NIRN | NATIONAL IMPLEMENTATION RESEARCH NETWORK
FRANK PORTER GRAHAM CHILD DEVELOPMENT INSTITUTE

Inter-rater Reliability Report:

Rivet's Professional Learning Partner Guide (PLPG) Rubric

Executive Summary

Rivet Education developed a Professional Learning Partner Guide (PLPG), a searchable database of learning providers with expertise in the adoption and implementation of High-Quality Instructional Materials (HQIM), in response to the lack of resources and guidance available to support districts with evaluation of professional learning (PL) services. To be selected as a PL provider (PLP) in the PLPG, PL organizations have to submit an application that is reviewed and scored by Rivet Education using a PLPG Rubric (formerly known as the Scoring and Evidence Guide or SEG).

To ensure confidence in the PLPG as a robust database of high-quality PL services, the PLPG Rubric needs to be evaluated as a reliable and valid assessment instrument. The National Implementation Research Network (NIRN) is conducting psychometric testing of the PLPG Rubric. PLPG Rubric's content validity was established in Phase I (Spring 2024). The current report summarizes the research on testing inter-rater reliability or the degree to which reviewers (or raters) are consistent with each other when rating the same application. Inter-rater reliability was assessed across all 3 Gateways of the PLPG, and across subtypes of Gateways 1 and 2, for a total of 7 Gateway categories.

Overall, results for inter-rater reliability varied by the Gateway category of the Rubric. Specifically, inter-rater agreement was found to be adequate (good) for two Gateway categories: 1) Gateway 1: Applicants for Initial Implementation, Ongoing Implementation Support for Teachers, and/or Implementation Support for Leaders; and 2) Gateway 3: Using Data to Improve. Results were not adequate for the four Gateway 2 types or Gateway 1: Adoption. A comparison of qualitative and quantitative findings at the item level suggested that qualitative trends contributing to disagreement aligned with indicators that had poor rater agreement. Overall, review of trends suggested that key areas for disagreement included:

- Evidence for examining assumptions and beliefs within PL
- The use of student work as part of the PL adoption, initial implementation, and ongoing support
- Evidence for explicit connections to instructional vision
- Use of data to drive improvements

Recommendations include revising indicators and/or their scoring criteria as well as revising the training for reviewers to include time for practice. In addition, the practice of using consensus calls among reviewers during the scoring process should be retained by Rivet Education to produce Version 3.0 of the SEG.

Introduction

Professional learning (PL) is recognized widely in the field of education as a critical implementation strategy to support teacher knowledge and skill development. Quality instructional content, access to high quality instructional materials (HQIM) and support with effective instructional practices are foundational components of a coherent instructional strategy needed by teachers. PL providers typically design, deliver, and facilitate PL opportunities for educators. However, not all PL services are created equal, and districts are often ill equipped to select PL that best fits their needs.

To address this challenge, Rivet Education developed a Professional Learning Partner Guide (PLPG), a searchable database of PL providers with expertise in the adoption and implementation of High Quality Instructional Materials (HQIM). To be selected as a PL provider (PLP) in the PLPG, PL organizations and vendors have to submit an application that is reviewed and scored by Rivet Education using the PLPG Rubric (formerly known as the Scoring and Evidence Guide or SEG).

To ensure confidence in the PLPG as a robust database of high-quality PL services, the PLPG Rubric needs to be evaluated as a reliable and valid assessment instrument. The National Implementation Research Network (NIRN) is conducting psychometric testing of the Rubric, using a 3-phase methodology. The first phase, completed in the Spring of 2024, established the Rubric's content validity. Data was used to inform changes to the Rubric, with the new Rubric published in June 2024 for use by reviewers. The Phase I report and additional deliverables summarizing Phase I findings are available in the form of a brief, a webinar presentation, and a blog post. Phase II focuses on exploring the reliability and structural validity of the Rubric, based on inter-rater reliability, internal consistency, and structural validity. The third phase will involve investigation of the Rubric's convergent and predictive validity.

This report summarizes the data from the inter-rater reliability testing conducted as part of Phase II. Inter-rater reliability refers to the degree to which reviewers (or raters) are consistent with each other when rating the same application. Internal consistency and structural validity testing will be established in Fall 2025 and Spring 2026.

Phase I Deliverables

Phase I Report

Brief: The Professional Leanrng Partner Guide (PLPG) Rubric: Does it Measure curriculum-based learning?

Blog: A Conversation with Joslyn Richardson on Supporting High-Quality Instructional Materials

Presentation: Investigating the PLPG Rubric

Methods

A mixed methods approach was used to assess and interpret the inter-rater reliability of the PLPG Rubric Version 3.0. The PLPG Rubric Version 3.0 is divided into three Gateways, two of which include subtypes, for a total of seven Gateway categories that were assessed for inter-rater reliability. Data sources included (1) quantitative scores of applications submitted by trained Rivet reviewers and (2) qualitative comments explaining any indicator items on the applications that did not receive full credit when scored. The study was conducted between July 2024 and January 2025. Upon completion of the analyses and preparation of the summary documents, the findings were discussed in detail with the Rivet team during two 2-hour virtual meetings in January 2025.

Participant Sample

Participant Recruitment

Study reviewers were recruited from a pool of trained Rivet reviewers by the Rivet and NIRN team. Rivet invited their full reviewer pool to participate in the study, and asked for reviewers to reply to indicate interest. Interested reviewers were required to participate in a 1-hour training with Rivet and the NIRN team on the PLPG 3.0 rubric prior to participating in the study. The first 16 reviewers who contacted the team and subsequently attended the training were selected to participate. Three additional reviewers also attended the training to serve as "backup" reviewers if the initial reviewers could not complete their applications. During the course of the study, one initial reviewer withdrew from the study prior to completing any reviews, and one "backup" reviewer took their place. Reviewers were compensated with $75 per completed application review.

Sample Description

The 16 reviewers comprising the final pool completed a brief demographic survey. The majority of the reviewers were female (94%, N = 15). No reviewers identified as Hispanic or Latine. 81% self-identified as White, 13% as multiracial, and 6% as Black. Half of the reviewers (50%) reported having 20 or more years of experience in education, with 38% reporting 10 to 19 years of experience, 6% reporting 5 to 9 years of experience, and 6% reporting less than 5 years of experience. Most reviewers also reported having a master's degree or doctorate (56% masters; 13% doctorate; 18% other, all of whom described having at least a master's degree as well) with 13% reporting their highest level of education as a bachelor's degree. 25% of reviewers identified as professional learning providers, 25% as teachers or professors, 6% as state Department of Education staff, 31% as district administrators, and 13% as school administrators.

Power Analysis for Application Sample Size

A power analysis for Kappa was conducted in R using the kappaSize package to determine the number of applications needed for review (R Core Team, 2021; Rotondi, 2018). Rivet provided a pass/fail proportion derived from prior submission cycles to estimate proportions for the power analysis (pass rate = .61, fail rate = .39). Alpha was set to .05 and power was set to .80, with an estimated kappa of .7. The power analysis suggested a sample size of 17 applications. Based on reviewer load and application availability, 15 to 17 applications were selected per gateway category.

Data Analysis

Agreement Coefficient Analysis

For each gateway category, three statistics for agreement were calculated from pass/fail scores: Percent agreement, Fleiss's kappa (Fleiss, 1971), and Gwet's AC1 (Gwet, 2008). All were calculated using the irrCAC package in R (Gwet, 2019; R Core Team, 2021).

Notably, generally high pass rates were seen across gateway categories in the reliability study, resulting in few failed applications in several gateways. As such, for these gateways, a relatively high percentage agreement was obtained, whereas the kappa values were low. This phenomenon, known as the kappa paradox, has been documented in previous literature (e.g., Feinstein & Cicchetti, 1990) and reflects that an imbalance in prevalence across categories (i.e., few fails and many passes) impacts the expected rates for each category. This creates a large expected chance agreement that is then accounted for in the kappa calculation, rendering it low even in instances where the absolute agreement is high (Derksen et al., 2024).

To provide additional information about rater agreement, Gwet's AC1 was also calculated. In contrast to assuming that all ratings might agree simply by chance and calculating expected agreement rates (as kappa does), Gwet's AC1 uses a chance agreement probability that is adjusted to be proportional to a subset of ratings where a random rating is most likely, which is based on the expected disagreement rates (Gwet, 2008; Vach & Gerke, 2023). Gwet's AC1 is robust to the paradox issue (Gwet, 2008) and some have argued it should be used as a preferred statistic for agreement over kappa (e.g., Gwet, 2008; Silviera & Siqueira, 2023). However, others have suggested that Gwet's AC1 should not be used in place of kappa, but rather as an additional piece of information, because Gwet's AC1 is likely to produce systematically larger values than kappa and therefore should not be interpreted using the same benchmarks (Vach & Gerke, 2023).

Item Analysis

After reviewing initial agreement scores by gateway category, an individual item analysis was conducted to better understand items that might be contributing to challenges in reviewer agreement. First, for each indicator, the number of applications for which scores on that item were in disagreement was identified. Discrepancy scores (range: 0-2) for that indicator between each reviewer pair were then calculated and averaged to an overall discrepancy score for that indicator.

In addition, qualitative comments were reviewed to better understand patterns of ratings which may have contributed to discrepancies. Qualitative comments were reviewed by gateway category, with more attention paid to comments that were made when raters were in disagreement. After trends and patterns were identified by gateway, this was compared to item level agreement values.

Additional Information about Gwet's

Descriptive benchmarks for Gwet's AC1.

In lieu of applying standard interpretation benchmark scales for kappa (e.g., Altman, 1991; Fleiss et al., 2003; Landis & Koch, 1977) to AC1, Gwet (2021) proposed an alternative way of benchmarking the AC1 coefficient. In short, Gwet recommended selecting one of the standard benchmark scales, and then calculating the probability of the true coefficient falling within each interval on that scale (i.e., from poor to very good in Altman's scale) using a calculation incorporating both the obtained sample coefficient and standard error. Cumulative probabilities are then summed across intervals on the scale, using the highest agreement interval (i.e., "very good") as a starting point. Once cumulative probability exceeds a threshold of .95 (i.e., the true agreement coefficient is expected to be at or above that threshold 95% of the time), then that associated descriptive label (e.g., "good") is applied to describe the agreement.

Gwet (2021) noted this procedure can be applied with any of three commonly used interpretation thresholds for agreement: Landis & Koch, Altman, or Fleiss. However, Gwet recommended the use of Landis & Koch or Altman rather than Fleiss due to Fleiss having fewer intervals for interpretation with wider ranges. Therefore, for this study, Altman's benchmarking thresholds (1991) were used in conjunction with Gwet's recommended procedure to determine final descriptive benchmark labels for the obtained Gwet's AC1 values. See Table A1 in the Appendix for all interval probabilities and cumulative probabilities for AC1 by gateway category.

Key Insights

Overall Inter-rater Reliability by Gateway Category

Overall, findings suggested that inter-rater reliability was

- Good for:
  - Gateway 1: Applicants for Initial Implementation, Ongoing Implementation Support for Teachers, and/or Implementation Support for Leaders
  - Gateway 3
- Moderate for Gateway 2: Adoption
- Fair for Gateway 1: Adoption
- Poor for three Gateway 2 Professional Learning Types:
  - Initial Implementation
  - Ongoing Implementation Support for Teachers
  - Ongoing Implementation Support for Leaders

The number of applications reviewed per gateway, number of applications in agreement, number of agreements that were pass or fail, percent agreement, Fleiss's kappa, Gwet's AC1, and descriptive benchmark for each gateway are presented in Table 1. More on kappa and Gwet's AC1 can be found in Table A2 in the Appendix.

Table 1

Agreement Coefficients for Gateway Categories

Gateway Category	# apps reviewed	# apps in agreement (sum of next two columns)	# of pass agree [1]	# fail agree [1]	Percent agreement	Fleiss's kappa[2]	Gwet's AC_1	IRR Descriptive Benchmark [3]
Gateway 1: Content and HQIM Expertise								
Applicants for Initial Implementation, Ongoing Implementation Support for Teachers, and/or Implementation Support for Leaders	15	13	13	0	86.67%	-.07	.85	Good
Applications for Adoption	15	11	10	1	73.33%	.17	.61	Fair
Gateway 2: Quality of Professional Learning Design								
Adoption	15	12	12	0	80%	-.11	.76	Moderate
Initial Implementation	17	11	11	0	64.7%	-.21	.50	Poor
Ongoing Implementation Support for Teachers	17	10	9	1	58.8%	-.06	.33	Poor
Ongoing Implementation Support for Leaders	17	9	7	2	52.9%	-.03	.13	Poor
Gateway 3: Using Data to Plan and Improve								
Using Data to Plan and Improve	15	13	12	1	86.67%	.42	.83	Good

Note: [1]The number of pass agreements and fail agreements is provided for reference as the balance of the two affects kappa. [2]Fleiss's kappa interpretation (Fleiss et al., 2003): > 0.75 as excellent; 0.40-0.75 as fair to good; < 0.40 as poor. [3]See Appendix Table A1 for Gwet's AC_1 descriptive benchmark calculations.

Item Analysis

A comparison of qualitative and quantitative findings suggested that qualitative trends contributing to disagreement aligned with indicators with poor rater agreement.

Overall, review of trends suggested several key areas that emerged across gateways for revising indicators or training on indicators including:

- Evidence for examining assumptions and beliefs within PL
- The use of student work as part of the PL adoption, initial implementation, and ongoing support
- Evidence for explicit connections to instructional vision
- Use of data to drive improvements

Across all indicators, the number of pairs of reviewers in disagreement ranged from 0 (i.e., total agreement on the indicator) to 10 pairs. Full descriptive statistics by item, including number of pairs in disagreement and average discrepancy score between reviewer ratings, can be found in Table A3 in the Appendix.

Full descriptive statistics by item, including number of pairs in disagreement and average discrepancy score between reviewer ratings, can be found in Table A3 in the Appendix.

Additional areas emerged specific to each gateway as well, which are further summarized below.

Gateway 1

Vague descriptions and inaccuracy in knowledge of the curriculum were described in qualitative comments as reasons for lower scores.

Disagreement was higher for the Adoption indicator (1.2) than the Initial Implementation, Ongoing Implementation Support for Teachers, and Ongoing Implementation Support for Leaders indicator (1.1).

Gateway 3

Evaluating the impact of services and evaluating facilitator/coaching services showed higher disagreement than the other indicators.

Gateway 2

Disagreement emerged in areas of both the overarching indicators and the specific indicators.

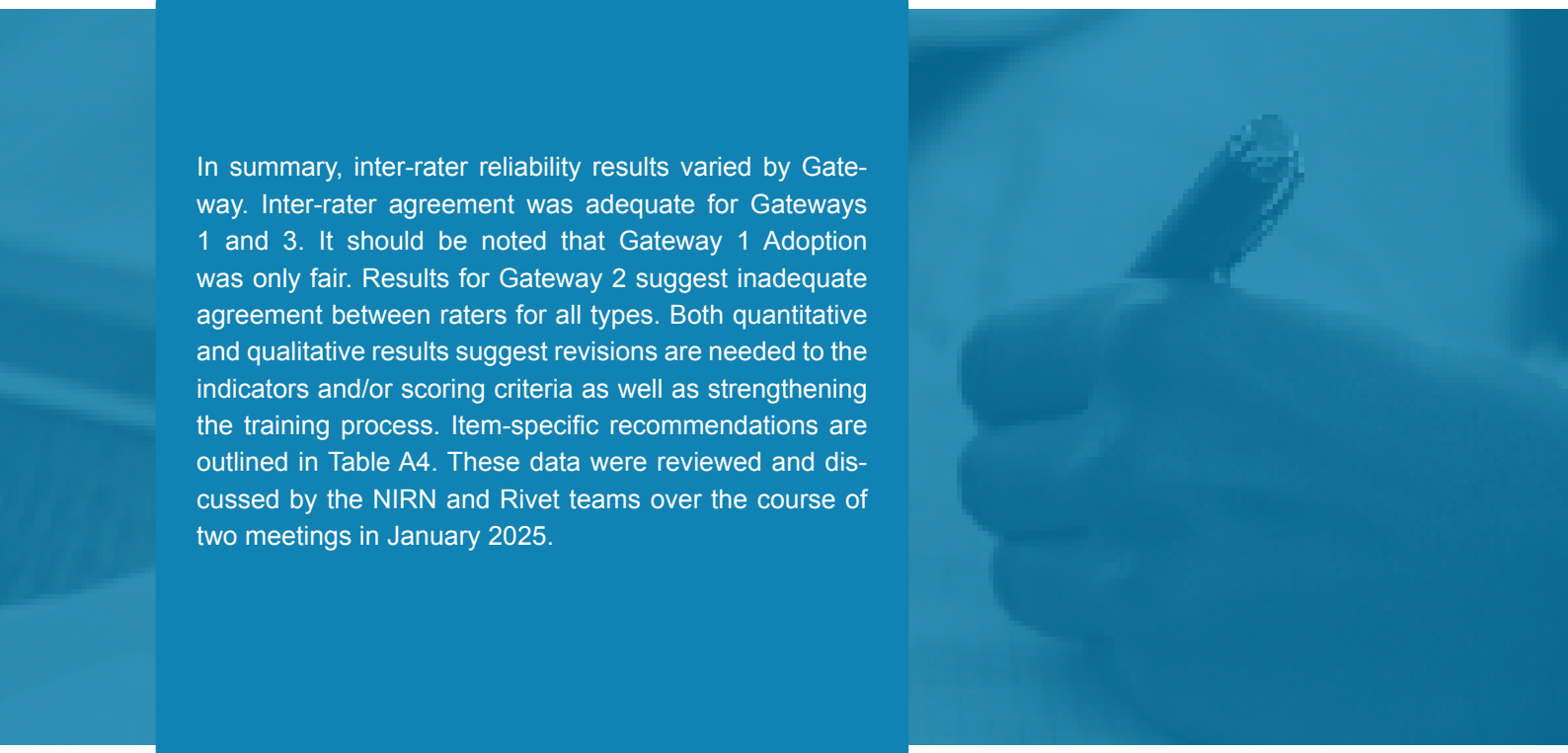
Overarching Indicators.

- Evaluating the role of using student work as part of the process (Overarching indicator 3) was a major area of rating dis agreement.
- Level of specificity of PL (overarching indicator 1) and how assumptions and beliefs are examined within PL (overarching indicator 2) also posed challenges to agreement.

Specific Indicators.

- For Adoption, explicit connection and alignment to the instructional vision (indicator 2A.6) had the highest disagreement.
- For Initial Implementation, leaders equipped to allocate time and resources (indicator 2I.9) showed the greatest disagreement.
- For Ongoing Implementation Support for Teachers, evidence of ongoing capacity building and a coaching model (indicator 2T.8) had the highest disagreement.
- For Ongoing Implementation Support for Leaders, all indicators showed high disagreement (7 pairs or more disagreeing out of 17 possible pairs).

Conclusion, Recommendations, and Next Steps



In summary, inter-rater reliability results varied by Gateway. Inter-rater agreement was adequate for Gateways 1 and 3. It should be noted that Gateway 1 Adoption was only fair. Results for Gateway 2 suggest inadequate agreement between raters for all types. Both quantitative and qualitative results suggest revisions are needed to the indicators and/or scoring criteria as well as strengthening the training process. Item-specific recommendations are outlined in Table A4. These data were reviewed and discussed by the NIRN and Rivet teams over the course of two meetings in January 2025.

Specific recommendations include:

01 Revisit the scoring evidence criteria within each indicator and revise indicators for which disagreement was high. This will be particularly important for Gateway 2 and several of the overarching indicators.

- Specific recommendations related to editing items and/or revising the scoring criteria are outlined in Table A4.
- In addition, it was recommended that additional information be collected through Rivet's reviewer survey in February on the following indicators: 1.2 (focus is on obtaining feedback from reviewers on Gateway 1 Adoption template, especially on questions 1 and 3), 2T.6 (focus is on obtaining additional information from reviewers on materials they expect to use to demonstrate "embedded supports"), and 2T.8 (focus is on obtaining additional information on challenges faced by reviewers when scoring this indicator).

02 In addition to providing an overview of changes and rationale during the training time, build in time for reviewers to practice scoring indicators from the revised rubric and for discussion of that practice, especially as it relates to:

- *new* indicators when training on the updated rubric
- *more* complex indicators, such as 2I.3.
- *Indicators with underlying assumptions and expectations*, such as indicator 3.2 (demonstrating anything less than three levels of Guskey is an automatic 0-point score) and 2L.6 (a 2-point score needs further operationalization)

03 Retain consensus calls among reviewers based on variability in reviewers' scores so that reviewers have an opportunity to discuss differences in scoring and come to an agreement.

Next steps include making identified revisions to indicators and scoring criteria and updating the training process. In addition, plans will be solidified for conducting internal consistency and structural validity using the revision of the guide.

References

Altman, D. G. (1991). Practical Statistics for Medical Research. London: Chapman and Hall.

Derksen, B. M., Bruinsma, W., Goslings, J. C., & Schep, N. W. L. (2024). The kappa paradox explained. The Journal of Hand Surgery, 49(5), 482-485.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. Journal of Clinical Epidemiology, 43(6), 543-549.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin, 75(5), 378-382.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). Statistical methods for rates and proportions (3rd ed.). John Wiley & Sons.

Gwet, K. L. (2008). Computer inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology, 61(1), 29-48.

Gwet, K. L. (2019). _irrCAC: Computing Chance-Corrected Agreement
Coefficients (CAC)_. R package version 1.0, <https://CRAN.R-project.org/package=irrCAC>

Gwet, K. L. (2021). Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters, Volume 1, Analysis of categorical ratings (5th ed.). Publisher Services.

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159-174.

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rotondi, M.A. (2018). _kappaSize: Sample Size Estimation Functions for Studies of Interobserver Agreement_. R package version 1.2, https://CRAN.R-project.org/package=kappaSize

Silveira, P. S. P., & Siqueira, J. (2022). Better to be in agreement than in bad company: A critical analysis of many kappa-like tests. Behavior Research Methods, 55, 3326-3347.

Vach, W., & Gerke, O. (2023). Gwet's AC1 is not a substitute for Cohen's kappa - A comparison of basic properties. MethodsX, 10, Article 102212.

Appendix

Table A1: *Full Calculations for Descriptive Benchmarks for Gwet's AC_1 Using Altman (1991) Benchmarks*

Gwet AC_1 value (coefficient)	SE	Low Benchmark	High Benchmark	Probability in Category	Cumulative Probability	Category Label
Gateway 1 Applicants for Initial Implementation, Ongoing Implementation Support for Teachers, and/or Implementation Support for Leaders						
.84772	.11746	0.8	1	0.620786564	0.620786564	Very Good
		0.6	0.8	0.359854172	**0.980640736**	**Good**
		0.4	0.6	0.019282803	0.999923539	Moderate
		0.2	0.4	7.64414E-05	0.999999981	Fair
		-1	0.2	1.93885E-08	1	Poor
Gateway 1 Applicants for Adoption						
Gwet AC_1 value (coefficient)	SE	Low Benchmark	High Benchmark	Probability in Category	Cumulative Probability	Category Label
.60784	.21221	0.8	1	0.15530983	0.15530983	Very Good
		0.6	0.8	0.34322701	0.49853684	Good
		0.4	0.6	0.33230966	0.8308465	Moderate
		0.2	0.4	0.14093077	**0.97177726**	**Fair**
		-1	0.2	0.02822274	1	Poor
Gateway 2 Adoption						
Gwet AC_1 value (coefficient)	SE	Low Benchmark	High Benchmark	Probability in Category	Cumulative Probability	Category Label
.7561	.15581	0.8	1	0.35093481	0.35093481	Very Good
		0.6	0.8	0.48098548	0.83192029	Good
		0.4	0.6	0.15624178	**0.98816207**	**Moderate**
		0.2	0.4	0.01164764	0.99980971	Fair

		-1	0.2	0.00019029	1	Poor

Gateway 2 Initial Implementation

Gwet AC_1 value (coefficient)	SE	Low Benchmark	High Benchmark	Probability in Category	Cumulative Probability	Category Label
.50244	.22265	0.8	1	0.07898819	0.07898819	Very Good
		0.6	0.8	0.24301831	0.32200649	Good
		0.4	0.6	0.35111199	0.67311848	Moderate
		0.2	0.4	0.23858464	0.91170312	Fair
		-1	0.2	0.08829688	**1**	**Poor**

Gateway 2 Ongoing Implementation Support for Teachers

Gwet AC_1 value (coefficient)	SE	Low Benchmark	High Benchmark	Probability in Category	Cumulative Probability	Category Label
.32578	.25957	0.8	1	0.02929568	0.02929568	Very Good
		0.6	0.8	0.11205591	0.14135159	Good
		0.4	0.6	0.24322264	0.38457423	Moderate
		0.2	0.4	0.29995397	0.6845282	Fair
		-1	0.2	0.3154718	**1**	**Poor**

Gateway 2: Ongoing Implementation Support for Leaders

Gwet AC_1 value (coefficient)	SE	Low Benchmark	High Benchmark	Probability in Category	Cumulative Probability	Category Label
.13376	.27163	0.8	1	0.0063794	0.0063794	Very Good
		0.6	0.8	0.03597736	0.04235677	Good
		0.4	0.6	0.12055258	0.16290934	Moderate
		0.2	0.4	0.24033996	0.40324931	Fair
		-1	0.2	0.59675069	**1**	**Poor**

Gateway 3

Gwet AC_1	SE	Low	High	Probability	Cumulative	Category

value (coefficient)		Benchmark	Benchmark	in Category	Probability	Label
.82659	.1327	0.8	1	0.53492579	0.53492579	Very Good
		0.6	0.8	0.41657425	**0.95150005**	**Good**
		0.4	0.6	0.04777799	0.99927803	Moderate
		0.2	0.4	0.00072068	0.99999871	Fair
		-1	0.2	1.2921E-06	1	Poor

Table A2: *Full Statistics for Kappa and Gwet's AC_1*

Gateway Category	kappa	SE	p-value	Gwet's AC_1	SE	p-value
Gateway 1 Applicants for Initial Implementation, Ongoing Implementation Support for Teachers, and/or Implementation Support for Leaders	-.07	.05	.19	.85	.12	<.001
Gateway 1 Applicants for Adoption	.17	.30	.59	.61	.21	.01
Gateway 2 Adoption	-.11	.07	.11	.76	.16	<.001
Gateway 2 Initial Implementation	-.21	.09	.03	.50	.22	.04
Gateway 2 Ongoing Implementation Support for Teachers	-.06	.24	.81	.32	.26	.23
Gateway 2 Ongoing Implementation Support for Leaders	-.03	.25	.90	.12	.27	.63
Gateway 3	.42	.35	.25	.83	.13	< .001

Table A3: *Item-level Quantitative Analysis for Indicator Agreement*

Indicator Number	Indicator Text	# pairs in disagreement	Average discrepancy
1.1	Applicants for Initial Implementation, Ongoing Implementation Support for Teachers, and/or Implementation Support for Leaders Professional learning provider demonstrates an understanding of the HQIM's approach, design principles, and structure/components.	2/15	0.13
1.2	Applicants for Adoption Professional learning provider demonstrates an understanding of the content standards and shifts, and the HQIM that align with them.	5/15	0.4
2A.1	Professional learning materials are specific to educators' roles (e.g., position, subject area, and grade level) and levels of expertise.	4/15	0.2667
2A.2	Professional learning prioritizes equity by helping educators examine how assumptions and practices can impact instruction, and professional learning builds or reinforces educators' beliefs that each and every student should have access to rigorous, grade-level instruction.	5/15	0.4
2A.3	Professional learning is grounded in examples of student work and provides opportunities for leaders to examine sample student work for alignment to a vision for excellent and equitable instruction.	5/15	0.4667
2A.4	Professional learning incorporates opportunities for active engagement and collaboration and uses appropriate adult learning strategies in a variety of formats.	2/15	0.1333
2A.5	Professional learning supports school and/or district leaders in defining or refining a shared, content-specific vision for excellent and equitable instruction, communicating that vision, and understanding the role HQIM plays in achieving that vision.	1/15	0.0667
2A.6	Professional learning supports school and/or district leaders in developing and executing an adoption plan that results in the selection and procurement of HQIM aligned to a vision for excellent, equitable instruction.	6/15	0.4667
2I.1	Professional learning materials are specific to an HQIM, educators' roles (e.g., position, subject area, and grade level), and their levels of expertise.	4/17	0.2353
2I.2	Professional learning prioritizes equity by helping educators examine how their assumptions and practices can impact instruction, and professional learning builds or reinforces educators' beliefs that each and every student should have access to rigorous, grade-level instruction.	4/17	0.2353
2I.3	Professional learning is grounded in examples of student work and provides opportunities to examine sample student work for alignment to the district vision for excellent and equitable instruction.	9/17	0.6471

2I.4	Professional learning incorporates opportunities for active engagement and collaboration and uses appropriate adult learning strategies in a variety of formats.	0/17	0
2I.5	Professional learning connects the vision for instruction to the HQIM.	4/17	0.2941
2I.6	Professional learning builds teachers' and leaders' understanding of what it means to implement their HQIM skillfully, including design principles and the arc of learning, and connects it back to a content-specific vision for excellent and equitable grade-level instruction.	3/17	0.1765
2I.7	Professional learning provides time for teachers and leaders to understand best practices for preparing to teach and internalize lessons and units in the HQIM.	4/17	0.2353
2I.8	Professional learning equips teachers and leaders to account for and navigate any publisher-specific logistical and technological considerations involved in classroom use of the HQIM, such as the components of the materials, how they are organized, and how teachers and students can access them.	2/17	0.1176
2I.9	Professional learning equips leaders to allocate essential resources and time necessary for strong HQIM implementation.	7/17	0.7059
2T.1	Professional learning materials are specific to an HQIM, educators' roles (e.g., position, subject area, and grade level), and their levels of expertise.	5/17	0.2941
2T.2	Professional learning prioritizes equity by helping educators examine assumptions and practices, focuses on specific actions that impact instruction, and reinforces that each and every student should have access to rigorous, grade-level instruction.	3/17	0.1765
2T.3	Professional learning is grounded in evidence of student learning and works to support teachers to reflect on and analyze student work from the HQIM.	6/17	0.5294
2T.4	Professional learning incorporates opportunities for active engagement and collaboration and uses appropriate adult learning strategies in a variety of formats.	2/17	0.1176
2T.5	Professional learning supports teachers with internalizing and rehearsing units and lessons with colleagues who teach the same content and HQIM, focusing on anticipating student thinking and responses and using HQIM-embedded supports to help each student access grade-level-appropriate content.	3/17	0.1765
2T.6	Professional learning equips teachers to address the needs of students with diverse and/or individualized learning needs by leveraging HQIM-embedded supports.	5/17	0.3529
2T.7	Professional learning reinforces teachers' understanding of what skillful implementation of their HQIM looks like and connects it back to a content-specific vision for excellent and equitable grade-level instruction.	3/17	0.1765

2T.8	Professional learning delivers a coaching model embedded in the larger professional learning plan that provides teachers and leaders with coaching grounded in the HQIM and builds the capacity of district and school building leaders to maintain the coaching model over time.	9/17	0.7059
2L.1	Professional learning materials are specific to an HQIM, educators' roles (e.g., position, subject area, and grade level), and their levels of expertise.	4/17	0.3529
2L.2	Professional learning prioritizes equity by helping educators examine how their assumptions and practices can impact instruction, and professional learning builds or reinforces educators' beliefs that each and every student should have access to rigorous, grade-level instruction.	4/17	0.4118
2L.3	Professional learning is grounded in evidence of student learning, especially student work, and equips leaders to monitor and identify trends within students' achievements of grade-level content and teachers' implementation of their HQIM.	6/17	0.5294
2L.4	Professional learning incorporates opportunities for active engagement and collaboration and uses appropriate adult learning strategies in a variety of formats.	2/17	0.1765
2L.5	Professional learning supports leadership to define and refine a vision for strong implementation of the HQIM that aligns with a broader vision for excellent and equitable grade-level instruction.	7/17	0.6471
2L.6	Professional learning prepares leaders to build coherence across their systems by examining and adjusting systems-level procedures, policies, and processes to monitor and support the implementation of the HQIM.	10/17	0.7647
2L.7	Professional learning develops leaders' abilities to provide and support professional learning, including coaching, observations, and feedback that are anchored in the HQIM.	7/17	0.5294
2L.8	Professional learning equips leaders to allocate essential resources and time necessary for a strong HQIM implementation.	8/17	0.6471
2L.9	Professional learning supports district and school leaders to use relevant data to monitor, support, and improve implementation.	9/17	0.7059
3.1	Professional learning provider has specific systems and processes in place to learn about clients' goals, resources, and requirements in order to tailor approaches and/or services to meet clients' needs.	4/15	0.2667
3.2	Professional learning provider evaluates the impact of its services to ensure participants' learning and to drive improvements.	5/15	0.3333
3.3	Professional learning provider evaluates facilitators for knowledge of content, content pedagogy, HQIM, and adult learning practices. Professional learning provider has systems and processes in place to provide facilitators with training as needed.	3/15	0.2667

3.4	Professional learning provider has a process to evaluate facilitator/coach effectiveness and uses that data to improve overall services and address individual facilitators' needs.	7/15	0.5333
3.5	Professional learning provider has a process in place to differentiate materials for HQIM published across multiple platforms, to stay up to date on changes to publication formats and content, and/or to stay informed on current HQIM available on the market.	3/15	0.2

Table A4: *Item-specific Recommendations for improving the PLPG Rubric and Reviewer Training*

Indicator Number	Recommendations
Gateway 1	
1.2	More information needed before making changes. Via a survey in February 2025, obtain feedback from reviewers about the Gateway 1 Adoption template that providers use to submit their application, paying particular attention to the first and third questions on the template.
Gateway 2	
2A.2	For 1-point criteria, lead with: "Provider has one of the following:". For 0-point criteria, lead with: "Provider has none of the following:". Consider defining the word "assumptions" in the glossary on the PLPG website. Consider what resources or guidance could be given to PL providers to assist in their understanding of what "assumptions" means and how evidence of this shows up in their work.
2A.3	Remove the "grounded in" piece and "student work" from the indicator. Change the indicator wording to, "Professional learning provides opportunities for leaders to examine student experiences within the curriculum for alignment to the vision for excellent and equitable instruction."
2A.6	Remove "procurement" from the indicator, criteria, and evidence. Keep "selection" in the indicator. Update wording of the 1-point criteria to be clear that it's one of the bullet points but not both.
2I.1	The second bullet point in 2A.1's 1-point response should replace the second bullet point in this indicator's 1-point response. It should read: ● Professional learning is specific to a content area but ○ Does *not* delineate for grade-level bands as called for by the standards. ○ Is *not* specific to participants' levels of expertise.
2I.3	In the indicator, remove "is grounded in examples of student work" and streamline to just "provides opportunities to examine sample student work". Remove "some" from the first bullet under the 1-point criteria, so that it exactly matches the first bullet under the 2-point criteria. Provide opportunities for reviewers to practice scoring on this indicator during training, given its complexity.
2I.9	Flip the order of the bullet points in the 2, 1, and 0-point criteria so the overview bullet is first. Remove "limited" from the overview bullet point in the 1-point criteria and remove "or" between the two bullet points. Changing the order and removing the "or" should make it clear that, for 1 point, the PL provides an overview but does not equip leaders to allocate resources and time.
2T.1	In the first bullet point in the 1-point response, remove "leaders" and change to "teachers with some experience". The second bullet point in 2A.1's 1-point response should replace the second bullet point in this indicator's 1-point response. It should read:

	● Professional learning is specific to a content area but ○ Does *not* delineate for grade-level bands as called for by the standards. ○ Is *not* specific to participants' levels of expertise.
2T.3	In the indicator, remove "is grounded in examples of student work" and streamline to just "provides opportunities to examine sample student work".
2T.6	More information needed before making changes. Via a survey in February 2025, obtain feedback from reviewers about what they expect to see for embedded supports when reviewing applications.
2T.8	Via a survey in February 2025, obtain more information from reviewers about challenges faced when scoring with this indicator. Recognizing that this was a new indicator, pull reviewer scores for this item after the current review cycle closes to see if scoring was more consistent in this round than in the reviews completed on old applications. Remove "evidence-based" from the coaching model in the indicator, so it reads "...has a coaching model…"
2L.3	Re-word the indicator to align with the focus on support for leaders, including replacing "is grounded in" with "provides opportunities to examine". Include "monitor and identify trends" in indicator 2L9 instead.
2L.5	Add examples of what "instances" might look like to the third evidence bullet.
2L.6	In the 1-point criteria, remove "a few of" and "adjust". Change the wording to, "...prepares leaders to examine but not adjust the systems-level structures…" Provide more training for reviewers on what adjustment looks like in order for an application to receive a 2-point score on this indicator.
2L.7	Reword the indicator to say, "Professional learning develops leaders' abilities to provide and support a comprehensive professional learning plan anchored in the HQIM." The criteria should be where the structures are identified. Adjust the criteria to reflect how the leaders are equipped to support teachers through coaching and feedback. The 2-point criteria could be inclusive of workshops, collaborative planning, and coaching. 1-point criteria could say that "PL equips leaders to provide and support PL inclusive of some, but not all, CBPL structures" for the first bullet point; remove "is not anchored in the HQIM" from the wording of the first bullet point. 0-point criteria should say that PL does not equip them at all; it could have an "or" for to say that that it is equipping them to do this *but not anchored in an HQIM*.
2L.8	For the 2-point, 1-point, and 0-point criteria, switch the order of the bullets so the overview bullet comes first. The 2-point criteria should say that the PL provides an over and does equip leaders, the 1-point criteria should say that the PL provides an overview but does *not* equip leaders, and the 0-point response should say that the PL neither provides an overview nor does it equip leaders.

	Edit to add “monitor and identify trends” which was originally part of 2L3 In the 1-point criteria, use positive language in both bullet points (removing “not”) and add an “or” between the two bullets. For 1-point, the provider demonstrates doing one but not both of these. In the third bullet point in the evidence, say “this can include” or “some examples include” to imply it’s some but not all of these data.
2L.9	This was a new indicator, so reviewers should practice scoring it during training.

Gateway 3

3.2	The 1-point criteria should say that the provider evaluates on three or more Guskey levels *but* does not have a process for sharing and debriefing data with clients. Remove the second bullet point for the 0-point criteria. Emphasize in reviewer training that having anything less than three levels of Guskey is an automatic 0-point rating.
3.4	For the 2-point criteria, change to a single bullet saying that the provider has “a defined process to evaluate…” *and* they use that data to improve overall services. For the 1-point criteria, change to a single bullet saying that the provider has a process for evaluation, but *does not* use it to improve. Remove any wording about concrete details.
Overall	Go through the criteria for all indicators and ensure that the logic is in place for the 0-point, 1-point, and 2-point criteria on all indicators.