Teaching and Learning in Communication Sciences & Disorders

Volume 3 | Issue 2 Article 8

2019

Toward the Development of a Quick, Reliable Assessment Tool for Reflective Journals

April Garrity
Georgia Southern University, Armstrong Campus, agarrity@georgiasouthern.edu

Casey Keck
Georgia Southern University, Armstrong Campus, ckeck@georgiasouthern.edu

Janet L. Bradshaw
Georgia Southern University, Armstrong, jbradshaw@georgiasouthern.edu

See next page for additional authors

DOI: https://doi.org/10.30707/TLCSD3.2Garrity

Toward the Development of a Quick, Reliable Assessment Tool for Reflective Journals

Abstract

Reflective practice, including reflective writing, can facilitate enriched learning, especially when implemented as part of a service-learning (SL) experience. Reflection is a central component of service-learning (SL) experiences. Students' reflective abilities are often measured through reflective journaling; however, assessment of students' reflective journals is not always efficient and straightforward. The goal of the present study was to establish a simple, reliable, and relatively quick tool for use by busy college instructors seeking to encourage students' deep learning through reflective writing. A total of 258 reflective journals from 43 graduate students in speech-language pathology were evaluated by three raters using a three-tier assessment framework (*nonreflection, reflection, and critical reflection*) after Mezirow (1990) and Plack et al. (2005). Although previous studies found moderate to high interrater agreement and reliability, the current study did not support this finding. Strengths and weaknesses of the assessment framework and qualitative observations of the assessment process are discussed.

Keywords

service-learning, reflective practice, reflective journals, health professionals, speech-language pathology, assessment

Cover Page Footnote

ACKNOWLEDGEMENTS The authors wish to acknowledge our former student, Ellisa Davis, B.S., who assisted tremendously with data coding and analysis; and, Kaye Wimberly, who provided valuable feedback on an earlier draft of the manuscript.

Authors

April Garrity, Casey Keck, Janet L. Bradshaw, and Keiko Ishikawa

A number of scholars have developed theoretical frameworks to help us understand how various metacognitive skills relate to the process of learning how to think critically, and how we, as college instructors, might teach and assess these skills (Bloom & Krathwohl 1956; Dewey, 1916, 1938; Kolb, 1984; Schön, 1983,1987). One practice that is prominently represented in these frameworks is that of reflection. The ability to engage in reflective thinking is an important aspect of, and possibly a precursor to, critical thinking (Choy & Oo, 2012). Students who engage in reflection are cognizant and attuned to their learning by evaluating what they understand, what they need to learn, and how to apply the knowledge (Sezer, 2008).

Reflective practice is also a critical component of service-learning (SL). SL is an instructional technique that utilizes community-based opportunities to highlight and augment the academic content presented in a course (Stacey, Rice, & Langer, 2001). In communication sciences and disorders, reflective practice may give richer meaning to clinical experiences within the SL paradigm. Reflective practice, including reflective writing, can facilitate students' descriptions and analytical reflections on clinical experiences (Boud, 2001; Jarvis, 1992, 2001). Supporters of reflective writing suggest that the practice can assist new clinicians in cultivating a proficiency in content knowledge and clinical application (Kerka, 2002; Schön, 1987).

Understandably, much of the research dedicated to the assessment of students' written reflections has focused on student learning outcomes. Indeed, the evaluation of reflective journals is not always a systematic process (Boud, 2001; Woodward, 1998) and may be considered too subjective (Bourner, 2003), calling into question the validity and reliability of current assessment frameworks and tools. Instructors might also encounter difficulty with the time commitment required for reading and evaluating reflective writing assignments. The process of grading these types of assignments may be quite time-consuming, even prohibitive, on top of other course-related responsibilities.

Given the complexities associated with the assessment of students' reflective journals, and in light of the evidence supporting the benefits of their use to facilitate enriched learning, especially in the implementation of SL experiences, we were compelled to ask if current reflection assessment frameworks might be used more objectively and efficiently. In this paper, we will discuss our investigation of this question, first considering relevant theoretical frameworks and evidence related to speech-language pathology in this area of inquiry, then describing the use and subsequent modification of an existing framework to evaluate student reflections on a service-learning (SL) experience.

Literature Review

In order to determine how to best assist students with developing their reflection skills, thus maximizing significant learning opportunities, instructors must utilize a valid and reliable assessment framework (Hill, Davidson, & Theodoros, 2012). Although the difficulties of evaluating students' reflections are well-documented (Boud, 2001; Plack & Greenberg, 2005), a number of assessment frameworks have been proposed and studied (see Ash & Clayton, 2004; Boud, Keogh, & Walker, 1985; Hatton & Smith, 1995; Kember, McKay, Sinclair, & Wong, 2008; Mezirow, 1990; Plack, Discoll, Blisset, Mckenna, & Plack, 2005; Scanlan & Chernomas, 1997; Valli, 1997). These frameworks tend to share many similarities in terms of the characteristics that

are indicative of different levels of reflection. However, they differ on factors including complexity (e.g., the number of elements or components in the framework) as well as how they have been implemented and studied (e.g., the tasks upon which reflections were based and whether writing prompts were used). Relatively few studies have addressed the assessment of reflective writing in our discipline of speech-language pathology. Therefore, this review will focus on several different frameworks that have been used to assess and study reflective journals among students in other health professions as well as our own.

Seminal Frameworks: Boud et al. (1985), Mezirow (1990), And Schön (1987). Three frameworks that have influenced the evaluation of students' reflection journals in speech-language pathology and related fields are Boud et al. (1985), Mezirow (1990), and Schön (1987). Boud et al. (1985) and Mezirow (1990) proposed assessment frameworks based on three levels of reflection. The first and lowest level of reflection in these frameworks involves a re-telling of the objective experience. The second level requires the presence of more subjective components, typically including a description of emotions relating to the experience. The third level of reflection in these frameworks includes an evaluation or analysis of the objective and subjective components of the experience.

Although the levels of reflection are similarly organized, the frameworks differ in their descriptions of these levels. Specifically, Boud et al. (1985) described these levels of reflection as: *returns to experience*, *attends to feelings*, and *reevaluates the experience*. Mezirow's (1990) model includes three reflection types (*content reflection, process reflection*, and *premise reflection)* as well as three holistic rating categories (*nonreflection, reflection,* and *critical reflection).* These frameworks have been found to be at least moderately reliable when applied to written reflections of health professions students (Williams, Sundelin, Foster-Seargeant, & Norman, 2000; Wong, Kember, Chung, & Yan, 1995).

Lastly, Schön (1987) proposed a three-level framework based on when the reflection takes place. This framework includes the descriptors *reflection in action*, *reflection on action*, and *reflection for action*. *Reflection in action* refers to objective observations during the experience. *Reflection on action* takes place following the experience and is the consideration of newly learned information in conjunction with the applied experience. Finally, in *reflection for action,* an individual consolidates emotions and information from past experiences to make predictions about and plan for future experiences.

Comprehensive Applications of the Seminal Frameworks. At least two studies (Hill et al., 2012; Plack et al., 2005) have examined the application of these seminal frameworks in the evaluation of reflective journals of health professions students. These studies tested the reliability of an integrated approach, in which the seminal frameworks were combined into two levels of assessment.

The two-level approach assessed students' reflection skills at elemental and holistic levels. Level I was used to evaluate the textual level (words, sentences, and paragraphs), and included three general themes (time, content, and stage), which were organized into a total of nine different elements. Time-dependent elements (Schön, 1987) were used to assess when the reflection took place, and therefore, when a student might use the reflection to change a behavior. The time-

dependent elements included *reflection in action, reflection on action,* and *reflection for action*. Content-dependent elements (Mezirow, 1990) indexed the student's understanding of an experience, strategies used for problem solving, and ability to think critically about his or her own biases. These elements consisted of *content reflection, process reflection,* and *premise reflection.* Finally, stage-dependent elements (Boud et al., 1985) indicated a student's level of engagement with an experience and the degree to which he or she constructed and integrated meaning from the experience. The stage-dependent elements were: *returns to experience, attends to feelings*, and *reevaluates the experience*.

Level II coding assigned an overall or holistic rating to each reflective journal: *no evidence of reflection, evidence of reflection,* or *evidence of critical reflection*. These ratings were operationally defined by the Plack et al. (2005)(see Table 1).

Plack et al. (2005) used this integrated approach to retrospectively analyze the reflective journals of 27 physical therapy students' clinical experiences. The data set was comprised of 43 reflective journals. Prior to writing their reflections, the students were given information about the importance of reflection in practice and were directed to make "consistent journal entries" (Plack et al., 2005, p. 202). Students were provided with "reflection starters" (Plack et al., 2005, p. 202) to assist with their writing to use as needed.

Several measures of interrater reliability, including percent agreement and intraclass correlation coefficient (ICC), were applied to the data. Across all three raters for the Level I ratings (nine elements of reflection), percent agreement ranged from 65.1% to 93.0%. For the Level II three-tier rating system (evidence of reflection), agreement between pairs of raters ranged from 67.4% to 85.7%. Their statistical analyses indicated that both Level I (the comprehensive, nine-element framework), and Level II (the holistic three-level framework) yielded moderate to high levels of interrater reliability.

This same two-level integrated approach was used by Hill et al., 2012 to code the reflective journals of 52 speech-language pathology students. Level I of coding was referred to as *breadth of reflection* and Level II as *depth of reflection*. These authors used the same operational definitions as those used by Plack et al. (2005)(see Table 1).

Each of the students in the study wrote a total of three reflective journals, which followed interviews with three different standardized patients (SPs). The students were provided with guiding questions to facilitate their reflections. The reflective journals of ten students ($n = 70$) were used to calibrate the raters, leaving the reflective journals of 42 students ($n = 126$) to be included for analysis.

Table 1
Evaluation Rubric with Definitions (Mezirow, 1990; Plack et al., 2005) and Additional Criteria/Examples By The Authors of This Study

Rating	Definition	Additional criteria	Examples
Nonreflection	No evidence of reflection is present within the journal. May see description of experiences with no evidence of evaluation/questioning of the experience.	N/A	"I learned a lot about all the different strategies she uses to overcome the challenges of daily activities of living around the house and outside the home setting."
Reflection	Evidence of reflection is present in the journal. Writer reflects so to better understand the situation, or decide how best to perform; writes beyond describing/reporting experiences.	Writer provides a label and description for emotions evoked by the experience. Writer provides a general judgment or prediction about the experience. Writer makes a comparison but does not draw conclusions based on the comparison.	"Many of the students, including myself, addressed some of the things that we do when we are in loud and crowded restaurants and grocery stores. I believe this may help them feel as if they are not the only ones who need compensatory strategies in our daily lives."

Critical Reflection	Evidence of critical reflection is present within the journal. Writer explores the existence of the problem, where the problem stems from, or the assumptions underlying the problem. Writer will critique his/her experiences, assumptions and may begin to show evidence of modifying his/her own biases or assumptions.	Writer provides specific details and/or examples of how/what will be modified. Writer provides underlying reasons for why or why not something happened Writer draws specific conclusions based on experience	"This discussion made me realize that in therapy, I should try harder to focus more on the functional communication that targets their hobbies and activities that are important in their daily lives." "Therapy was client-centered which was different from any other session thus far. We decided to do this because we want to know more about what the participants struggle with since they are the ones that had the stroke. They know more about their difficulties than we do because they experience them every day."

Hill et al. (2012) reported high rates of agreement for both Level I (*breadth of reflection*) ratings and Level II (*depth of reflection*) across reflective journals for all three SP interviews. Level I interrater agreement ranged from 81.48% to 98.77% (*M* = 91%), with ICC kappas ranging from -0.033 to 1. Level II interrater agreement ranged from 33.33% to 100% (*M* = 96.03%), with kappas ranging from 0.481 to 1. The authors used the Landis and Koch (1977) interpretation of their kappa values, which indicated poor agreement to almost perfect agreement for Level I ratings and fair agreement to almost perfect agreement for Level II ratings. Hill et al. (2012) replicated the significant interrater agreement reported by Plack et al., (2005), lending further evidence to support the use of the comprehensive two-level assessment framework.

Which Framework is "Best"? All of the assessment frameworks used in these previous studies were reported as having good interrater agreement and/or reliability. One of the major differences among them was the complexity of the framework (e.g., the number of elements to be rated). Determining the "best" assessment framework to use may be a particularly difficult undertaking. College instructors may need to give considerable thought to the time commitment required for implementation of an assessment framework. Neither of the studies reviewed here provided information about the time required to use the two-level framework they studied. However, given that this framework included 12 elements, one might assume the time commitment would be significant. College instructors have different assessment philosophies, needs, and workload demands, which complicates the question as to which framework is "best", and creates a situation in which we will almost certainly find differing answers to that question.

As busy college instructors at a teaching-focused institution who are interested in promoting deep transformative learning through experiential means with a significant reflective component, specifically an SL experience, we set out to determine if a previously established assessment framework would meet the needs of instructors in similar situations. Specifically, we were searching for the most parsimonious framework that would allow for relatively quick evaluation and provide an acceptable level of reliability.

Method

This study was approved by the Institutional Review Board of Georgia Southern University (Protocol ID: H18328). All students whose data were included in this study provided informed consent.

Context. The context for this study was a 12-week SL experience in the area of adult language impairment (Communication Help for Adults after Stroke; CHATS). The experience was coordinated with an existing community stroke survivors' group. Second-semester students in a speech-language pathology graduate program developed and facilitated weekly modules with the stroke group. These modules were designed to be fun and interactive, and typically focused on topics of functional communication for activities of daily living. Activities emphasized the use of any available functional communicative modality – including speaking, writing, drawing, and gesturing – in conversation.

Per the course syllabus provided to students (Garrity, 2013), students were required to submit journal-style reflections about the SL experience. Reflections were to be at least one, but no more

than two typed pages, double-spaced, in 12-point Times New Roman or Arial font. Students were encouraged to write their entries after each session attended rather than waiting until the end of the semester when they were due. The instructor did not provide feedback to students about the quality of their reflections during the course of the semester.

Students were instructed to follow the What? So what? Now what? Protocol (Rolfe, Freshwater, & Jasper, 2001) for completing their journals. The What? portion was to include a brief description of the events of the session. The So what? portion was to provide the student's interpretation, explanation, emotions, opinions about the events described in the What? portion. Finally, for the Now what? portion, students were instructed to "tie it all together—make the objective and subjective portions come together", and describe specific short-term insights as well as general (long-term) ones (Garrity, 2013, p. 14). Students were further instructed that their journals "must truly be reflective and/or contemplative in nature and demonstrate your personal integration of academic content with your own thoughts and experiences within the service-learning project" (Garrity, 2013, p. 14).

Rating Framework. For the purposes of coding and analysis, all journals were redacted of personal information and randomly assigned an identification number. The de-identification procedures were conducted by Rater 1, who was also the course instructor. De-identification was completed at the end of the semester in which the reflections were written. The reflections that were analyzed for the current study were written by students in 2012 – 2015, and coding began on these reflections in 2016. The intervening time period between the course, de-identification, and analysis was sufficiently long to prevent rater bias, as the course instructor/Rater 1 did not remember specifics about students' reflections by the time they were analyzed for the study.

Raters 1 and 2 conducted the first round of coding based on Plack et al.'s (2005) Level II coding, which assigns an overall or holistic three-tier rating to each reflection journal: *nonreflection, reflection,* or *critical reflection*. This framework was selected because it is one of the more parsimonious, as it includes few rating categories, and had previously been found to have relatively high rates of interrater reliability (Hill et al., 2012; Plack et al., 2005; Wong et al., 1995).

Participants. A total of 43 first-year speech-language pathology graduate students from two different cohorts contributed data for this study. Participants included one male and 42 females, with a mean age of 25.24 years (range = 22;10 - 42;11). Each participant submitted six reflective journals.

Piloting Phase. For the piloting phase, a subset of the data set was evaluated (participant $n = 19$; journal $n = 114$) by Raters 1 and 2. Initial coding attempts yielded unacceptably low and variable rates (19% - 50%) of agreement between the independent raters. In an attempt to improve agreement on future rounds of coding, Raters 1 and 2 came to consensus on a portion of the dataset (participant $n = 9$; journal $n = 54$). Through discussing these reflective journals, more details, including specific examples from the current dataset, were added to the evaluation rubric to more clearly delineate the differences among *nonreflection*, *reflection*, and *critical reflection* (see Table 1). Following the revision of the rubric, the third rater was trained on the updated rubric and completed ratings using the three-tier framework.

Data Analysis. Three raters (the first three authors of this study) conducted relevant readings, familiarized themselves with the rating framework, and discussed the rating framework collectively. The raters did not specifically measure length of time required for training. However, post hoc estimates based on personal calendar records indicated that training in the rating protocol took place over two sessions of approximately one hour each. All three raters were instructors in a speech-language pathology program, and all had experience with reading and evaluating students' written assignments.

The three raters independently rated the remainder of the journals (participant $n = 33$; reflection $n = 198$) as either *nonreflection*, *reflection*, or *critical reflection* based on the rubric criteria. The raters extracted passages from the journals to support their ratings of *reflection* or *critical reflection*. Raters were not required to extract passages from journals rated as *nonreflection* because, by definition, there was no evidence of reflection to extract. In addition, this practice was regarded as a time-saving measure, because the authors/raters were attempting to apply a reliable framework in a relatively short amount of time per journal. Raters used a standard rating form to record their reflection ratings and passages providing support for the ratings for each journal. For data analysis purposes, categorical reflection ratings were converted to numerical ratings, where 1 = *nonreflection*, 2 = *reflection*, and 3 = *critical reflection*. The numerical ratings were used for statistical analysis.

Interrater agreement was computed using Cohen's kappa, Fleiss' kappa, and overall percent agreement. The kappa statistic was selected because it allowed for comparisons of raters on ordinal data and calculates the number of agreements among raters that are beyond chance (McHugh, 2012; Sim & Wright, 2005). Cohen's kappa was used to calculate interrater agreement among rater pairs (i.e., 1 and 2, 1 and 3, 2 and 3), while Fleiss' kappa was used to calculate interrater agreement among the three raters.

Overall percent agreement was computed because it provided a straightforward measure of agreement by dividing the total number of agreed upon journals by the total number of rated journals (McHugh, 2012) and was used by Plack et al. (2005) to examine interrater agreement of reflection journal ratings. Interrater reliability was computed using ICC (ICC [2,1]). ICC measures both the degree of correlation and agreement between raters (Koo & Li, 2016; Shrout & Fleiss, 1979) and was selected because it examined the reliability of more than two raters on ordinal data. Interrater agreement (kappa statistics only) and reliability measures were calculated using R statistical package *irr v0.84* (Gamer, Lemon, Fellows, & Singh, 2012).

The criteria established by Cicchetti (1994) were used to determine the strength of the interrater agreement and reliability results. According to Cicchetti (1994), measurements of less than 0.40 are considered poor agreement, measurements between 0.40 and 0.59 are considered fair agreement, between 0.60 and 0.74 are considered good agreement, and between 0.75 and 1.00 are considered excellent agreement.

Results

Forty-three students each completed six journals for a total of 258 journals. Although 43 students completed journals, nine students' journals (n = 54) were excluded from the analysis because raters came to consensus on the journals' reflection ratings for training purposes and one student's journals (n = 6) were excluded due to file corruption. Journals from 33 students (n = 198) were analyzed for interrater agreement and reliability measures.

Interrater agreement and reliability was poor. For the pairs of raters, Cohen's kappa values ranged from (K) 0.16-0.39 with $p < 0.05$, indicating that these results were not due to chance (McHugh, 2012). Poor agreement was revealed between raters 1 and 2, K = 0.289, $p < 0.01$, 2 and 3, K = 0.391, $p < 0.01$, and 1 and 3, K = 0.166, p =0.019. Fleiss' kappa results for all three raters was (K) .15, $p < 0.05$. Overall percent agreement between the three raters was also poor at 31.8% agreement. Lastly, interrater reliability was poor among the three raters was poor, ICC = 0.29, 95% CI [0.20, 0.38], $p < 0.05$ (see Table 2).

Table 2
Percent Agreement and Interrater Reliability among Raters

	Raters 1 and 2	Raters 2 and 3	Raters 1 and 3	Raters 1, 2, and 3
Percent agreement (%)	51.01	55.56	53.54	31.82
Kappa	0.289 $p < 0.01$	0.391 $p < 0.01$	0.166 $p = 0.019$	0.154 $p < 0.01$
ICC				0.29 [CI:0.20, 0.38] $p < 0.01$

A secondary analysis was conducted to examine the raters' disagreement between the various reflection rating levels. The purpose of this secondary analysis was to determine which reflection rating levels the raters were the least consistent in rating. Disagreement was defined as journals in which two raters assigned the same rating, but a third rater assigned a different rating.

Three disagreement comparisons were made: *reflection* vs. *nonreflection*, *critical reflection* vs. *nonreflection*, and *critical reflection* vs. *reflection*. For example, a *reflection* vs. *nonreflection* disagreement occurred for participant 211. Raters 1 and 2 rated journal 1 for participant 211 as *nonreflection* while Rater 3 rated it as *reflection*. Journals in which all three raters assigned different ratings were analyzed as part of the *critical reflection* vs. *nonreflection* comparison. Percent agreement was calculated for each disagreement comparison by dividing the number of agreement journals by the total number of rated journals. The raters' percent agreement for *reflection* vs. *nonreflection* was 83.8%, for *critical reflection* vs. *nonreflection* was 95.5%, and for *critical reflection* vs. *reflection* was 52.5%.

The raters' extracted journal passages were analyzed to investigate potential reasons for the inconsistent ratings between the various reflection rating levels. The analysis consisted of the three raters' independently reviewing and re-rating the disagreement journals based solely on the extract passages. To minimize bias, the journals were de-identified of rater information (e.g., rater name). The raters then discussed the ratings and passages until consensus was reached.

Three types of disagreement journals were analyzed. These were journals in which: (1) two raters rated *nonreflection* and one rater rated *reflection*, (2) two raters rated *reflection* and one rater rated *critical reflection*, or (3) one rater rated *nonreflection*, one rater rated *reflection*, and one rater rated *critical reflection*. Raters were not required to extract passages to support their ratings of *nonreflection*; therefore, journals in which two raters rated *reflection* or *critical reflection* and one rater rated *nonreflection* were not analyzed. As examples, participant 317, journal 1 was analyzed because Raters 2 and 3 rated it as *nonreflection* while Rater 1 rated it as *reflection*. For Participant 212, journal 1 was not analyzed because Raters 1 and 2 rated the journal as *reflection* while Rater 3 rated it as *nonreflection*. A total of 13 disagreement journals and 14 corresponding passages were analyzed. Raters extracted different passages for one of the disagreement journals; therefore, both passages were analyzed. Consensus was reached on 10 of the 14 disagreement journal passage ratings and/or rating rationales. For three of the 14 disagreement journal passages, only two of the three raters reached consensus on the journal passage ratings. Two raters agreed to ratings of *critical reflection* for the three disagreement journal passages, while one rater maintained ratings of *reflection*. For one of the 14 disagreement journal passages, all three raters agreed on the rating, but only two of the raters agreed on the rationale for the rating. Table 3 provides the disagreement journal passages for which consensus on the ratings and/or rationales for the ratings was reached by two of three raters.

While the raters did not specifically record times of rating sessions, post hoc estimates indicated that the time required to read and rate reflective journals was 15 to 25 minutes per student. This time frame was based on raters who had extensive experience evaluating students' written work as well as with the assessment rubric used for the assignment. Raters who are less experienced with evaluating students' written work or who are not familiar with this specific assessment rubric used here may require more time than this.

Discussion

The purpose of this study was to examine the utility of a previously published assessment framework for evaluating the reflective journals of graduate-level speech-language pathology students in a SL experience. Several previous studies of assessment of reflective journals among students in speech-language pathology and related disciplines have yielded successful reliability across and within raters. Overall, our attempts to use a simple and efficient framework was successful, but also revealed weaknesses in the evaluation rubric. Table 4 provides a comparison of several characteristics of the current study in relation to the four studies reviewed. We included a fourth study for comparison in Table 4 (Chabon & Lee-Wilkerson, 2006), because those authors provided a measure of the time commitment required, where the others did not. That study will be discussed further in this section. Whereas previous studies used anywhere from four to a total of 12 elements of assessment to consider both the textual and abstract levels of reflection journals, in the interest of parsimony, we chose to use and study a simple three-tier framework based on the work of Mezirow (1990).

The studies reviewed here found moderate to high levels of interrater agreement using more complex rating frameworks. Others have also used the three-tier system used in the current study with acceptable levels of agreement/reliability, which was one of the reasons this specific

framework was selected for examination. Wong et al. (1995) obtained 88% agreement using the holistic three-tier framework. However, overall agreement and reliability among the three raters was poor to fair for the current study.

Several factors could account for the lower agreement and reliability among raters in this study. Some of these may be related to the students and the reflection task itself. Over the course of a semester, students experience varying levels of motivation and time, as well as energy and cognitive capacity. Although the task and reflection prompt itself seems straightforward, it might not actually have provided enough structure (e.g., specific questions) for some students to successfully and consistently document their reflections. Previous studies have noted the need to teach students how to reflect before asking them to reflect, the role of instructor feedback, and that student responses varied significantly based on the question or question type (Chabon & Lee-Wilkerson, 2006; Dyment & O'Connell, 2010).

Similar factors in the raters might have been responsible for the low agreement and reliability. The raters in this study are full-time faculty in a speech-language pathology program at a teaching-intensive university. Faculty members, too, experience fluctuations in levels of motivation, time, energy, and cognitive capacity. While the time commitment for each individual reflection journal was not extensive, the ratings reported here were assigned over a period of approximately two years, during which a number of external factors might have interfered. In addition, rater assessment personalities (i.e., easy grader versus hard grader) might have also negatively affected interrater reliability. Although it had not seemed problematic for previous raters, the authors completed a secondary analysis to attempt to identify weaknesses in the rubric that might account for the low agreement and reliability. Recall that the secondary analysis revealed high rates of agreement for *reflection* vs. *nonreflection* (83.8%) and for *critical reflection* vs. *nonreflection* (95.5%), which were consistent previous studies. Of particular concern was the essentially chance rate of percent agreement for *critical reflection* vs. *reflection* (52.5%). This analysis focused on raters' reaching consensus, using both the ratings and the evidence provided for the ratings.

Table 3

Disagreement Journal Passages for which Consensus was Reached by Two of the Three Raters: Ratings with Rating Explanations

Student ID #	Passage	Consensus rating	Explanation
220	I was surprised to hear some of the foods that the group members were eating on daily basics [sic]. I thought that this module was very informative and helpful to the group members. I am not sure that all of the stroke group members have ever had someone talk to them about the health risks they face after having a stroke, and the likelihood for another one to occur. Maintaining a healthy diet will greatly improve the chance of not suffering another stroke or other health complications.	*†Reflection*	Rubric criterion: *Writer provides a label and description for emotions evoked by the experience.* *Writer provides a general judgement or prediction about the experience.*
307	I think this was a good way to start the discussion because the other group members seemed to open up more and want to talk about their experiences. Betty, in particular, opened up more than I have seen over these past couple of weeks. She shared how emotional the past few years have been post-stroke and how it affected different areas of her life. It was a very somber time during the discussion, but I believe it was a good venting time and acknowledgment period for the group.	**Critical reflection*	Rubric criterion: *Drawing specific conclusions based on experience.* Raters' comments: The phrase "I believe it was a good venting time and acknowledgment period for the group," indicates that the student drew a specific conclusion based on his/her CHATS experience.
314	This service-learning experience has shown me how much of an impact I can have on the clients that I work with in therapy.	**Critical reflection*	Rubric criterion: *Drawing specific conclusions based on experience.* Raters' comments: The student's conclusion, that his/her therapeutic services can impact a client's life, is based on his/her CHATS experiences.
320	Even if they did not experience the same things post stroke they were connected in that they understood the hardships adjustment	**Critical reflection*	Rubric criterion: *Drawing specific conclusions based on*

DOI: https://doi.org/10.30707/TLCSD3.2Garrity

	has been. As a whole they made it very aware that people who have not had a stroke can never understand the true understanding of life post stroke. This was really amazing to watch and listen to, and made me realize just how much they truly connect with one another. Sharing this information with us also meant they were more open to Group-A as a whole compared to past visits.	*experience.* Raters' comments: The student's conclusion, that the students in Group-A have developed a therapeutic relationship with the CHATS members, is based on his/her CHATS experience.

Note. * denotes journal passage ratings for which two of the three raters reached consensus.
† denotes journal passage ratings for which all three raters reached consensus on the rating, but only two of the three raters agreed on the rationale for the rating. Raters' comments are based on the revised definitions of *reflection* and *critical reflection*.

Table 4

Comparison of Current and Previous Studies

Study	Journal *n*	Participant *n*	Journal entry length	Assessment framework	Estimated rating time per rater per journal entry	% agreement/reliability statistic
Current study	258	43	1-2 typed, doubled spaced pages, 12 point font	Three levels of reflection	2.5 - 5 minutes	31.8% agreement ICC: 0.29
Chabon & Lee Wilkerson (2006)	95	18	NR	Evidence of learning objectives plus four levels of reflection	25 minutes	84% agreement Reliability NR
Hill et al. (2012)	156	52	NR	Nine elements of textual evidence of reflection plus three levels of reflection	NR	77.78% agreement ICC Range: 0.143 to 0.5 [a]
Plack et al. (2005)	43	27	NR	Nine elements of textual evidence of reflection plus three levels of reflection	NR	67.4%-85.7% agreement ICC: 0.74 [a]
Williams et al. (2000)	848	53	NR	Six levels of reflection	NR	Agreement NR Reliability coefficient: 0.68

Note. NR = Not reported [a] Agreement/reliability is reported for three-tier reflection ratings only.

To reach consensus, raters had to further define *reflection* and *critical reflection* criteria terminology. Raters often struggled to decide if a passage exemplified the *reflection* criteria of "the writer provides a general judgment or prediction about the experience," or the *critical reflection* criteria of the writer "draws specific conclusions based on experience" because they had difficulty distinguishing between "general" and "specific" experiences. Through discussion, the *reflection* term "general" was defined as relating to a broad group of individuals that were not involved in the CHATS experience. For example, based on the agreed upon definition of "general," two of the raters concurred that the following passage exemplified a rating of *reflection*. In the following passage, the journal entry begins with a discussion of a specific CHATS experience, but the student ends the entry with a general conclusion regarding stroke prevention:

> I was surprised to hear some of the foods that the group members were eating on daily basics [sic]. I thought that this module was very informative and helpful to the group members. I am not sure that all of the stroke group members have ever had someone talk to them about the health risks they face after having a stroke, and the likelihood for another one to occur. Maintaining a healthy diet will greatly improve the chance of not suffering another stroke or other health complications.

The *critical reflection* term "specific" was defined as relating to the student's CHATS experiences. For example, based on the agreed-upon definition of "specific," two of the raters concurred that the following passage exemplified a rating of *critical reflection*, "It was a very somber time during the discussion, but I believe it was a good venting time and acknowledgment period for the group." The decisive element of this passage was the phrase "I believe it was a good venting time and acknowledgment period for the group," because it indicates that the student drew a specific conclusion based on his/her CHATS experience.

While the secondary analysis did not allow for all raters to come to consensus on all disagreements, it did illuminate a major weakness of the rubric that likely led to the lower rates of agreement and reliability. Even though several rounds of training had taken place and examples from the dataset were included in the rubric, raters still did not have operational definitions and examples that clearly illustrated each level of the three-tier framework. This was particularly problematic when trying to differentiate *reflection* from *critical reflection*. Further specification of each level and improved examples are expected to increase the rubric's reliability.

The time commitment required to complete the reading and rating of reflective journals may be a relative strength of the current method. Only one other study that we are aware of (Chabon & Lee-Wilkerson, 2006) reported the time required to complete their ratings. Those authors evaluated the reflective journals of 18 graduate students in speech-language pathology using a framework that consisted of four tiers based on the work of several previously published reflection models (Anderson & Krathwohl, 2001; Fink, 2003; Kerka, 2002; Wlodkowski, 1999): *descriptive*, *empathic*, *analytic*, and *metacognitive* (see Chabon & Lee-Wilkerson, 2006 for detailed descriptions and examples of each tier). They reported that each entry took 25 minutes to rate, translating to a total of approximately 40 hours for all entries in the dataset, which also accounted for follow up agreement discussions between the raters. The three-tier framework utilized in the current study required approximately two hours or training followed by a time commitment of 15-25 minutes per entry.

Another factor that could influence the time required to assess reflective journals is their length. While the reflective journals in the dataset examined here were relatively short (1 - 2 pages in length, 12-point font, and double-spaced), the other studies reviewed do not report the length of the entries. In addition, while the reflective journals in the current study were assessed just for depth of reflection, Chabon and Lee-Wilkerson (2006) were evaluating their reflective journals for depth of reflection as well as evidence of learning of course objectives. Considering the dual purpose of their evaluation, the time commitment of their framework and the one used in the current study appears to be comparable, a finding that speaks to the complexities of finding a singular "best" reflective journal assessment framework mentioned earlier.

Despite its weaknesses, the strengths of this rubric were the parsimony of the three-tier system, the relative efficiency with which ratings could be assigned, and the substantial agreement for differentiating entries rated *reflection* and c*ritical reflection* from those rated as *nonreflection*. Instructors seeking to evaluate reflective journals have several frameworks from which to choose, keeping in mind that they need to consider several factors before selecting one. These factors include, but are not necessarily limited to, the purpose(s) of the assessment, assignment parameters, and complexity of the framework. In addition, since evidence suggests that details such as student characteristics and reflection questions/prompts might also play a role in the quality of reflective journals, further inquiry into these aspects will help instructors to better engage our students in high level reflection for deep engagement and meaningful learning experiences.

On a final note, considering that the assessment of reflective journals is potentially affected by a number of extraneous factors, the authors have attempted to modify our practice in this area. Although not yet fully formed, we recognize the value of our rubric as a foundation for the assessment of reflective journals. While keeping in mind the lessons learned about the limitations of this method, as well as its strengths, we will continue to study and refine it, applying specific modifications based on evidence from the literature and from our own students' reflective journals.

The findings of this investigation have led to changes in practices related to the assessment of reflective journals within the context of this SL experience. We are attempting to control for some of the extraneous variables we have identified by providing students with the reflection rubric at the beginning of the course, so they are aware of the ratings and criteria, and by sharing their reflection ratings after each reflective journal is submitted. In addition, we now ask students to provide, as part of every reflective journal, a rating (1 = very low, 5 =very high) regarding their state of mind in the following areas: level of interest in the week's topic; overall level of motivation during the week; level of preparation for the week's session; and level of focus when writing their weekly reflection.

As we continue to develop this method of reflective journal assessment, we also plan to address the limitations that were revealed within the rubric itself. We need to explore and craft improved operational definitions that will allow users of this rubric to better distinguish *reflection* from *critical reflection*. In addition, just as we are currently collecting information from students regarding their interest, motivation, preparation, and focus, we must examine the role that similar variables among raters might play in their assessment of reflective journals, as well as personal characteristics such as being an "easy grader" as opposed to being a "hard grader". As our goal is

to create a reflective journal rating framework that is valid and reliable across contexts, we need to determine the factors that influence students and raters in this process in order to realize a more representative view of students' reflection skills.

Author Disclosures

April Garrity: I am a full-time faculty member at Georgia Southern University, for which I receive a salary. I am the founder and director of the Communication Help for Adults (CHATS) Project discussed in this article, however I do not receive additional compensation for that work as it is part of my assigned workload at Georgia Southern University. I have no other relevant financial or nonfinancial relationships to disclose.

Casey Keck: I have no relevant financial or nonfinancial relationships to disclose.

Janet Bradshaw: I have no relevant financial or nonfinancial relationships to disclose.

Keiko Ishikawa: I have no relevant financial or nonfinancial relationships to disclose.

References

Anderson, L., & Krathwohl, D. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York, NY: Longman.

Ash, S., & Clayton, P. (2004). The articulated learning: An approach to guided reflection and assessment. *Innovative Higher Education, 29*, 137-154.

Bloom, B., & Krathwohl, D. (1956). *Taxonomy of educational objectives: The classification of educational goals, by a committee of college and university examiners. Handbook I: Cognitive domain.* New York, NY: Longmans, Green.

Boud, D. (2001). Using journal writing to enhance reflective practice. *New Directions Adult Continuing Education*, *90*, 9-18.

Boud, D., Keogh, R., & Walker, D. (1985). Promoting reflection in learning. In D. Boud, R. Keogh, & D. Walker (Eds.), *Reflection: Turning experience into learning* (pp. 18-40). London, England: Koran Page.

Bourner, T. (2003). Assessing reflective learning. *Education Training, 45*, 267-272.

Chabon, S. S., & Lee-Wilkerson, D. (2006). Use of journal writing in the assessment of CSD students' learning about diversity: A method worthy of reflection. *Communication Disorders Quarterly*, *27*(3), 146–158.

Cicchetti, V. D. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284-290.

Choy, S. C., & Oo, P. S. (2012). Thinking and teaching practices: A precursor for incorporating critical thinking into the classroom? *International Journal of Instruction, 5*(1), 167-182.

Dewey, J. (1916). *Democracy and education: An introduction to the philosophy of education*. New York, NY: Macmillan.

Dewey, J. (1938) *Experience and education.* New York, NY: Macmillan.

Dyment, J.E. & O'Connell, T.S. (2010) The quality of reflection in student journals: A review of limiting and enabling factors. *Innovative Higher Education, 35(4)*, 233-244.

Fink, L. (2003). *Creating significant learning experiences*. San Francisco, CA: Jossey-Bass.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2012). irr: Various coefficients of interrater reliability and agreement (R package version 0.84)[Computer software].

Garrity, A. (2013). *CSDS 7151: Aphasia and related neurogenic disorders* [Syllabus]. Savannah, GA: Communication sciences and disorders program, Armstrong Atlantic State University.

Hatton, N., & Smith, D. (1995). Reflection in teacher education: Towards definition and implementation. *Teaching and Teacher Education, 11*, 33-49.

Hill, A., Davidson, B., & Theodoros, D. (2012). Reflections on clinical learning in novice speech-language therapy students. *International Journal of Language and Communication Disorders, 47*, 413-426.

Jarvis, P. (1992). Reflective practice and nursing. *Nurse Education Today, 12*, 174-181.

Jarvis, P. (2001). Journal writing in higher education. *New Directions for Adult and Continuing Education, 90*, 79-86.

Kember, D., McKay, J., Sinclair, K., & Wong, F. (2008) A four-category scheme for coding and assessing the level of reflection in written work. *Assessment & Evaluation in Higher Education, 33*, 369-379.

Kerka, S. (2002). *Journal writing as an adult learning tool.* (Practice Application Brief No. 22). ERIC Clearinghouse on Adult, Career, and Vocational Education. Office of Educational Research and Improvement. ERIC Number ED470782.

Kolb, D. (1984). *Experiential learning: Experience as the source of learning and development.* Englewood Cliffs, NJ: Prentice Hall.

Koo, T. K., & Li, M. Y. (2016). A guideline for 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

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

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282.

Mezirow, J. (1990). *Fostering critical reflection in adulthood: A guide to transformative and emancipatory learning.* San Francisco, CA: Jossey-Bass Inc.

Plack, M., & Greenberg, L. (2005). The reflective practitioner: Reaching for excellence in clinical practice. *Pediatrics, 116*, 1546-1552.

Plack, M., Discoll, M., Blisset, S., Mckenna, R., & Plack, T. P. (2005). A method for assessing reflective journal writing. *Journal of Allied Health*, *Winter 34* (4), 199-208.

Rolfe, G., Freshwater, D., & Jasper, M. (2001). *Critical reflection in nursing and the helping professions: A user's guide*. London, England: Palgrave Macmillan, Ltd.

Scanlan, J., & Chernomas, W. (1997). Developing the reflective teacher. *Journal of Advanced Nursing, 25*, 1138-1143.

Schön, D. (1983). *The reflective practitioner*: *How professionals think in action*. New York, NY: Basic Books.

Schön, D. (1987). *Educating the reflective practitioner: Toward a new design for learning in the professions*. San Francisco, CA: Jossey-Bass.

Sezer, R. (2008). Integration of critical thinking skills into elementary school teacher

education courses in mathematics. *Education*, *128*(3), 349-362.
Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.
Sim, J., & Wright, C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, *85*(3), 257-268.
Stacey, K., Rice, L. R., & Langer G. (2001). *Academic service-learning. Faculty development manual.* USA Eastern Michigan University: Office of Academic Service-Learning.
Valli, L. (1997). Listening to other voices: A description of teacher reflection in the United States. *Peabody Journal of Education, 72*, 67-88.
Williams, R., Sundelin, G., Foster-Seargeant, E., & Norman, G. (2000). Assessing the reliability of grading reflective journal writing. *Journal of Physical Therapy Education, 14*, 23-26.
Wlodkowski, R. (1999). *Enhancing adult motivation to learn: A comprehensive guide for teaching all adults.* San Francisco, CA: Jossey-Bass.
Woodward, H. (1998). Reflective journals and portfolios: Learning through assessment. *Assessment and Evaluation in Higher Education*, *23*, 415-423.
Wong, F., Kember, D., Chung, L., & Yan, L. (1995). Assessing the level of student reflection from reflective journals. *Journal of Advanced Nursing, 22*, 48.