Intern Evaluation – Mid-Experience: Pilot Data: Brief Report (Draft)

The following report briefly summarizes basic item-level, reliability, and descriptive information for the revised intern evaluation. The obtained data are based on the second experience, mid-experience evaluation of 33 candidates, rated by mentors and completed during the fall, 2017 semester.

Item-Level Information

Item-level information is summarized in the following table. Across items on the intern evaluation, the response of 'Proficient' was most often obtained. Interns demonstrated strong performance with respect to learning environments, flexibility and responsiveness, and ethical practice based on the number of 'Exemplary' responses obtained on the respective items. A response of 'Unsatisfactory' (*n*=1) was obtained for the item evaluating performance with respect to learner assessment. The response 'No opportunity to observe' was obtained more frequently on the following items: impact on learning, communication with families, and leadership and collaboration. Across broad categories, this indicates that the response 'No opportunity to observe' was obtained more often in the area of professional responsibility. Other trends can be viewed in the table.

Item-Level Information

Item	How Often Each Response Was Selected (Frequencies)				
	No opportunity	Unsatisfactory	Developing	Proficient	Exemplary
Learner Development[a]	0	0	5	18	10
Learner Differences[a]	1	0	6	16	10
Learning Environments[a]	0	0	1	13	19
Managing Classroom Procedures[a]	1	0	7	15	10
Content Knowledge[b]	2	0	5	18	8
Content Application[b]	1	0	4	20	8
Pedagogical Procedures[b]	1	0	4	19	9
Flexibility and Responsiveness[b]	1	0	2	14	16
Learner Assessment[c]	4	1	6	16	6
Learner Feedback[c]	3	0	2	16	12
Impact on Learning[c]	5	0	6	16	6
Reflection on Teaching[c]	0	0	3	18	12
Instructional Resources[c]	4	0	4	20	5
Planning for Instruction[c]	2	0	3	19	9
Instructional Strategies[c]	1	0	5	18	9
Instructional Technology[c]	4	0	2	15	12
Communication with Families[d]	15	0	4	9	5
Ethical Practice[d]	0	0	1	9	23
Professional Development[d]	3	0	4	13	13
Leadership and Collaboration[d]	8	0	1	13	11

Note. [a]Items contribute to The Learner and Learning category; [b]Items contribute to the Content Knowledge category; [c]Items contribute to the Instructional Practice category; [d]Items contribute to the Professional Responsibility category.

Reliability Evidence (Internal Consistency)

To provide a preliminary evaluation of reliability evidence for the intern evaluation, internal consistency reliability was examined. Internal consistency reliability is commonly used to evaluate the reliability of a set of test or questionnaire items. It provides an indication of an instrument's reliability by estimating the extent to which items on an instrument consistently measure the same construct (e.g., intern performance).[1] Values exceeding 0.70 are considered 'adequate', while values exceeding 0.80 are preferred for pilot work. For comparative purposes and because the response option 'No opportunity to observe' can feasibly be regarded as missing data, internal consistency reliability was examined both with the 'No opportunity to observe' response option included as well as with the 'No opportunity to observe' response option recoded as missing data.

Reliability With 'No opportunity to observe' Response Option Included

With data based on the 'No opportunity to observe' response option included in the analysis, internal consistency reliability of the intern evaluation was strong (α=0.93). Reliability evidence was then examined for items measuring each intern evaluation/InTASC category: the learner and learning, content knowledge, instructional practice, and professional responsibility. Evidence of internal consistency reliability was adequate for the learner and learning (α=0.78), content knowledge (α=0.77), and instructional practice (α=0.88) categories, and was low for the professional responsibility (α=0.52) category.

Reliability With 'No opportunity to observe' Response Option Excluded

With data based on the 'No opportunity to observe' response option excluded (i.e., recoded) from the analysis, internal consistency reliability of the intern evaluation was strong (α=0.97). Likewise, evidence of internal consistency reliability was adequate for the learner and learning (α=0.84), content knowledge (α=0.86), instructional practice (α=0.91), and professional responsibility (α=0.87) categories.

Intern Evaluation/InTASC Categories – Composite Scores

Descriptive Information About The Four Intern Evaluation/InTASC Categories

To examine performance across the four intern evaluation/InTASC categories, average scores were created for each candidate and compared across each category. The response option 'No opportunity to observe' was excluded, and items were recoded such that 1 corresponded to 'Unsatisfactory' and 4 corresponded to 'Exemplary'. As a result, for the purposes of this section, higher scores are indicative of better performance on each category of the evaluation. The following table presents relevant descriptive statistics for these average scores for each of the four categories.

Category Descriptive Information

Area	Mean	Median	*SD*	Min-Max
The Learner and Learning	3.24	3.25	0.55	2.00-4.00
Content Knowledge	3.24	3.25	0.51	2.00-4.00
Instructional Practice	3.25	3.13	0.48	2.00-4.00
Professional Responsibility	3.48	3.75	0.52	2.00-4.00

Note. Mean reflects the mean of the average scores for each area; Median reflects the median of the average scores for each area; *SD*=standard deviation; Min-Max=range of average scores from minimum to maximum score.

Overall, candidates demonstrated similar performances on the four intern evaluation/InTASC categories. Median values for each category largely corresponded to proficient performance. Descriptively based on analyses conducted but not displayed in the table, these performances were also similar regardless of candidate major (early childhood compared with elementary education).

Correlations Among The Four Intern Evaluation/InTASC Categories

To examine the extent to which scores on each of the intern evaluation/InTASC categories were related to one another, correlations were computed among the four categories.[2] Correlation coefficient values ranged from 0.78 to 0.89[3], indicating that scores on the categories of the learner and learning, content knowledge, instructional practice, and professional responsibility were moderately to strongly correlated with one another.[4] These correlations could again be used to support validity evidence based on relations to other variables; such findings would support convergent evidence that these four categories are and should be related. In other words, the findings suggest that, as expected, these four categories reflect related dimensions of teaching performance.

Intern Evaluation – Final Experience: Pilot Data: Brief Report (Draft)

The following report briefly summarizes basic item-level, reliability, and descriptive information for the revised intern evaluation. The obtained data summarized in this section are based on the second final-experience evaluation of 33 candidates, rated by both mentors and supervisors and completed during the fall, 2017 semester.

Item-Level Information

Item-level information is summarized in the following tables, first for mentors and then for supervisors.

Item-Level Information For Mentors

Item	How Often Each Response Was Selected (Frequencies)				
	No opportunity	Unsatisfactory	Developing	Proficient	Exemplary
Learner Development[a]	0	0	3	9	21
Learner Differences[a]	0	0	3	11	19
Learning Environments[a]	0	0	1	7	25
Managing Classroom Procedures[a]	0	0	2	9	22
Content Knowledge[b]	0	0	3	13	17
Content Application[b]	0	0	3	12	18
Pedagogical Procedures[b]	1	0	2	15	15
Flexibility and Responsiveness[b]	0	0	1	7	25
Learner Assessment[c]	1	0	2	15	15
Learner Feedback[c]	1	0	1	10	21
Impact on Learning[c]	1	0	1	15	16
Reflection on Teaching[c]	0	0	0	11	22
Instructional Resources[c]	2	0	1	19	11
Planning for Instruction[c]	0	0	3	14	16
Instructional Strategies[c]	1	0	2	10	20
Instructional Technology[c]	2	0	0	12	19
Communication with Families[d]	5	0	0	17	11
Ethical Practice[d]	0	0	0	8	25
Professional Development[d]	0	0	0	12	21
Leadership and Collaboration[d]	4	0	0	9	20

Note. [a]Items contribute to The Learner and Learning category; [b]Items contribute to the Content Knowledge category; [c]Items contribute to the Instructional Practice category; [d]Items contribute to the Professional Responsibility category.

Mentors' ratings largely fell within the 'Proficient' and 'Exemplary' categories. The rating of 'No opportunity to observe' was provided primarily for items assessing the areas of Communication with Families and Leadership and Collaboration. No ratings of 'Unsatisfactory' were provided. In general, compared with the mid-experience evaluation, mentors provided fewer ratings of 'Developing'.

Supervisors' ratings were slightly more varied but also largely fell within the 'Proficient' and 'Exemplary' categories. No responses of 'Unsatisfactory' were provided. A large number of 'No opportunity to observe' responses were obtained for the items assessing Communication with Families and Ethical Practice in the category of Professional Responsibility, again suggesting the need to review the applicability of these items to the supervisor version of the intern evaluation.

Item-Level Information For Supervisors

Item	How Often Each Response Was Selected (Frequencies)				
	No opportunity	Unsatisfactory	Developing	Proficient	Exemplary
Learner Development[a]	0	0	1	7	25
Learner Differences[a]	0	0	2	15	16
Learning Environments[a]	0	0	0	4	29
Managing Classroom Procedures[a]	0	0	0	11	22
Content Knowledge[b]	0	0	2	10	21
Content Application[b]	0	0	2	15	16
Pedagogical Procedures[b]	0	0	2	19	12
Flexibility and Responsiveness[b]	0	0	1	6	26
Learner Assessment[c]	1	0	2	20	10
Learner Feedback[c]	0	0	1	8	24
Impact on Learning[c]	1	0	1	16	15
Reflection on Teaching[c]	0	0	0	18	15
Instructional Resources[c]	3	0	2	16	12
Planning for Instruction[c]	0	0	1	14	18
Instructional Strategies[c]	0	0	2	13	18
Instructional Technology[c]	1	0	0	16	16
Communication with Families[d]	20	0	0	4	9
Ethical Practice[d]	6	0	0	5	22
Professional Development[d]	3	0	0	15	15
Leadership and Collaboration[d]	3	0	1	15	14

Note. [a]Items contribute to The Learner and Learning category; [b]Items contribute to the Content Knowledge category; [c]Items contribute to the Instructional Practice category; [d]Items contribute to the Professional Responsibility category.

Reliability Evidence (Internal Consistency and Inter-Rater)

To examine evidence for reliability of scores on the intern evaluation, both internal consistency and inter-rater reliability methods were used. Internal consistency evidence was first examined for both mentors and supervisors[1]. In the following table, estimates of internal consistency reliability are provided by assessor for the intern evaluation overall as well as for items measuring each intern evaluation/InTASC category: The Learner and Learning, Content Knowledge, Instructional Practice, and Professional Responsibility. Evidence of internal consistency reliability was adequate for scores on the intern evaluation overall as well as for each category.

Internal Consistency Reliability By Rater

Scale/InTASC Category	Mentors	Supervisors
Intern Evaluation Overall	0.96	0.96
The Learner and Learning	0.95	0.80
Content Knowledge	0.91	0.85
Instructional Practice	0.93	0.91
Professional Responsibility	0.78	0.82[a]

Note. [a]Internal consistency reliability for this scale is based only on items 19 and 20.

Inter-rater reliability was next examined through two approaches. In the first approach, based on the nature of the data obtained, item-level correlations were examined[5]. This approach provided a measure of how strongly mentors' and supervisors' responses to each evaluation item were related. Higher values indicate a stronger relationship between scores, and are indicative of stronger inter-rater reliability. Correlation values ranging from 0.35 to 0.59 are generally described as being moderate in strength, while values ranging from 0.60 to 0.79 are considered strong; values exceeding 0.80 would be described as very strong. In the following table, item-level correlation values delineating the relationships between mentors' and supervisors' ratings are presented.

Inter-Rater Reliability, Approach 1 (Item-Level Correlations)

Item	Correlation value	Interpretation
1. Learner Development[a]	0.49	Moderate
2. Learner Differences[a]	0.41	Moderate
3. Learning Environments[a]	0.42	Moderate
4. Managing Classroom Procedures[a]	0.31	Weak
5. Content Knowledge[b]	0.56	Moderate
6. Content Application[b]	0.55	Moderate
7. Pedagogical Procedures[b]	0.60	Moderate
8. Flexibility and Responsiveness[b]	0.44	Moderate
9. Learner Assessment[c]	0.45	Moderate
10. Learner Feedback[c]	0.51	Moderate
11. Impact on Learning[c]	0.38	Moderate
12. Reflection on Teaching[c]	0.39	Moderate
13. Instructional Resources[c]	0.80	Moderate
14. Planning for Instruction[c]	0.32	Weak
15. Instructional Strategies[c]	0.63	Moderate
16. Instructional Technology[c]	0.68	Moderate
17. Communication with Families[d]	NC	-
18. Ethical Practice[d]	NC	-
19. Professional Development[d]	0.48	Moderate
20. Leadership and Collaboration[d]	0.68	Moderate
Average correlation value across items	0.52[6]	Moderate

Note. NC=Not calculable based on number of 'No opportunity to observe' responses.
[a]Items contribute to The Learner and Learning category; [b]Items contribute to the Content Knowledge category; [c]Items contribute to the Instructional Practice category; [d]Items contribute to the Professional Responsibility category.

Overall, moderate evidence of inter-rater reliability was obtained based on the item-level correlation values. Obtained correlation values suggest that there may be a need for further review of or calibration on items 4 (Managing Classroom Procedures) and 14 (Planning for Instruction). Inter-rater reliability evidence for items 17 (Communication with Families) and 18 (Ethical Practice) was not examined based on the number of supervisors who did not feel there was adequate opportunity to rate the candidate(s) on the two items.

In the second approach, consistency between mentors' and supervisors' item-level ratings was examined using the intraclass correlation coefficient[7]. This approach works well with ordered score categories and accounts for systematic differences in scores that may be based on raters. Values ranging from 0.50 to 0.74 are generally described as supporting moderate reliability, while values ranging from 0.75 to 0.90 support strong reliability; values exceeding 0.90 indicate excellent reliability.

Inter-Rater Reliability, Approach 2 (Intraclass Correlations)

Item	Intraclass correlation value	Interpretation
1. Learner Development[a]	0.68	Moderate
2. Learner Differences[a]	0.64	Moderate
3. Learning Environments[a]	0.48	Weak/Moderate
4. Managing Classroom Procedures[a]	0.44	Weak
5. Content Knowledge[b]	0.76	Strong
6. Content Application[b]	0.74	Moderate
7. Pedagogical Procedures[b]	0.80	Strong
8. Flexibility and Responsiveness[b]	0.73	Moderate
9. Learner Assessment[c]	0.69	Moderate
10. Learner Feedback[c]	0.68	Moderate
11. Impact on Learning[c]	0.62	Moderate
12. Reflection on Teaching[c]	0.56	Moderate
13. Instructional Resources[c]	0.89	Strong
14. Planning for Instruction[c]	0.53	Moderate
15. Instructional Strategies[c]	0.84	Strong
16. Instructional Technology[c]	0.81	Strong
17. Communication with Families[d]	NC	-
18. Ethical Practice[d]	NC	-
19. Professional Development[d]	0.65	Moderate
20. Leadership and Collaboration[d]	0.81	Strong

Note. NC=Not calculable based on number of 'No opportunity to observe' responses.
[a]Items contribute to The Learner and Learning category; [b]Items contribute to the Content Knowledge category; [c]Items contribute to the Instructional Practice category; [d]Items contribute to the Professional Responsibility category.

Overall, moderate evidence of inter-rater reliability was again obtained based on the intraclass correlation values. Obtained values suggest that there may be a need for further review of or calibration on items 3 (Learning Environments), 4 (Managing Classroom Procedures) and 14 (Planning for Instruction). Inter-rater reliability evidence for items 17 (Communication with Families) and 18 (Ethical Practice) was again not examined based on the number of supervisors who did not feel there was adequate opportunity to rate the candidate(s) on the two items.

Across analyses, the findings suggest adequate internal consistency and inter-rater reliability of scores on the intern evaluation based on the pilot data obtained. The findings suggest the need to explore revision of and calibration and training efforts on specific items (3, 4, and 14), as well as review of the applicability of specific items (17 and 18) for supervisors.

Intern Evaluation/InTASC Categories – Composite Scores

Descriptive Information About The Four Intern Evaluation/InTASC Categories

To examine performance across the four intern evaluation/InTASC categories, average scores were created for each candidate and compared across each category. The response option 'No opportunity to observe' was excluded, and items were recoded such that 1 corresponded to 'Unsatisfactory' and 4 corresponded to 'Exemplary'. As a result, higher scores are indicative of better performance on each category of the evaluation. The following table presents relevant descriptive statistics for these average scores for each of the four categories by assessor.

Category Descriptive Information by Assessor

Category	Mean	Median	*SD*	Min-Max
Mentors				
The Learner and Learning	3.59	4.00	0.58	2.00-4.00
Content Knowledge	3.52	3.50	0.53	2.00-4.00
Instructional Practice	3.53	3.63	0.45	2.25-4.00
Professional Responsibility	3.66	3.75	0.35	3.00-4.00
Supervisors				
The Learner and Learning	3.67	3.75	0.39	2.50-4.00
Content Knowledge	3.51	3.50	0.48	2.00-4.00
Instructional Practice	3.50	3.56	0.42	2.50-4.00
Professional Responsibility	3.47	3.50	0.45	2.50-4.00

Note. Mean reflects the mean of the average scores for each area; Median reflects the median of the average scores for each area; *SD*=standard deviation; Min-Max=range of average scores from minimum to maximum score.

Descriptively, mentors and supervisors provided similar ratings on the four intern evaluation/InTASC categories. Similar ratings were also obtained for each assessor across categories.

Correlations Among The Four Intern Evaluation/InTASC Categories

To examine the extent to which scores on each of the intern evaluation/InTASC categories were related to one another, correlations were computed among the four categories for each assessor role.[2] Based on mentors' ratings, correlations among the four intern evaluation/InTASC categories were significant[4]. Correlations among the scores ranged from *r*=0.83 to 0.92, indicating moderate to strong relationships among the intern evaluation categories. Based on supervisors' ratings, correlations among the four intern evaluation/InTASC categories were also significant. Correlations among the scores ranged from *r*=0.66 to 0.86, again indicating strong relationships among the intern evaluation categories. Similar to the results of the mid-experience evaluation analyses, these findings suggest that the four intern evaluation categories capture related dimensions of teaching performance.

Correlations between mentors' and supervisors' ratings are presented in the following table. These correlations provide information about how strongly scores on each category were related across assessors.

Correlations Among Ratings By Assessor

Category	Mentors' ratings:			
	TLL	CK	IP	PR
Supervisors' ratings:	-	-	-	-
The Learner and Learning	0.55	-	-	-
Content Knowledge	0.54	0.88	-	-
Instructional Practice	0.35	0.73	0.81	-
Professional Responsibility	0.42	0.71	0.72	0.69

Note. TLL=The Learner and Learning; CK=Content Knowledge; IP=Instructional Practice; PR=Professional Responsibility.

The findings indicate that scores based on mentors' and supervisors' ratings were significantly and moderately to strongly related. In particular, these correlations provide some degree of convergent evidence for the intern evaluation/InTASC categories.

Examining Differences In Mentor Ratings Over Time

Because mentors' ratings of the candidates were available from both mid- and final-experience evaluations, analyses were also conducted to examine gains in candidates' performance over time. Specifically, analyses were conducted to evaluate whether scores on the categories of The Learner and Learning, Content Knowledge, Instructional Practice, and Professional Responsibility increased from mid- to final-experience evaluations[8]. The results of the analyses indicated that scores increased significantly on each intern evaluation/InTASC category, suggesting development in performance as measured by the intern evaluation over time (based on mentors' ratings).

Conclusions

The results of the pilot analyses suggest, overall, that the intern evaluation holds continued promise as an instrument measuring intern performance. Many of the findings support basic evidence of both reliability and validity of scores. Based on the findings, however, additional revision of and work on specific items is warranted. Next steps include: evaluating and revising the language of items 3 (Learning Environments), 4 (Managing Classroom Procedures) and 14 (Planning for Instruction); again critically evaluating the applicability of items 17 (Communication with Families) and 18 (Ethical Practice) to the supervisor version of the intern evaluation; conducting training and calibration on specific items or the instrument as a whole to ensure adequate understanding and use of rubric items; and ensuring general applicability of the instrument to the entire unit.

Notes

[1] Internal consistency reliability was evaluated through calculation of Cronbach's alpha as a lower-bound estimate of reliability. Cronbach's alpha effectively evaluates the mean of all possible split-half correlations among items in an instrument. Standardized item alpha values were also computed and compared with Cronbach's alpha values.

[2] Given the nature of the data analyzed, both Pearson correlation coefficients and Spearman rank-order correlation coefficients were computed and compared. Findings with respect to the significance and magnitude of obtained correlations were similar across correlation type.

[3] Given the recoding procedure employed with the 'No opportunity to observe' response option, missing data were present, particularly on the instructional practice and professional responsibility categories. To account for this, correlations were computed and compared based on both pairwise and listwise missing data procedures. In no instance did significance of correlations change based on the missing procedure conducted, nor did interpretation of the magnitude of the obtained correlations change substantively.

[4] One-tailed tests were conducted.

[5] Given the nature of the item-level data analyzed, Spearman rank-order correlation coefficients were computed and are reported. This coefficient accounted for the ordinal nature of the ratings.

[6] Fisher's *z* transformations were conducted prior to averaging and reporting of the mean correlation value.

[7] A two-way mixed model based on the consistency type was estimated for all item-level intraclass correlation analyses. Across analyses, the average measures intraclass correlation is reported.

[8] Based on the nature and distributions of the data obtained, Wilcoxon signed rank tests were conducted as opposed to paired-samples/dependent samples *t*-tests.