10.1097/ACM.0000000000005794>.
10. Lewis KL, Bohnert CA, Gammon WL, Hölzer H, Lyman L, Smith C, Thompson TM, Wallace A, Gliva-McConvey G. The Association of Standardized Patient Educators (ASPE) Standards of Best Practice (SOBP). *Adv Simul.* 2017;2:10. <https://doi.org/10.1186/s41077-017-0043-4>.
11. van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. *J Stat Softw.* 2011;45(3):1–67. <https://doi.org/10.18637/jss.v045.i03>.
12. Enders CK. *Applied missing data analysis.* New York: Guilford Press; 2010.
13. Pinheiro J, Bates D. *Mixed-effects models in S and S-PLUS.* New York: Springer; 2000.
14. Rubin DB. *Multiple imputation for nonresponse in surveys.* New York: Wiley; 1987.

15. Hauer KE, Park YS, Bullock JL, Tekian A. My assessments are biased! Measurement and sociocultural approaches to achieve fairness in assessment in medical education. *Acad Med*. 2023;98(11S):S16–27. <https://doi.org/10.1097/ACM.0000000000005245>.
16. van der Vleuten CPM, Schuwirth LWT, Driessen EW, Dijkstra J, Tigelaar D, Baartman LKJ, van Tartwijk J. A model for programmatic assessment fit for purpose. *Med Teach*. 2012;34(3):205–14. <https://doi.org/10.3109/0142159X.2012.652239>.
17. Humphrey-Murto S, Smee S, Touchie C, Wood TJ, Blackmore DE. A comparison of physician examiners and trained assessors in a high-stakes OSCE setting. *Acad Med*. 2005;80(10 Suppl):S59–62. <https://doi.org/10.1097/00001888-200510001-00017>.
18. Bandura A. *Social foundations of thought and action: a social cognitive theory*. Englewood Cliffs: Prentice-Hall; 1986.
19. Arellano M. Computing robust standard errors for within-groups estimators. *Oxf Bull Econ Stat*. 1987;49(4):431–4. <https://doi.org/10.1111/j.1468-0084.1987.mp49004006.x>.
20. Croissant Y, Millo G. Panel data econometrics in R: the plm package. *J Stat Softw*. 2008;27(2):1–43. <https://doi.org/10.18637/jss.v027.i02>.

Figures

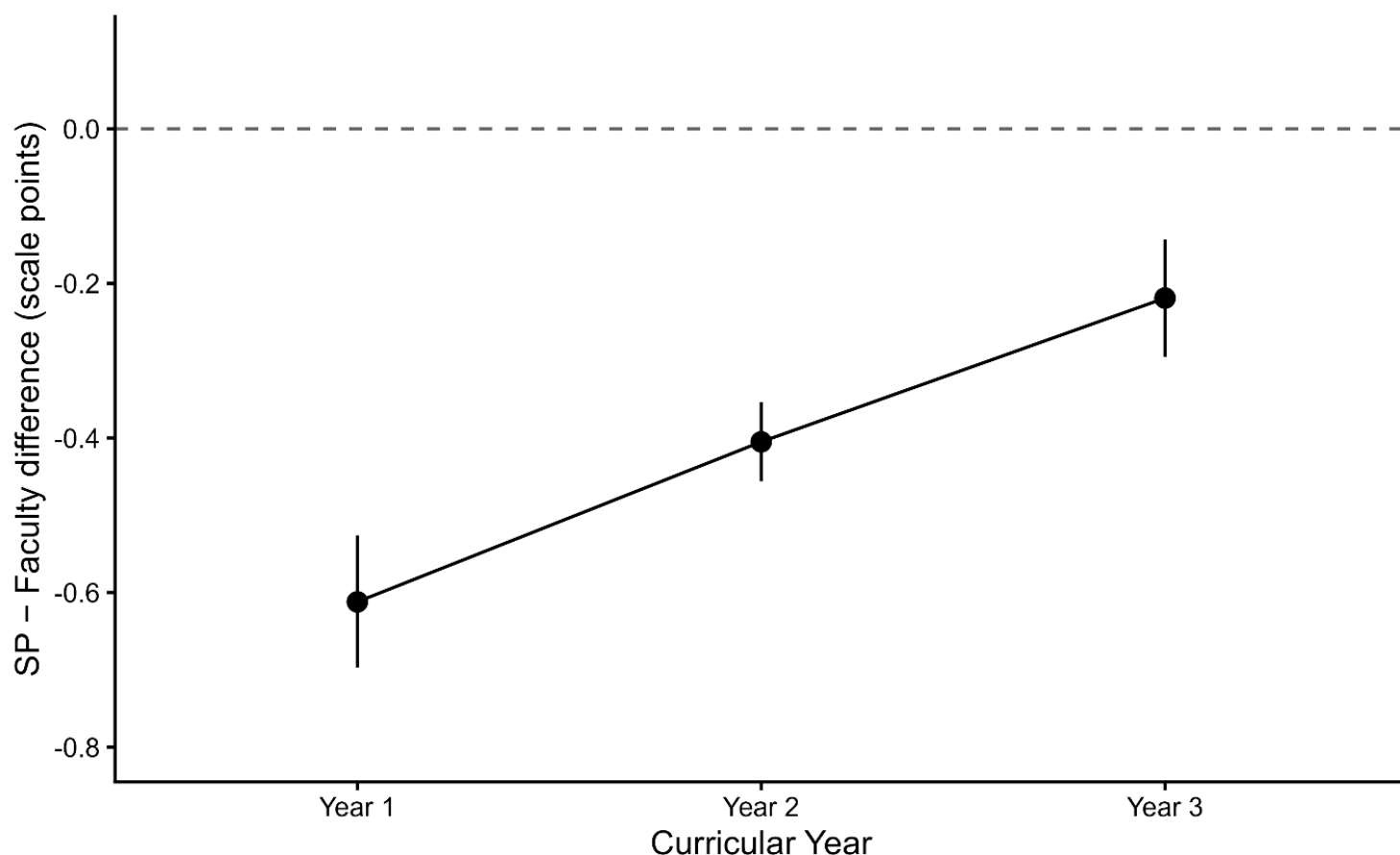


Figure 1

Estimated SP–faculty difference in observer-rated perspective-taking by curricular year.

Note. Points represent model-based SP minus faculty contrasts with 95% confidence intervals. Values below the dashed line at zero indicate that SPs rated perspective-taking lower than faculty.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementalMaterial.docx](#)