Journal of Educational Sciences
Journal homepage: https://jes.ejournal.unri.ac.id/index.php/JES

JES
P-ISSN 2581-1657
E-ISSN 2581-2203

# Establishing Reliability and Validity of a High Leverage Practice Rubric for Assessment (HLPR-A)

**Deana J. Ford*, Sara E. Luke[2], S. Michelle Vaughn[3]**
*1Mercer University, Atlanta, 30341, USA*
*2Mercer University, Lithia Springs, 30122, USA*
*3Mercer University, McDonough, 30253, USA*

**ARTICLE INFO**

*Article history*:

Received: 30 Aug 2022
Revised: 28 Sept 2022
Accepted: 04 Oct 2022
Published online: 24 Oct 2022

*Keywords:*

High Leverage Practices;
Reliability;
Rubric;
Validity

**ABSTRACT**

The purpose of the current research study was to establish reliability and validity of an assessment high-leverage practice rubric (HLPR-A). Five raters scored 33 participants' instructional videos two times using the HLPR-A. Interrater and intrarater reliability was established using an intraclass correlation coefficient. Content validity was established through expert review. Construct validity was presented using a nomological network. Internal validity was confirmed using principal axis factor analysis. The results revealed the HLPR-A to be a reliable and valid tool for assessing preservice teachers' performance when interpreting and communicating assessment information to a parent during a parent-teacher conference using mixed reality virtual simulations. Future research could explore using the HLPR-A on other populations, with a variety of raters, in other environments, and programs.

## 1. Introduction

Teacher education special education is an essential field where preservice teachers must develop basic proficiency in using effective teaching practices to support improved outcomes for students with disabilities. High leverage practices (HLPs) are foundational teaching practices that can be used to effectively instruct students with and without disabilities across grade levels, schools, and content areas (McLeskey et al., 2017). High leverage practices are divided into instructional, assessment, collaboration, and social-emotional domains and represent the "essence of effective practice in special education" (McLeskey et al., 2017, p. 9). Providing preservice teachers with repeated opportunities to engage in rehearsals of HLPs is imperative for developing their fluency in using these practices once in the K-12 environment. While the importance of teaching preservice teachers HLPs is noted in the literature (McLeskey et al., 2017), formal assessments of preservice teachers' use of HLPs is absent.

* Corresponding author.
E-mail: ford_dj@mercer.edu

Doi: https://doi.org/10.31258/jes.6.4.p.707-722

Rubrics are a common form of assessment in higher education that make learning expectations clear, facilitate learners' self-assessment, and provide feedback to learners (Firmansyah et al., 2020; Jonsson, 2014). A rubric is an assessment tool used for the rating, or scoring, of authentic, or complex, student work (Firmansyah et al., 2020; Van Helvoort et al., 2017). Rubrics include specific criteria for rating dimensions of performance and typically describe levels of the quality of performance (Firmansyah et al., 2020; Jonsson, 2014). In higher education, performance assessments, authentic assessments, or other assessments that model real-world activities, can be evaluated using rubrics (Gallardo, 2020; Jonsson & Svingby, 2007). Educational researchers proposed that using rubrics reduces subjective and unfair grading processes and quantifies aspects of students' behaviors (Gallardo, 2020; Jonsson & Svingby, 2007; Van Helvoort et al., 2017). Rubrics are also beneficial for providing students with detailed feedback, explicit expectations of their instructors, and peer and self-assessment (Johnson et al., 2019; Gallardo, 2020; Jonsson & Svingby, 2007; Reddy & Andrade, 2010).

The continued use of rubrics in higher education makes establishing the validity and reliability of rubrics important because it answers one of the most difficult aspects of learning: making judgments about performance (Firmansyah et al., 2020; Gallardo, 2020). The validity of rubrics is important because establishing the appropriateness of the rubric is critical to trusting its use as a credible evaluation tool (Firmansyah et al., 2020; Johnson et al., 2019). Likewise, reliability is important because consistent scoring across raters, time, and students is essential in quality assessment. However, despite the noted importance of HLPs in teacher education as well as the frequent use of rubrics as an effective assessment tool, only one reliable rubric evaluating the HLP explicit instruction has been identified (Johnson et al. 2019). Johnson and colleagues (2019) developed a rubric for assessing the high leverage practice explicit instruction (HLP16) among special education inservice teachers. Reliability and validity estimates were determined using percentage agreements (total agreement, adjacent agreement, and Cohen's kappa), consistency estimates (Pearson's correlations, Cronbach's alpha, and Spearman's rho), and measurement estimates (generalizability theory and many-facets Rasch model). No reliable or valid rubric could be found in the literature that evaluated the assessment HLP5 and could be used with preservice teachers. More reliable and valid rubrics used for teaching and assessing HLPs should be developed.

Many researchers have established the credibility of rubrics through reliability and validity studies. In a systematic review of the literature, Jonsson and Svingby (2007) explored 75 research articles pertaining to the reliability and validity of rubrics. They found only seven studies reported on intrarater reliability and most of the studies used Cronbach's alpha. More than half the articles reported interrater reliability using either consensus estimates (agreements), consistency estimates (correlations), or measurement estimates (generalizability theory, many-facet Rasch measurement, etc.). Their systematic review of the literature also showed that most of the research articles reported some form of validity. The most reported validity measures were criterion validity, content validity (expert review), and construct validity. Likewise, Reddy and Andrade (2010) found that many researchers reported the interrater reliability but omitted the process of the development of the rubrics. They also found that training raters to use the rubrics was needed to “achieve acceptable levels of reliability” (p. 445). Reddy and Andrade discovered that many researchers did not report on the validity of the rubrics, and recommended content validity, construct validity, and criterion validity as important aspects to support the quality of rubrics.

The aim of the current research was to develop and establish reliability and validity of a rubric that was used during mixed reality virtual simulations to evaluate Assessment HLP5: interpret and communicate assessment information with stakeholders to collaboratively design and implement educational programs. The rubric was based on the following four domains a) assessment description, b) assessment interpretation, c) SMART goal, and d) collaborative plan. The domains could be used for evaluating the level of assessment HLPs implemented in preservice teacher programs. The following research question was investigated: How was the HLPR-A sufficiently reliable and valid for use with preservice teachers?

## 2. Methodology

### *Rubric Development*

The HLPR-A has gone through several changes since its original development in 2019. Initially, the second author developed the first draft of the rubric based on operationalized skills she wanted students to learn and demonstrate during a mixed-reality simulation in an assessment course she taught. The skills were identified from assessment course objectives in conjunction with the literature on how to work collaboratively with families (Jones & Peterson-Ahmad, 2017) and included the constructs of personal exchange, transition, assessment description, assessment interpretation, student performance, targeted areas of concern, collaborative plan, and follow up. After using the original rubric with students, the second author realized there were too many skills to practice during the 8- 10-minute assessment conference, so the rubric was revised and “transition” was removed. Then, after the second draft of the rubric was used in another assessment course in 2020 to evaluate and teach the targeted skills, three researchers revised the rubric for the third time. The “personal exchange” construct was removed to allow the preservice teachers to focus more on the targeted skills of “assessment description” and “assessment interpretation” and less on building rapport through a personal exchange (see Table 1). While the researchers certainly acknowledge building rapport with families is important in establishing positive relationships, the course objectives and purpose of the rubric and simulations emphasized assessment description and interpretation to stakeholders. Furthermore, the researchers noticed that the preservice teachers were spending more time practicing the “personal exchange” during the simulations than discussing assessment information and wanted to help the preservice teachers focus on practicing the targeted HLP. Researchers also collapsed “student performance” into “assessment interpretation,” “follow up” into "collaborative plan," and removed the "exceeds" column as preservice teachers were learning the skills on the rubric and were not demonstrating exceeding behaviors, nor were they expected to (see Table 2).

Table 1. Brief Description of the Four Constructs in the Rubric

| Construct | Description |
|---|---|
| Assessment Description | Preservice teachers will briefly describe the two assessments including the names of assessments and the skills assessed on both assessments, as well as explain the purpose of both assessments. |
| Assessment Interpretation | Preservice teachers will discuss the child's current level of performance on assessments, will compare to grade level, and normed scores for context, and will identify both strengths and work areas. |
| SMART Goal | Preservice teachers will create 1 SMART goal to support area(s) of concern that is specific, measurable, attainable, realistic, and time-bound (SMART). |
| Collaborative Plan | Preservice teachers will collaboratively design an action plan that includes identification of strategy for school and for home. Students will plan to follow up with the parent by a specific date. |

Table 2. An Example of the Grade Levels Within a Construct of the Rubric

| | Proficient (3) | Developing (2) | Needs Improvement (1) |
|---|---|---|---|
| **Assessment Description** | □ Identified both assessment names | □ Identified one of the assessment names | □ Did not identify any assessment names |
| | □ Described the skills assessed on both assessments | □ Described some of the skills assessed on one assessment | □ Did not describe any of the skills assessed on the assessments |
| | □ Explained the purpose for both assessments | □ Explained the purpose of only one assessment | □ Did not explain the purpose for giving the assessments |

### *Participants*

Participants were traditional students, ages 18-22, and non-traditional students, over the age of 22, enrolled in the teacher education Elementary Special Education program at a private university in the southeast. There were 33 participants that varied in gender, race, and ethnicity. Participants were enrolled in the EDUC 451, Assessment and Evaluation in Special Education, and EDUC 333, Curriculum-based Assessment, courses. The two courses used a blended format in which instruction took place face-to-face in a classroom as well as synchronously online via Zoom. All participants provided permission to allow the researchers to watch and score their video recordings for this study. To maintain confidentiality, the researchers assigned ID numbers to all participants. After data analysis was completed, the researchers deleted the video recordings.

### Context

The study took place in two courses and while the courses were in different programs at the university, they had similar course objectives. Table 3 presents a crosswalk that aligns the Council for Exceptional Children Initial Preparation standard, high leverage practice, course objectives, and targeted skills on the rubric.

Table 3. Crosswalk of Standards, HLP, Course Objectives, & Targeted Skills

| 4.2 CEC K-12 Initial Practice Based Professional Standards | High Leverage Practice | Assessment Course Objectives | Targeted Skills |
|---|---|---|---|
| 4.1 Candidates collaboratively develop, select, administer, analyze, and interpret multiple measures of student learning, behavior, and the classroom environment to evaluate and support classroom and school-based systems of intervention for students with and without exceptionalities.<br>4.3 Candidates assess, collaboratively analyze, interpret, and communicate students' progress toward measurable outcomes using technology as appropriate, to inform both short- and long-term planning, and make ongoing adjustments to instruction. | Interpret and communicate assessment information with stakeholders (i.e., other professionals, families, students) to collaboratively design and implement educational programs (McLeskey et al., 2017, p. 45). | 1. Demonstrate knowledge of measurement terms and principles and interpreting assessment results<br>2. Explain the use of assessments in the area of academic achievement | 1. Preservice teachers will briefly describe the two assessments including the names of assessments and the skills assessed on both assessments, as well as explain the purpose of both assessments.<br>2. Preservice teachers will discuss the child's current level of performance on assessments, will compare to grade level, and normed scores for context, and will identify both strengths and work areas.<br>3. Preservice teachers will create 1 SMART goal to support area(s) of concern that is specific, measurable, attainable, realistic, and time- bound (SMART).<br>4. Preservice teachers will collaboratively design an action plan that includes identification of strategy for school and for home. Students will plan to follow up with the parent by a specific date. |

The rubric was used as a teaching tool to help preservice teachers learn how to communicate assessment results to a parent by breaking down the HLP into targeted skills that aligned with course objectives. The rubric was also used as a self, peer, and instructor evaluation tool to provide preservice teachers feedback on their performance of the assessment conference practices. The preservice teachers conducted a simulated parent-teacher conference where they discussed the results of several literacy assessments of a struggling student. The preservice teachers engaged in a practice roleplay simulation where they were the teacher and had to explain the assessment results to another preservice teacher who was acting as the parent of a student. Then, the preservice teachers engaged in mixed-reality virtual simulations where they interacted with an avatar parent who was controlled by an actor through a simulator called TeachLivE.

### *Raters*

The five raters who evaluated the videos using the HLPR-A included an associate professor, two assistant professors, a post-doctoral scholar, and a doctoral student. The associate professor specialized in literacy research and contributed to the design of the research study, development of the rubric, evaluation of the videos, and procuring expert review. One of the assistant professors specialized in quantitative research methodology. The other assistant professor was a special education expert and taught the courses. Both assistant professors contributed to the design of the research study, development of the rubric, evaluation of the videos, and data analysis. The post-doctoral scholar was an expert in mixed-reality virtual simulation and contributed to evaluating the videos. The doctoral student specialized in quantitative research methodology and contributed to evaluating the videos and analyzing the data. All five raters scored the 33 videos twice using the HLPR-A.

### *Rater Training*

All five raters participated in training of the HLPR-A. Rater training took approximately three hours. First, the researchers described and explained the study's design and virtual simulation scenario. Next, the researchers presented, described, and explained the rubric. All five raters then watched one video and independently scored the HLPR-A. Each rater shared their score with the other raters. The raters discussed discrepant criteria scores until all raters agreed. The same procedure continued for four more videos until there was 90% agreement.

### *Procedure*

The researchers implemented eight procedures. (1) The researchers trained the raters on the content of the parent teacher conferences and the HLPR-A. (2) The participants practiced a parent teacher conference emphasizing assessment description and interpretation of data to a parent about their child in a virtually simulated environment. (3) The virtually simulated parent teacher conferences were video recorded. (4) The video recordings of the participants' parent teacher conferences were watched and scored by all the raters using the HLPR-A. (5) The researcher consolidated the raters' scores into one SPSS (Statistical Package for Social Science) data file for analysis. (6) Six weeks later, the raters watched and scored the video recordings of the participants' parent teacher conferences a second time using the HLPR-A. (7) The researchers consolidated the raters' scores into one SPSS data file for analysis. (8) The researchers analyzed the data to establish reliability and validity of the HLPR-A.

### *Data Analysis*

Data was collected from the HLPR-A. All raters watched and scored the videos twice using the specific criteria on the HLPR-A. The HLPR-A scale of 1-3 (needs improvement, developing, or proficient) was used. The scores on the rubric were averaged by rater to calculate an intraclass correlation coefficient (ICC). The HLPR-A data was also used to calculate the standard error of measurement (SEm), to determine minimal detectable change (MDC), and to conduct a principal component analysis (PCA).

## 3. Results and Discussion

### *Reliability*

Reliability of a rubric refers to the consistency of scoring (Field, 2018, Johnson & Morgan, 2016). The researchers calculated interrater reliability, similar scores among different raters, and intrarater reliability, similar scores by the same rater (Field, 2018; Von Helvoort, et al., 2017), by average score and among each construct on the HLPR-A. Independently, each rater scored the 33 videos using the HLPR-A and then each rater scored the 33 videos a second time six weeks later.

### *Interrater Reliability*

The researchers established interrater reliability using an ICC (Field, 2018) among all five raters' average scores on the HLPR-A from all 33 participants. The ICC estimates and their 95% confidence intervals (CI) were calculated using SPSS statistical package version 28 based on the mean rating ($k$ = 5), absolute-agreement, 2-way mixed-effects model (Koo & Li, 2016). "ICC values less than 0.5 are indicative of poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.9 indicate good reliability, and values greater than 0.90 indicate excellent reliability" (Koo & Li, 2016, p. 158). The average scores' ICC (3, 5) was .90, 95% CI [*.83, .95*], suggesting good to excellent interrater reliability between the five raters and their average scores on the HLPR-A. The raters' average scores ranged from a minimum score of 5.49 to a maximum score of 5.90. The results suggest that the raters scored the HLPR-A in an equivalent manner.

The researchers established interrater reliability using an ICC among all five raters' construct scores (assessment description, assessment interpretation, SMART goal, and collaborative plan) on the HLPR-R from all 33 participants. The ICC estimates and their 95% CI were calculated based on the mean rating (*k* = 5), absolute-agreement, 2-way mixed-effects model. The assessment description ICC (3, 5) was 0.73, 95% CI [.*56*, .*85*], suggesting moderate to good interrater reliability. The assessment interpretation ICC (3, 5) was .78, 95% CI [.*64*, .*88*], suggesting moderate to good interrater reliability. The SMART goal ICC (3, 5) was .78, 95% CI [.*62*, .*88*], suggesting moderate to good interrater reliability. The collaborative plan ICC (3, 5) was .92, 95% CI [.*86*, .*96*], suggesting good to excellent interrater reliability.

### *Intrarater Reliability*

Intrarater reliability refers to consistent scoring among the same rater over time, otherwise known as test-retest reliability (Jonsson & Svingby, 2007). The researchers calculated intrarater reliability using an ICC, CI, SEm, and MDC for each rater by average score and for each rater by construct on the HLPR-A. SEm refers to how a person's multiple scores on the same instrument are distributed around their true score (Alghadir et al., 2015). MDC estimates the amount a rater's score needs to change to show significant differences from the first to the second rating while accounting for measurement error (Alghadir et al., 2015). Table 4 shows the ICC, CI, SEm, and MDC for each rater by average score and each construct on the HLPR-A.

Table 4. ICC, CI, SEm, and MDC by Rater

| | **ICC (3,1)** | **CI** | **SEm** | **MDC** |
|---|---|---|---|---|
| Rater 1 | | | | |
| Average Score | .92 | .84, .96 | 0.36 | 0.51 |
| assessment description | .90 | .81, .95 | 0.63 | 0.90 |
| assessment interpretation | .65 | .40, .81 | 1.09 | 1.55 |
| SMART goal | .72 | .51, .85 | 0.46 | 0.65 |
| collaborative plan | .88 | .77, .94 | 0.89 | 1.26 |
| Rater 2 | | | | |
| Average Score | .89 | .74, .95 | 0.44 | 0.63 |
| assessment description | .55 | .27, .75 | 1.21 | 1.72 |
| assessment interpretation | .83 | .69, .91 | 0.88 | 1.25 |
| SMART goal | .75 | .52, .88 | 0.82 | 1.16 |
| collaborative plan | .71 | .48, .85 | 1.62 | 2.30 |
| Rater 3 | | | | |
| Average Score | .85 | .18, .95 | 0.64 | 0.91 |
| assessment description | .64 | .38, .80 | 0.97 | 1.37 |
| assessment interpretation | .49 | .19, .71 | 1.72 | 2.44 |
| SMART goal | .60 | .20, .80 | 1.26 | 1.78 |
| collaborative plan | .78 | .54, .89 | 1.79 | 2.53 |
| Rater 4 | | | | |
| Average Score | .86 | .60, .94 | 0.66 | 0.93 |
| assessment description | .35 | .01, .62 | 1.56 | 2.20 |
| assessment interpretation | .50 | .20, .71 | 1.56 | 2.20 |
| SMART goal | .38 | .04, .64 | 1.13 | 1.59 |
| collaborative plan | .90 | .82, .95 | 1.05 | 1.48 |
| Rater 5 | | | | |
| Average Score | .95 | .90, .98 | 0.31 | 0.44 |

| | | | | |
|---|---|---|---|---|
| assessment description | .87 | .75, .93 | 0.75 | 1.06 |
| assessment interpretation | .86 | .73, .93 | 0.82 | 1.17 |
| SMART goal | .94 | .87, .97 | 0.24 | 0.34 |
| collaborative plan | .85 | .70, .93 | 1.18 | 1.67 |

Analysis of intrarater reliability showed good to excellent reliability (.85 to .95) across all raters for average scores on the rubric. By construct, raters 1, 2, and 3 showed moderate to good interrater reliability across all constructs. Rater 4, however, showed poor interrater reliability across three of the four constructs. Rater 5, on the other hand, showed good to excellent interrater reliability across all constructs. Raters' SEm scores ranged from 0.24 to 1.79 suggesting most raters repeated scores were close to their true scores. Raters' MDC ranged from 0.34 to 2.53 suggesting some variation in the rubric. Overall, intrarater reliability was good proposing the HLPR-A is a reliable tool for assessing students' performance.

### *Validity*

The validity of a rubric determines if the rubric is accurately assessing what it is supposed to assess (Firmansyah et al., 2020; Jonsson & Svingby, 2007). The researchers established two types of validity: content validity and construct validity. Content validity is the extent that the assessment instrument measures what it is intended to measure (Firmansyah et al., 2020; Jonsson & Svingby, 2007). Content validity was established through expert review. Construct validity is the notion that rubric's constructs are measuring the domain it is supposed to measure (Cronbach & Meehl, 1955). Construct validity was established using a nomological network and principal axis factor analysis.

### *Content validity*

To establish content validity, do the constructs and criteria within the rubric represent the assessment HLP, five additional content experts (two professors, one K-5 practitioner, one K-5 administrator, and one parent of a student with a disability) were asked to review the HLPR-A and examine each construct and criteria. Experts provided feedback and suggestions about the HLPR-A about what they thought should be changed or included. The professors and the elementary practitioner stated that the HLPR-A appeared sufficient and made no further recommendations for change. The parent suggested changing the term *resource* to *strategy* and the research team decided to include both terms in the HLPR-A. The parent also suggested the HLPR-A include the option of using either normed scores or grade level scores. The principal feedback was a repair of a grammatical error. None of the recommendations changed the overall intent of the HLPR-A nor any of the constructs. An expert review of the rubric provided content validity evidence.

### *Construct validity*

There are three assessment HLPs. The assessment HLP 4 focuses on using multiple sources of information to develop an understanding of students' needs.

The assessment HLP 6 focuses on using students' assessment data to analyze and adjust instructional practices. The assessment HLP 5, and the rubric of interest for this study, focuses on interpreting and communicating assessment information with the stakeholder to collaboratively design and implement educational programs. HLP 5 states:

> Teachers interpret assessment information for stakeholders (i.e., other professionals, families, students) and involve them in the assessment, goal development, and goal implementation process. Special educators must understand each assessment's purpose, help key stakeholders understand how culture and language influence interpretation of data generated, and use data to collaboratively develop and implement individualized education and transition plans that include goals that are standards-based, appropriate accommodations and modifications, and fair grading practices, and transition goals that are aligned with student needs. (McLeskey et al., 2017, p. 45)

HLP 5 differed from the other two assessment HLPs, 4 and 6. While HLP 4 targeted developing an understanding of students' needs and HLP 6 highlighted instructional practices, HLP 5 emphasized interpreting and communicating assessment information for stakeholders. Although the researchers acknowledge that HLPs 4 and 6 are both important, the researchers were more interested in participants' ability to interpret, explain, and communicate assessment results and work collaboratively with parents and/or caregivers.

Construct validity of the HLPR-A was established using a nomological network (Cronbach & Meehl, 1955). The nomological network provided a method to organize and structure the themes, constructs, domains, and criteria of the rubric (see Figure 1). The researchers identified five broad themes in the assessment HLP 5: collaboration, understanding, communication, goal setting, and interpretation. The researchers decomposed the five broad themes into seven explicit and measurable constructs: personal exchange, transition, assessment description, assessment interpretation, target areas of concern, SMART goal, collaborative plan, and follow up. Understanding and communication themes were linked to all seven constructs and dual arrows were used to indicate the reciprocal nature of communication between teachers and parents and/or caregivers. The interpret theme was linked to the construct of assessment interpretation. Goal setting was connected to both SMART goal and collaborative plan because preservice teachers had to create a SMART goal and then discuss it with the parent as part of the collaborative planning process. The collaborate theme overlapped with the constructs of collaborative plan and follow up because the preservice teachers were practicing their collaboration skills during this time in the conference.

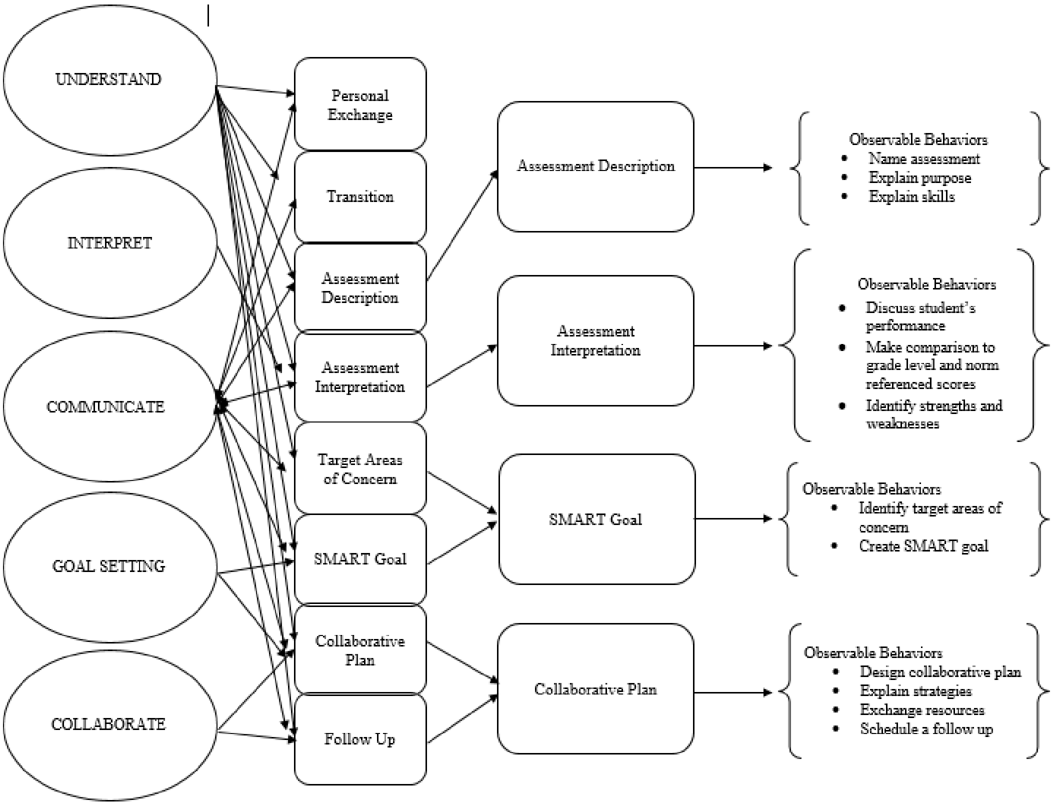

Figure 1. Nomological Network of Broad Themes, Constructs, Domains, and Criteria

The researchers collapsed and omitted the explicit and measurable constructs into a manageable rubric (HLPR-A) that could assess participant performance within the allotted time. Because the rubric was going to be used to assess a targeted practice experience focused on interpreting and communicating assessment that lasted only 8 to 10-minutes, the researchers chose to eliminate personal exchange and transition. Although participants naturally engaged in brief personal exchanges to begin their conference anyway, the emphasis was not on building rapport but interpreting and communicating assessment results. The removal of personal exchange and transition ensured that the participants would not spend too much time focusing on building rapport and not move on to the other aspects of the assessment conference. The researchers also collapsed target areas of concern and SMART goal into one domain labeled SMART Goal. The researchers used the rubric to help the preservice teachers build an understanding of the domains so combining target areas of concern with the SMART goal helped link those two constructs together more explicitly. Finally, the researchers also collapsed collaborative plan and follow up into one domain labeled collaborative plan because following up with the parent and/or caregiver was a part of making a collaborative plan and did not require its own domain.

The data consisted of 165 data points from the individual rubrics scored by the research team. The HLPR-A had 10 criteria within four general domains (assessment description, assessment interpretation, SMART goal, and collaborative plan). Each domain had three criteria except the SMART goal, which had only one criteria. Each participant could have scored a 1 (needs improvement), 2 (developing), or 3 (proficient) for each criteria. The researchers conducted a PCA to determine how many significant components there were using Kaiser's (1960) eigenvalue one criterion, the scree plot, and variance explained. Since the domains were correlated, the researchers used oblique rotation (direct oblimin) to estimate significant values (Field, 2018). The determinant value was .17 and Bartlett's test of sphericity, which tests the significance of all the correlations, was significant, □$^2$ (45) = 279.51, *p* < .001. The Kaiser-Meyer-Olkin (KMO) measures of sampling adequacy indicated that the strength of the relationships among the constructs was moderate, KMO = .67. Based on Kaiser's (1960) eigenvalue one criteria, two significant components accounted for 42.31% of the variance (see Table 5). The first component had an eigenvalue of 2.73 and accounted for 27.25% of the variance. The second component had an eigenvalue of 1.51 and accounted for a further 15.06% of the variance. The scree plot of the eigenvalues also supported two components (see Figure 2).

Table 5. Eigenvalues

| **Number** | **Eigen Value** | **Proportion Explained** | **Cumulative Proportion Explained** |
|---|---|---|---|
| 1 | 2.73 | 27.25 | 27.25 |
| 2 | 1.51 | 15.06 | 42.31 |

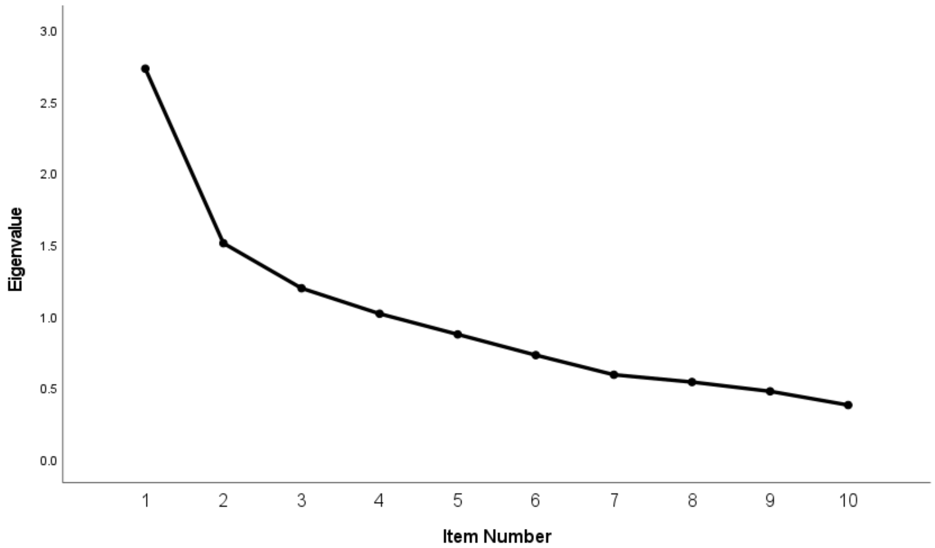

Figure 2. Scree Plot

The PCA of the 10 items on the HLPR-A used in the current study revealed two components were sufficient to explain the underlying structure of the assessment HLP. The pattern matrix in Table 6 revealed component one to consist of five items. This component was labeled Plan. The second component consisted of five items and was identified as Assessment. Overall, PCA of the HLPR-A items revealed that all items loaded into one of two components. The two components revealed through PCA allowed the researchers to collapse Assessment Description and Assessment Interpretation into a broader domain labeled Assessment and collapse SMART Goal and Collaborative Plan into a broader domain labeled Plan, which supported the validity of the constructs of the HLPR-A (see Appendix).

Table 6. Summary of Principal Component Analysis Results for the HLPR-A (N = 165*)*

| | Rotated Pattern Matrix Loadings | |
|---|---|---|
| Item | **Plan** | **Assessment** |
| Item 9 | **0.72** | 0.13 |
| Item 7 | **0.68** | -0.21 |
| Item 10 | **0.65** | -0.09 |
| Item 8 | **0.65** | -0.05 |
| Item 6 | **0.49** | 0.09 |
| Item 2 | -0.12 | **0.76** |
| Item 1 | -0.11 | **0.70** |
| Item 4 | 0.21 | **0.66** |
| Item 5 | 0.23 | **0.58** |
| Item 3 | -0.01 | **0.32** |

## 4. Conclusion

The purpose of the study was to establish reliability and validity of the high leverage practice rubric HLPR-A. High leverage practices are difficult to teach and assess and offering a reliable and valid rubric provides the field with a consistent measure of performance for both the learners and teachers. The HLPR-A was found to be both reliable and valid instrument for assessing the Assessment HLP5 during simulations based on the way that researchers decomposed the five broad themes into seven explicit and measurable constructs.

The current study addressed two of the gaps in the literature regarding using rubrics as instruments. First, the reliability and validity of rubrics used in research are often not reported (Firmansyah et al., 2020; Reddy & Andrade, 2010). The researchers provided details about the development of the HLPR-A and the reliability and validity procedures, analysis, and results. Reliability was established using interrater and intrarater reliability. Interrater reliability was good to excellent; when a rater scored low on the rubric, all the raters scored low on the rubric and vice versa. Intrarater reliability showed good reliability of the HLPR-A; the raters scored the rubrics in a similar fashion after six weeks. The researchers provided the ICC information as suggested by Koo and Li (2016) and rater-training procedures which, according to Johnson and Morgan (2016), leads to a well-designed and reliable rubric. Validity was established using expert review, nomological network, and PCA. Expert review provided content validity (Firmansyah et al., 2020). The nomological network allowed the researchers to align the criteria on the rubric with the themes, constructs, and domains being assessed (Beachcroft-Shaw & Ellis, 2020; Cronbach & Meehl, 1955; Firmansyah et al., 2020; Reddy & Andrade, 2010). PCA allowed the researchers to identify significant and relevant components within the rubric (Field, 2018).

Second, researchers report that rubrics are primarily used as an evaluation tool and the utility of rubrics for teaching purposes has been relatively ignored (Firmansyah et al., 2020; Reddy & Andrade, 2010). To address this gap in the literature, the HLPR-A was used as both a teaching tool and an evaluation tool. The rubric was given to preservice teachers as a representation of the deconstructed assessment HLP5 to aid them in developing their understanding and use of it. Additionally, the rubric served as a self, peer, and instructor evaluation tool that was used to provide feedback to preservice teachers on their performance. Using the HLPR-A as both a teaching and assessment tool supported the preservice teachers' use of self-regulatory behaviors as it allowed them to plan, monitor, and evaluate their use of their practice.

### *Limitations*

The researchers made the HLPR-A very specific to meet the criteria for this study. The researchers acknowledge that the specific criteria of the HLPR-A limits the generalizability of the rubric to other populations, for other purposes, or other raters. However, the researchers provided detailed descriptions of how to decompose, develop, and establish reliability and validity of any HLP rubric. The researchers also note that criterion-related validity could not be established because the researchers could find no other valid HLP rubric to compare the HLPR-A to. The researchers are also aware that the raters' mindset and attitude while scoring the rubric varied. Lastly, the researchers narrowed the PCA down to two broad terms because the preservice teachers used the terms description and interpretation interchangeably when discussing assessment scores with a parent. The lack of distinction could be attributed to the preservice teachers beginning understanding of the complex practice and may show different results if used with more veteran teachers.

### *Future Research*

Future research should explore using the HLPR-A on other populations, with a variety of raters, in other environments, and programs. Criterion-related validity should be established between and among other assessment HLP rubrics using the HLPR-A. The researchers recommend adding an exceeding column to the HLPR-A if working with inservice teachers. Finally, a review of the literature revealed no consistent way to establish reliability and validity of an instrument or rubric. Since there are a variety of ways to establish reliability and validity of rubrics, a consensus and systematic approach to rubric credibility is needed. The consistency of rubric reporting is important because rubrics are a common assessment tool in higher education.

Faculty members need credible evaluation choices for measuring preservice teachers' high leverage practices. High leverage practices are foundational teaching practices in special education that preservice teachers should learn, however, only one other reliable and valid rubric has been identified to support preservice teachers' learning and evaluation of HLPs (Johnson et al., 2019). The HLPR-A has been found to be reliable and valid and offers the field of teacher education a blueprint for not only evaluating preservice teachers' rehearsals of an HLP but also their learning of the practice. The HLPR-A provides a decomposed look at the assessment HLP5 that gives preservice teachers an explicit example of each component of HLP5 and how they might do it. While the HLPR-A is certainly not the only way assessment HLP5 could be represented, it is simply the researchers' attempt to add a reliable and valid tool to the field to support the use of HLPs. The researchers encourage further examination of developing credible HLP rubrics that can contribute to the development of preservice teachers and their practices.

## Acknowledgement

This research was funded by Mercer University Provost Grant and there was no conflict of interest.

## References

Alghadir, A., Anwer, S., & Brismee, J. M. (2015). The reliability and minimal detectable change of timed up and go test in individuals with grade 1–3 knee osteoarthritis. *BioMed Central Musculoskeletal Disorders, 16*(1), 174-174. https://doi.org/10.1186/s12891-015-0637-8

Beachcroft-Shaw, J. & Ellis, D. (2020). Finding common ground: Mapping the nomological networks of sustainability constructs for improved social marketing. *Sustainability Science, 15,* 745-758. https://doi.org/10.1007/s11625-019-00755-z

Council for Exceptional Children. (2020). Initial Practice Based Professional Standards for Special Education.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed). Sage.

Firmansyah, Dr. R., Nahadi, N, & Firman, H. (2020). Development of performance assessment instruments to measure students' scientific thinking skill in the quantitative analysis acetic acid levels. *Journal of Educational Science, 4*(3), 459-468 https://doi.org/10.31258/jes.4.3

Gallardo, K. (2020). Competency-based assessment and the use of performance-based evaluation rubrics in higher education: Challenges towards the next decade. *Problems of Education in the 21st Century, 78*(1), 61-79. https://doi.org/10.33225/pec/20.78.61

Johnson, R. L. & Morgan, G. B. (2016). *Survey Scales: A guide to development, analysis, and reporting.* Guildford Press.

Johnson, E. S., Zheng, Y., Crawford, A. R., & Moylan, L. A. (2019). Developing an explicit instruction special education teacher observation rubric. *The Journal of Special Education, 53*(1), 28-40. https://doi.org/10.1177/0022466918796224

Jones, B. A., & Peterson-Ahmad, M. B. (2017). Preparing new special education teachers to facilitate collaboration in the individualized education program process through mini conferencing. *International Journal of Special Education, 32*(4), 697-707. https://files.eric.ed.gov/fulltext/EJ1184062.pdf

Jonsson, A. (2014). Rubrics as a way of providing transparency in assessment. *Assessment and Evaluation in Higher Education, 39*(7), 840-852. https://doi.org/10.1080/02602938.2013.875117

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity, and educational consequences. *Educational Research Review, 2,* 130-144. https://doi.org/10.1016/j.edurev.2007.05.002

Kaiser, H. F. (1960). The application of electronic computer to factor analysis. *Educational and Psychological Measurement, 20*(1), 141-151. https://doi.org/10.1177/001316446002000116

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15,* 155-163. http://dx.doi.org/10.1016/j.jcm.2016.02.012

McLeskey, J., Barringer, M-D., Billingsley, B., Brownell, M., Jackson, D., Kennedy, M., Lewis, T., Maheady, L., Rodriguez, J., Scheeler, M. C., Winn, J., & Ziegler, D. (2017). *High-leverage practices in special education*. Arlington, VA: Council for Exceptional Children & CEEDAR Center. https://ceedar.education.ufl.edu/wp-content/uploads/2017/07/CEC-HLP-Web.pdf

Reddy, Y. M. & Andrade, H. (2010). A review of rubric use in higher education. *Assessment & Evaluation in Higher Education, 35*(4), 435-448. https://doi.org/10.1080/02602930902862859

Van Helvoort, J., Brand-Gruwel, S., Hyusmans, F., & Sjoer, E. (2017). Reliability and validity test of a scoring rubric for information literacy. *Journal of Documentation, 7*3(2), 305-316. https://doi.org/10.1108/JD-05-2016-006

How to cite this article:

Ford, D. J., Luke, S. E., & Vaughn, S. M. (2022). Establishing Reliability and Validity of a High Leverage Practice Rubric for Assessment (HLPR-A). *Journal of Educational Sciences, 6(4),* 707-722.