Setting Standards for a Portfolio Component Using Qualitative Methods

BY

KEVIN E. VAN KANEGAN
B.S., University of Illinois, Urbana, 1989
B.S., University of Illinois, Chicago, 1991
D.D.S., University of Illinois, Chicago, 1993

THESIS

Submitted as partial fulfillment of the requirements
for the degree of Master of Health Professions Education
in the Graduate College of the
University of Illinois at Chicago, 2014

Chicago, Illinois

Defense Committee:
Ilene Harris PhD, Chair and Advisor
Steven Downing PhD
Yoon Soo Park PhD

This thesis is dedicated to my wife, Mona, and our children, Neil and Rahel, whose forbearance and patient urging have encouraged me to complete this journey despite the lengthy path it became.

ACKNOWLEDGMENTS

To my Thesis Committee Members:

Ilene Harris, who provided consistent and unwavering encouragement and enthusiasm coupled with precise ideas and advice.

Steven Downing, who gave realistic assessment of the assessment methods used and results obtained all the while infusing the conversation with humor.

Yoon Soo Park, who willingly came into the picture at a late date and lent critical statistical support and asked insightful questions,

William Knight, who put me on this journey and provided the space and environment for it to occur,

The University of Illinois at Chicago, College of Dentistry Faculty, for their willingness and patience to participate in the study that forms the basis of this endeavor,

The administration and faculty at Midwestern University-College of Dental Medicine-Illinois, who continued the support I needed for the conclusion of this project.

KVK

TABLE OF CONTENTS

LIST OF TABLES

SUMMARY

Over the past two decades, there has been increasing interest and use of portfolios as an assessment method in health professions education. However, the use of portfolios to derive summative assessments has been thwarted by difficulties in obtaining high enough inter-rater reliability to support the decisions to be made. The lack of inter-rater reliability can be attributed to the often variable nature of what is included in the portfolios and the fact that at least some of these items are written essays and other constructed response type items. Some researchers have improved inter-rater reliability through application of prescriptive portfolio construction and assessment guidelines. The problem with these efforts is the concomitant removal of the self-selective aspects of portfolio construction which are considered essential for the evaluation of professionalism for which portfolios are most commonly used. Others have proposed the use of qualitative methods in portfolio assessment as being better suited to this purpose, of evaluating professionalism.

This study was designed to compare inter-rater reliability for grading of patient care based reflective essays, a common portfolio component, using a conventional scoring rubric and a rubric with standards set and anchors derived from a focus group discussion. Two groups of 10 essays from a regular offering of a fourth year comprehensive care course at the University of Illinois at Chicago, College of Dentistry were graded by a group of 4 full time clinical faculty members. The first group of 10 essays was graded using the scoring guidelines provided in the course syllabus and

used in the regular offering of the course. The second set of 10 similar essays was graded by the same 4 faculty using a scoring rubric that was the result of qualitative analysis of themes obtained from a focus group discussion among a similar but different group of 5 faculty members. These 5 faculty members, prior to the focus group discussion, were provided 2 borderline essays from the same course to read and grade. The focus group discussion was facilitated to describe the qualities in the essays that led the faculty to assign the grade they chose. After each set of essays was graded, exact agreement and intra-class correlations were calculated. Finally, a second focus group discussion was conducted with the faculty assessors to provide insight into their experiences with the two rubrics.

The results of this study did not demonstrate the improvement in inter-rater reliability needed to support summative assessment decisions. However, these results were promising in that this technique, applied using a larger sample of essays, might show even greater improvements in inter-rater reliability. Probably most encouraging was the faculty assessor perceptions of their experiences in grading, which clearly demonstrated a greater degree of confidence and defensibility in the grades given, using the focus group derived scoring rubric and standards. The results of this study indicate that the use of qualitative methods applied to the assessment of portfolios warrants further investigation.

I. Introduction

A. Background

Over the past two decades portfolios have increasingly been used within the health professions to make judgments regarding the competency of individuals to practice their chosen professions (McCready, 2006), obtain initial licensure (Chambers, 2004) and achieve re-certification (Wilkinson, et al., 2002). Professionalism is a domain for which portfolios are considered as particularly suitable for assessment. The self-selected and reflective aspects of portfolio construction are viewed as crucial demonstrations of learners' capacity to create accurate and plausible action plans to achieve learning goals in areas of self-identified deficiencies (Goldie, et al., 2007). The fact that the portfolio is constructed over a period of time is thought to be reflective of an authentic learning process (Friedman Ben David, et al., 2001). These qualities make portfolios potentially fertile ground from which to reap summative assessments, even though their use for this purpose also presents potential dangers, somewhat akin to farming on the slope of an active volcano.

B. Statement of the Problem

Unfortunately, although there is significant evidence to recommend portfolios for use in formative assessment, the evidence is insufficient to support their use for summative assessments (Roberts, et al., 2002). Most agree that portfolios are primarily useful in assessment of learners' abilities to accurately engage in reflection and self-assessment; however, there is little consensus regarding what comprises a "portfolio"

(Rees, 2005). Moreover, the majority of studies do not report the typical, or required, components of the portfolios they expect their learners to construct (Buckley, et al., 2009; Driessen, et al., 2007). One element, however, that seems to be common among many portfolios is a patient experience-based reflective essay. The thorniest assessment challenges, among others, are mostly related to assessment of what is frequently a collection of these narratives. The grading of portfolios inherently involves a degree of subjective interpretation. Given the often substantial amount of resources required to create and use portfolios in assessment, many have raised questions regarding whether it is worth the effort (McCready, 2006; Pinsky, et al., 2004). Do portfolios contribute so strongly or uniquely to the assessment system to justify the effort or are they just another passing season in the pedagogical almanac?

As reliability is considered an essential precursor for inferences of validity, many have attempted to increase the reliability of portfolio interpretation by increasing the structure of guidelines for their construction and assessment (Pitts, et al., 2002; Rees, et al., 2004; Karlowicz, 2010). These efforts have met with moderate success, but the related decrease of self-selection compromises the validity of assessment interpretations (Driessen, et al., 2003). These attempts to increase the reliability of assessment, by providing more specific guidelines for portfolio construction, also perhaps require an inordinate amount of resources to achieve only moderately improved reliability levels. The primary finding of many of these studies is the improvement to inter-rater reliability that may occur when there is discussion among multiple independent assessors (Pitts, et al., 2002; Rees, et al., 2004).

It is into this situation that others have more recently recommended the use of qualitative methodology for portfolio assessment, in place of the more traditional quantitative approaches (Driessen, et al., 2005; Webb, et al., 2003). Because a portfolio is, by its very nature, a qualitative work, these educators/authors suggest that qualitative methods are a more natural fit than quantitative methods for their assessment and that forcing a quantitative assessment framework yields perhaps, not unsurprisingly, dismal results. We don't experience learning as a stable quantifiable characteristic, so why expect it to be so, and for it to always be amenable to assessment using quantitative and statistical analysis? Driessen, et al. at Maastricht University, using the foundational work of Lincoln and Guba (1985), recommend and use in practice, a portfolio framework with dependability as a parallel to reliability and credibility as a parallel to validity (Driessen, et al., 2005).

Credibility, similar to validity, according to Lincoln and Guba, is defined as the "truth value" of both the study process itself and the interpretation of the study findings (Lincoln, et al., 296). The techniques recommended for establishing credibility include persistent observation, prolonged engagement of the evaluator, triangulation, negative case analysis, peer debriefing, member checking and referential adequacy (Lincoln, et al., 301). Dependability, like reliability, focuses on the reproducibility of both the study and results, but also relates to accounting for the often inherent instability of what is being observed in the qualitative setting (Lincoln, et al., 299). The recommended technique for establishing dependability is an inquiry audit (Lincoln, et al., 317-318). The relationship between credibility and dependability is also considered to be the same

as that between validity and reliability; there is no credibility without dependability as there is no validity without reliability (Lincoln, et al., 316). The specific techniques for establishing credibility and dependability in this study will be discussed further as they apply to the method of inquiry.

Driessen and his colleagues propose the use of triangulation, prolonged engagement, member checking, audit trail and dependability audit in the evaluation of their portfolios (Driessen, et al., 2005). A reasonable evaluation and interpretation of these results is that the effort may be too extensive, that the evidence is not yet strong enough for use of portfolios for summative assessments and that the solution to the appropriate balance between costs and benefits, as is often the case, probably lies somewhere in between. The quest, therefore, is to create portfolio assessment approaches with enough structure to provide guidance, but not so much that there is no freedom for students to choose what to include, while at the same time expending resources efficiently. The authors of this paper suggest seeking credibility and dependability in the construction of grading guidelines, which may lead to validity and reliability in the interpretation of the results, when effectively communicated to learners and assessors.

When attempting to create a grading construct for portfolios, it is important to keep in mind that there are several variables to consider. Norcini writes that, “specifically standards need to be set by the right number and kind of standard-setters, the method

to set them must meet certain criteria, and the outcomes of applying the standard should be reasonable" (Norcini, et al., 2002). As previously discussed, qualitative methods hold great promise for enhancing the use of portfolios in summative assessment. Qualitative methods allow for a collection of rich textual data. These methods are particularly applicable in situations in which the object of study is embedded in a complex social environment and, through rigorous application, allow for credible interpretations to be drawn (Harris, 2002). The decision was made for this study to set grading standards for a portfolio component, using a qualitative approach, because it is in keeping with the character of portfolios, namely the construction of portfolios using rich narrative data. Specifically, a focus group was used to derive themes that were used as anchors for determining grades. The goal was not to create a detailed checklist of criteria, but rather to characterize what a passing, and then a high passing, portfolio "looks" like. What does one, as a faculty reviewer, see when assessing a portfolio that gives evidence for her or his decision at each of these levels? This is the essential question the answer to which was used to anchor the grading levels, i.e., to provide descriptive anchors for use in grading. The desired outcome is significant improvement in inter-rater reliability. Was the resulting grading still subjective? Certainly, but subjective with mutually agreed upon understandings of construct relevant variables, is a continent away from subjective assessment without mutually agreed upon standards. And with further testing, the use of focus group methodology to set standards may result in portfolios having the credibility and dependability worthy of the summative interpretations we are attempting to make.

In summary, currently portfolio assessment is still not reliable or only reliable with restrictive guidelines and structure imposed, that at the same time diminish the value of portfolios for assessment of student self-directed learning. Setting standards for assessment, using a focus group/grounded theory approach, may be a means to obtain reliable holistic assessment without undue structure. The inter-rater reliability for this study was compared to levels published for similar assessments. Furthermore, the faculty assessors' grading experiences with two different grading standards were elicited, recorded, and examined, to compare their experiences with using each of the standards.

C. **Purpose of the Study**

The purpose of this study was to set grading standards and anchors for a patient experience-based reflective essay, a common component of portfolios, using focus group methodology, and to evaluate the effectiveness of this approach for raising inter-rater reliability to an acceptable level to permit valid summative assessments.

The reason for choosing a focus group for this purpose was a desire to engage participants in a dialogue which could be recorded and analyzed, to develop new portfolio grading standards and anchors based on the themes formulated. Given that there are few currently provided anchors for portfolio grading at the University of Illinois at Chicago, College of Dentistry (UIC-COD), a brainstorming type environment, characteristic of focus groups, was selected as the qualitative method used to elicit standards. This semi-structured milieu for faculty discussion also seemed appropriate

for eliciting standards for assessment of the type of portfolios we have had our learners construct.

II. Methods

A. The Study Setting

This study took place at the University of Illinois at Chicago, College of Dentistry. The comprehensive care course from which the sample essays were obtained was offered in the Fall semester of students' fourth year and taken by 68 students. The essays were graded using a rubric of High Pass, Pass and No Pass. This grading structure was chosen, as it seemed consistent with the degree of confidence we have in portfolio assessment at this time and is consistent with what others report in the literature (Grant, et al., 2007; Usherwood, et al., 1992). One of the authors (KVK) was the director for the fourth year comprehensive care courses and the only grader of the essays for the course.

B. The Study Design

Nine experienced full time undergraduate clinical faculty members were asked to participate in the study. All of these faculty provided direct clinical instruction to the students enrolled in the comprehensive care course. Four of the faculty members were designated to be assessors of essays and five participated in a standard setting focus group discussion. The Executive Associate Dean for Academic Affairs was asked to participate as auditor of the research methods and the results.

The four faculty assessors were given information regarding the type of learners being assessed (fourth year dental students), the objectives of the assignment (course

syllabus), and two sealed packets each containing 10 different anonymous sample essays. No other formal training was provided to these faculty, which was consistent with what was being done in the comprehensive care courses at the time. The essays were chosen to be a mix of essays of varying quality, based on grades the course director gave each essay at the actual time the course was administered. The faculty were not informed of the grades that were assigned these essays and the essays were purged of any information identifying the student authors. Each assessor was asked to create her/his own two digit code to be placed on each essay. The authors did not know or gather these codes, and upon collection of the graded essay packets, there was no way to trace a graded essay back to an individual faculty member assessor. The faculty assessors were then asked to read and grade independently each of the 10 essays in the first packet. Grades were assigned, using the categories High Pass, Pass, and No Pass. None of the faculty assessors had seen these essays prior to this invitation to participate in the study.

The general expectations for the essay included reflecting on a personal patient care experience in which the treatment recommended involved choices among at least two options. The focus of learner effort was expected to be on searching and appraising literature related to the prognoses for these treatment options for this patient and also recording the patient's reasons for choosing one of the treatment options. Students were asked to reflect about this education experience and propose how she or he might make changes in the presentation of treatment option information to a similar patient in the future. Students were instructed not to include any patient identifying information in

the construction of the essay and this was verified during course grading and again prior to inclusion in the packets.

A separate group of five faculty were given two sample essays, from the same administration of the course, to read one week prior to participation in a one hour focus group discussion. One of the essays was one for which the course director struggled to determine whether it warranted a grade of No Pass or Pass and the other essay was one the course director struggled in deciding whether to grade a Pass or High Pass. Both of the samples were chosen because of their potential to generate vigorous discussion focused on what criteria should be associated with particular grade categories. The student authors had addressed all the elements described in the syllabus, but had done so with varying levels of quality. Again, these were anonymous sample essays collected during the same regular administration of the course.

One of the authors (KVK) served as facilitator for the focus group and the session was audio recorded for later transcription. At the opening of the session, ground rules were established, indicating the role of the facilitator, the length of the session, and the recording and maintenance of anonymity of responses. The session was audio recorded, only, and the transcription was produced to distinguish between facilitator comments and non-facilitator comments. In the transcript, the individual participants were identified as Participant 1, Participant 2, etc., so as to trace distinct perspectives. However, these identifications were not linked to any participant's identity. The facilitator did not participate in the discussion other than to ask open-ended questions to

invite responses and to request clarification. The entire session was then transcribed and analyzed by the facilitator (KVK) to identify themes. The themes and supporting documents were then submitted to the Executive Associate Dean for Academic Affairs for an audit of process and product. The Executive Associate Dean for Academic Affairs was chosen to do this, as he is familiar with the learners and curriculum, and can provide a contextual perspective which may provide evidence of content validity.

The themes were then, upon completion of review by the auditor, submitted to the focus group participants in the form of an email survey, with the request to rate each theme as Strongly Agree (SA), Agree (A), Disagree (D), or Strongly Disagree (SD). Additional comments, to emphasize degree of agreement or disagreement and to add or delete themes, were also requested. The survey method was chosen as a means of member checking to allow the participants to evaluate the themes in an independent and reflective mode. No neutral choice was provided, to avoid equivocation. This process of surveying and re-surveying continued until no new themes or suggestions for modification were identified. By this means of member checking, an iterative process continued until saturation was achieved and the final list of themes to serve as grade anchors was developed.

These themes were then submitted back to the four faculty assessors to serve as grade category anchors for application in the assessment of the 10 essays in the second packet. Both packets were then sealed and returned to the facilitator. All data

was collected and stored without identifiers. The data for the two essays sets was analyzed and compared, using exact agreement and intra-class correlation statistics. Exact agreement is a measure of the proportion of exact agreement between raters. Intra-class correlation takes into account chance agreement and also penalizes for larger score differences between raters. SPSS software was used to perform the statistical analyses.

(http://support.spss.com/productsext/spss/documentation/statistics/macros/Mkappasc.htm).

The facilitator then conducted a second one hour focus group session with the faculty assessors using the same focus group methods. The purpose was to elicit descriptions and perspectives from the assessors about their experience in using the two grading rubrics to assess the essays. They were asked questions, such as: Do you have a preference for one process versus the other? Why? Probes: Efficiency? Greater confidence in the grade given? Describe the experience of grading the portfolios with the first set and the second set. This data was sought to provide a qualitative comparison of the two grading experiences and add a detailed analysis of the process itself. The themes the author derived from analysis of the second focus group transcript were also submitted for audit by the Executive Associate Dean of Academic Affairs and then, via email survey, member checked until saturation was achieved.

C. **Study Design Summary**

1. Four (4) faculty assessors graded the first sample of 10 essays and inter-rater reliability was calculated.

2. a. Five (5) focus group faculty were given 2 borderline sample essays to read.
   b. An audio recorded standard setting focus group was conducted.
   c. The focus group session was transcribed and analyzed to identify themes.
   d. The transcript and themes were submitted to the Executive Associate Dean for Academic Affairs for audit.
   e. The themes were submitted back to focus group participants, via e-mail survey, for member checking. This process was repeated until saturation was reached.
   f. The themes were used as anchors for development of a grading rubric.

3. The same original four faculty assessors were provided the new rubric with anchors and asked to grade a second set of 10 sample essays. Inter-rater reliability was calculated again.

4. Pre and post inter-rater reliability was compared using exact agreement, and intra-class correlation statistics, and subsequently compared with data found in the literature for similar assessments.

5. a. The four faculty assessors, without knowledge of changes to inter-rater reliability, participated in an audio recorded focus group session .
   b. The focus group session was transcribed and analyzed to identify themes.
   c. The transcript and themes were submitted to the Executive Associate Dean for Academic Affairs for audit.
   d. The themes were submitted back to faculty assessors, via e-mail survey, for member checking. This process was repeated until saturation was reached.
   e. Themes were identified to describe the experiences and perspectives of faculty evaluators in using the two rubrics.

This methodology is in keeping with recommended standard setting practices for constructed response items, such as use of multiple raters and model essays (Downing, et al., 2009) and is consistent with recommendations calling for discussion among multiple raters found in the portfolio assessment literature (Rees, et al., 2004). Norcini and Shea have presented clear guidelines for construction of credible standards which are widely accepted based on the evidence supporting their defensibility (Norcini and Shea, 1997). See Table I for how each of these guidelines was addressed in this study.

TABLE I

STANDARD SETTING METHODOLOGY

Evidence for Credibility of a Standard Setting Process (Norcini, et al., 1997).	**How this study provided this evidence.**
Standard-setters should have the right qualifications.	The standard-setters are licensed practitioners and clinical educators. They are experts without conflicts of interest. Both male and female standard setters were chosen.
Many standard-setters should be involved (5-10).	Five standard-setters were involved.
The judgments of other credible groups should be included.	Borderline essays were used to generate focus group discussion. Audit by Executive Associate Dean of Academic Affairs provided contextual perspective.
The standard should be absolute.	The purpose of this study was to develop anchors for absolute holistic rating of a patient experience-based reflective essay.
The standard should be based on informed judgment.	The standard-setters and auditor are familiar with the learners and the assignment expectations. Judgments were made based on discussion about borderline sample essays.
The process should demonstrate due diligence.	Once themes were derived, member checking continued via survey of standard-setters until saturation was achieved.
The method should be supported by research.	Published guidelines for standard setting practice and focus group research were applied.

Prior to conducting this study, the research protocol was submitted to the University of Illinois at Chicago Institutional Review Board and was granted an exemption.

III. Results

A. Standard Setting Results

The results of the standard setting focus group yielded the following themes which were then applied as anchors for the chosen grade categories. Six themes were developed to describe an essay that would receive a grade of PASS and six additional themes were developed and assigned to characterize an essay that would receive a grade of HIGH PASS. It was decided that an essay would receive a grade of NO PASS if it failed to demonstrate all the qualities of a PASSing essay. It took two total rounds of surveying (initial survey and then re-survey after modification) to reach saturation. The auditor found no issues with regards to process or adherence to protocol. In addition to the anchors for the grade categories, six additional themes were formulated which reflected general faculty concerns and these themes also are described below. These additional themes were also subjected to audit and then member checking, until saturation, at the same time as the grade category themes.

An essay receiving a grade of **NO PASS** does not demonstrate ALL of the qualities of an essay receiving a grade of PASS

An essay receiving a grade of **PASS** demonstrates ALL of the following qualities:

1. It follows all directions for format and content of the assignment, as presented in the course syllabus.

2. It is well organized and clear.
3. It presents enough patient specific information (patient history, description of clinical and radiographic findings) to support the proposed treatment plan options.
4. The treatment plan options presented are reasonable and present all viable options.
5. The evidence presented is of adequate quality upon which to base prognoses.
6. Communication with the patient regarding treatment options, prognoses and his/her values is adequately described to illustrate the basis for the patient's decision.

In summary, a PASSing essay presents the facts of the case, the evidence considered, and the communications with the patient.

An essay receiving a grade of **HIGH PASS** demonstrates all of the qualities of a PASSing essay and at least some (two or more) of the following:

1. The essay provides a high level of detail regarding the patient's specific needs and histories. The student seems to really know the patient.
2. The etiology and risk factors of the patient's presenting condition are clearly identified and thoroughly considered throughout the process described.

3. The evidence presented is of high quality and indicates a broad and deep search of the available evidence with strong consideration of patient specific histories and concerns (values).
4. The student brings his or her limited clinical expertise to bear upon the interpretation of the evidence related to this patient's specific needs. He or she recognizes the limits of the evidence and its application, and the value of context specific judgment.
5. The treatment options considered include a detailed and patient specific plan for follow up and maintenance.
6. The student describes how they assisted, or would assist, the patient with a treatment decision when the evidence is absent or of low level/poor quality.

In summary, the essay graded HIGH PASS goes beyond presenting the facts of the case to "bring the patient to life" by presenting detailed patient specific information and carefully considered evidence in light of the values elicited from this patient and the students' previous experiences.

Below are described additional themes that were identified which were not related to specific grade categories, but can be classified as general concerns that the standard setting faculty had when assessing an assignment such as this one.

1. The faculty participating want to increase the structure of the assignment instructions and grading guidelines by including requests in the syllabus for very specific information and using a “checklist” style grading rubric.
2. Faculty expressed a desire to use a relative scale to grade the essays, i.e., read all or several essays first, and then assign grades based upon a comparison of those read.
3. Faculty recognized both the advantages and disadvantages of implementing the above two processes, i.e., increasing structured guidelines and using a relative scale. In particular, they recognized that doing so would tend to diminish the potential to assess students’ reflection and critical thinking.
4. Faculty expressed strong emphasis on the importance of good faculty role modeling. In particular, there was great concern for the damage that occurs when faculty expect certain behaviors of students, but then themselves exhibit contrary behaviors. They thought this situation may lower faculty expectations for student performance.
5. Faculty questioned the use of a three category (High Pass, Pass, No Pass) grading scale. They thought that this was not all that much different from assigning conventional letter grades of A,B, C, and F. They thought that the number of grade categories chosen is important and closely related to the confidence a grader has in her/his ability to discriminate among different levels of performance.

6. Faculty emphasized that they should be careful to not have unrealistic expectations of learners. This was felt to be a particularly important precaution when asking learners to discuss a difficult issue, such as prognosis.

B. **Essay Rating Results**

The grades assigned by the faculty assessors for the two essay sets are presented in Table II.

TABLE II

FACULTY ASSESSOR GRADES

Essay Set 1

Assessor/Essay	**1-A**	**1-B**	**1-C**	**1-D**	**1-E**	**1-F**	**1-G**	**1-H**	**1-I**	**1-J**
1	NP	HP	HP	P	P	P	P	HP	P	NP
2	HP	P	HP	NP	HP	HP	NP	NP	NP	HP
3	P	P	HP	P	P	P	HP	P	P	HP
4	NP	NP	NP	NP	NP	NP	NP	NP	NP	NP

Essay Set 2

Assessor/Essay	**2-A**	**2-B**	**2-C**	**2-D**	**2-E**	**2-F**	**2-G**	**2-H**	**2-I**	**2-J**
1	NP	P	P	P	NP	P	NP	NP	NP	NP
2	P	HP	HP	HP	P	P	P	P	HP	P
3	P	P	P	P	P	P	P	P	P	P
4	P	P	P	P	P	HP	HP	P	P	P

HP=High Pass, P=Pass, NP=No Pass

Initial examination of the data from the comparison of the faculty ratings for the two essay sets yielded an exact agreement of 25.00% for Essay Set #1 and 43.33% for Essay Set #2. Intra-class correlation coefficients (ICC) were also calculated for all faculty assessors, combined, and yielded -0.042 (95% CI: -0.130, 0.209) for Essay Set 1, and 0.043 (95% CI: -0.061, 0.321) for Essay Set 2. Rater effect was found to be significant (p 0.0001). The intra-class correlation coefficients were calculated after coding the scoring categories both 0,1,2 and 1,2,3 and the results did not change. Due to the small sample size of essays, the intra-class correlation estimates cannot be considered stable (high standard error). This is indicated by the negative ICC value for Essay Set 1, and 95% confidence intervals (CI) that overlap across 0.

C. **Faculty Assessor Experience Results**

The results of the faculty assessor experience focus group yielded the following five themes. Quotes are included to demonstrate the themes. It took two total rounds of surveying (initial survey and then re-survey after modification) to reach saturation. The auditor found no issues with regards to process or adherence to protocol.

1. The Faculty Assessors felt that use of the second scoring rubric was more difficult to use than the first.

    -“I found it a little more arduous process the second time.”

    -“I had trouble internalizing the directions.”

    -“Yeah, I did find it challenging too, honestly.”

2. <u>The Faculty Assessors felt more confident in the grade they gave using the second scoring rubric.</u>

    -“I think I was better at grading the second part.”

    -“I think the second set, I was more confident.”

    -“I feel more confident about my second packet…..grading them.”

    -“Maybe just going back to the first one (referring to assessing the first set of essays), I would say I had very little confidence in my grade.”

3. <u>The Faculty Assessors felt the grade they gave using the second scoring rubric was more defensible.</u>

    -“You know the good thing about having this rubric (referring to the second scoring rubric), even though it seems like the kind of thing you might have to go through this course maybe a few years and then you’d finally have it internalized, but you have it and then if there were a question about the appropriateness of the grade, you would have something to fall back on.”

    -“I don’t know, I think like having the rubric (referring to the second scoring rubric) kind of gave a little bit more guidance to me. A little more concrete in terms of my evaluations, and so forth. So, I think that reflected in that.”

4. <u>The Faculty Assessors felt that using the second scoring rubric was more efficient, or had the potential to be more efficient, than the first.</u>

    -“The second for me was more efficient.”

-“Yes, I found both of them difficult, but it (the second scoring rubric) was probably more efficient.”

5. The Faculty Assessors expressed a need for calibration and practice to be able to evaluate assignments such as this, i.e., constructed responses/portfolios.

-“To identify all these things in that document was quite difficult and I am not sure I was even able to successfully do it.”

-“And like you said, if you are in a course time after time, the more that happens the more you can internalize it and kind of get used to utilizing it most efficiently, or most fairly.”

IV. Discussion and Analysis

The literature contains several studies in which improved inter-rater reliability in scoring portfolios is the goal. Unfortunately, these studies rarely indicate the methods used for developing the grading standards that were applied. Those that do describe these methods most commonly cite the use of discussion among experts and describe some difficulties with use of the scoring rubrics developed, related to differing opinions among raters regarding the importance of individual criteria (O'Sullivan, et al., 2004; Rees, et al., 2004). Other groups have used or adapted previously developed criteria from sources outside their institution and have applied these criteria, with varying results (Melville, et al., 2004; Tate, et al., 1999). Authors of a more recent systematic review of portfolios in medical education found that for the six studies that met the inclusion criteria, the average inter-rater reliability for a single evaluator was 0.63, with that number rising to 0.77 for two raters, 0.84 for three raters, and 0.87 for four raters (Driessen, E., et al., 2007). It is generally accepted that a minimum inter-rater reliability of 0.80 is required to make valid end of term summative decisions (Axelson, et al., 2009).

The statistical analysis indicates that applying the grade anchors derived from the focus group process did not improve inter-rater reliability to a degree to permit summative assessment decisions and that much of the variance was attributable to rater effect, indicating that variation in the scores assigned to the essays may be attributed to variations among the faculty assessors. Upon review of the study, and

reflection on the themes derived from faculty assessors' experiences, there appear to be reasonable explanations for these results and support for further use of this qualitative methodology for assessment of portfolios.

The literature is replete with recommendations for coupling of faculty mentoring and feedback during the portfolio construction process, and faculty calibration beforehand (Tochel, et. al., 2009; Dornan, et.al., 2011). These were both absent from this study due to the fact that the study was applied to sets of essays that were gathered from a previous administration of the course in which mentoring and calibration were not provided due to resource constraints. A strong theme, that was identified in the focus group discussions about faculty assessors' experiences, was a desire for training and practice using the rubric. It was uniformly agreed that this was important and would be beneficial. This training could take the form of the frame-of-reference training recommended by Bernardin and Buckley (1981) whereby the faculty would independently apply the scoring rubric to benchmarked essay samples and then follow this up with a course director facilitated group discussion of the grades assigned and how the rubric was applied. Both improved, and student performance more consistent with the rubric, might also be anticipated if the new rubric were communicated to the learners at the beginning of a future offering of the course.

Another possible source for the low inter-rater reliability was the relatively small number of essays (10/set) that were used. A different result might be obtained if each

set contained 60-70 essays, thereby increasing the precision and stability of inter-rater agreement measures. Again, the decision to use a smaller number of essays was based on what the authors thought to be reasonable expectations of faculty participants outside of regular course administration responsibilities.

It is interesting to note that the standard setting faculty expressed a desire to have a detailed and specific "checklist" style grading rubric. They wanted something to hold on to in which they felt confident. They all also recognized the risk of diminishing student use of the portfolio for self-assessment that would possibly occur with the use of such a checklist. Perhaps the use of a holistic grading rubric, such as the one developed for this study, coupled with benchmarked examples of essays at different grade levels, might be an acceptable compromise and a useful component of faculty calibration. This might also have the effect of allaying faculty concerns regarding unrealistic expectations of student performance, by providing more concrete expectations of the level of performance at each scoring category.

Other themes formulated in the discussion with the standard setting faculty were a tendency to want to read all the essays and rate them by comparison to each other, and a concern regarding the number of grade categories used. It seems natural that when presented with an assessment task that requires high inference that one would want to create a relative scale by ranking. This is not something that the focus group facilitator instructed the faculty to do, or not to do, and is possibly another source of reduced inter-

rater reliability. It became apparent during the faculty group discussion that some faculty had used this methodology to assign grades, and others had not. The choice of the number of grade categories for any assessment is critical. The number of categories should be consistent with the degree of confidence the evaluator(s) has in her/his ability to discriminate between performances or products (Downing, et.al., 2009). The grade categories selected for this essay were probably too many. For the high degree of inference necessary to assess a collection of constructed responses, such as a portfolio; it would probably be best to limit decisions to pass or no pass. This is arguably the most important decision we need to make when determining learner competence. Limiting the choices to this "yes/no" type decision would also potentially raise inter-rater reliability and simplify faculty calibration.

Researchers have described the detrimental effects of negative faculty attitudes about portfolios (Tochel, et. al., 2009) and the standard setting faculty commented about this as well. Not only were they concerned that faculty demonstrating negative attitudes about portfolios would lead to students producing lower quality work; they also indicated that these negative attitudes could lower faculty expectations of student performance. Part of any faculty calibration program should be a process by which convincing evidence is provided to support the process and faculty buy in sought. Faculty also need to feel free to express concerns to the portfolio coordinator(s) without fear of repercussion and should be urged to consider the harmful effects of expressing negative views of portfolios to students.

Perhaps the most encouraging results of this study were the descriptions of the experiences of the faculty assessors. Despite reporting greater difficulty using the second, focus group derived, rubric, they also indicated that they felt more confident in the grades assigned using this rubric and, related to this, the grades seemed more defensible. Confidence in grading based on defensibility is particularly important when attempting holistic assessment of constructed responses, such as those commonly included in portfolios. That these faculty evaluators, even without any calibration or practice, felt a greater degree of confidence in the grades they provided, to the degree that they indicated they could defend them based on the criteria developed, is promising. That they also found the second rubric to be possibly more efficient is even more encouraging that calibration, including practice using this rubric, would be viewed as a worthwhile endeavor.

V. Conclusion

A. Contributions of the Study

This is not the first study in health professions research to apply focus group methodology to portfolio assessment. There have been studies examining student and faculty perceptions and satisfaction regarding portfolio use (Ryan, 2011; Webb, et. al., 2012). Others have applied Delphi or modified Delphi techniques, which have varying degrees of similarity to focus group methodology, to determine portfolio content (Jenkins, et.al, 2012) and depth of student reflection (McNeill, et. al., 2010).

The uniqueness of this study lies in the use of a focus group as a standard setting methodology. A search of the literature yielded only one previous attempt to use focus groups for standard setting. An iterative process to refine standards was demonstrated to increase inter-rater reliability (Jaeger, 1982). However, this process was applied to a selected response (i.e., low inference) exam and the standard setters did not engage in group discussion as part of the process. Others have suggested that the resources expended, in this instance, were not necessary, given less costly and acceptable alternatives. As Norcini states, "the intent is to demonstrate due diligence, not endurance" on the part of the standard setters (Norcini, et al., 1997). In this study, the focus group process was applied to the development of anchors for a high inference format which may warrant the added expense, particularly if through further development and scrutiny, these standards may be shown to produce improved levels of inter-rater reliability on subsequent administrations of the same assessment and then

assessment of entire portfolios. The first step though, and the objective of this study, was to determine whether these methods have a useful effect at all. This study is potentially the beginning of the literature supporting this particular method.

B. Limitations of the Study

The limitations of this study included the examination of a component of a portfolio, but not the entire portfolio. It is therefore not possible to determine whether the same effects will hold true for the holistic assessment of a complete portfolio. This study also applied the standards created to only a small number of reflective essays. If more faculty time was available, such as if the standards were applied to the assessment of the 65-70 essays generated during the normal administration of the course, and faculty were calibrated to the standards, it would be interesting to see if similar results would be obtained. This might be more feasible if the study were conducted "in situ" as part of the actual offering of the course. The small sample of essays used most likely had the effect of exaggerating the statistical effects of disagreement among raters, while limiting the number of scoring categories (three) probably had the opposite effect. The decision was made to not include a control group in this study, but rather to compare the resulting inter-rater reliability to published results of similar studies. The variability of the portfolios constructed, and relative lack of information regarding standard setting processes available in these studies, makes the validity of this comparison open to speculation. This is the best available evidence at this time, though, and therefore the best comparison that can be made.

C. **Opportunities for Additional Inquiry**

The greater degree of confidence and defensibility the faculty assessors perceived when using the grading rubric, with the focus group derived anchors, make one inclined to continue to explore the use of this methodology. Given the opportunity to reproduce the study with faculty calibration, student mentoring/feedback during the process, and a larger number of essays, would the same results be obtained? Could this methodology then be successfully applied to larger constructed response products such as entire portfolios? How does this method compare with other methods with regards to use of resources? Is the degree of inter-rater reliability really the appropriate measure when considering standard setting for a high inference assessment such as portfolios, or should we be seeking the evidence provided by an accurate description of a credible and defensible process fairly applied? Would a better approach be to use two-three trained assessors for each portfolio, with the grade determined by consensus as the result of discussion, instead of pursuing "high enough" inter-rater reliability among single raters?

As yet, the proper balance between freedom and guidelines for learner selection of content to include in portfolios, and for faculty assessment of portfolios, eludes us. Perhaps the standard setting method described in this paper moves us closer by nudging the needle back slightly towards greater freedom. It is good to remember as the late 19th century Chicago architect John Wellborn Root wrote, "Reason should lead

the way, however, and imagination take wings from a height to which reason has already climbed." (Monroe, 1896). Let us then reason together to imagine and create standards that are truly authentic to what is being assessed and justifiably applied to the earnest efforts of our learners.

CITED LITERATURE

Axelson, R. D., (2009). Reliability. Assessment in Health Professions Education. edited by Downing, S. M. and Yudkowsky, R. Chapter 3: 57-73.

Bernardin, J. H. and Buckley, M. R., (1981). Strategies in Rater Training. The Academy of Management Review, 6: 205-212.

Buckley, S., et al., (2009). The educational effects of portfolios on undergraduate student learning: A Best Evidence Medical Education (BEME) systematic review. MedicalTeacher, 31: 282-298.

Chambers, D. W., (2004). Portfolios for determining initial licensure competence Journal of the American Dental Association, 135: 173-184.

Dornan, T., et. al., (2011). Medical Education Theory and Practice. edited by Dornan, T., et. al. Chapter 13: 222-225.

Downing, S. and Yudkowsky, R., (2009). Assessment in Health Professions Education. edited by Downing, S. M. and Yudkowsky, R. Chapter 7: 149-165.

Driessen, E. W., et al., (2003). Use of portfolios in early undergraduate medical training. Medical Teacher, 25: 18-23.

Driessen, E., (2005). The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: a case study. Medical Education, 39: 214-220.

Driessen, E., et al., (2007). Portfolios in medical education: why do they meet with mixed success? A systematic review. Medical Education, 41: 1224-1223.

Friedman Ben David, M., et al., (2001). AMEE Medical Education Guide No. 24: Portfolios as a method of student assessment. Medical Teacher, 23: 535-551.

Goldie, J., et al., (2007). Teaching professionalism in the early years of a medical curriculum: a qualitative study. Medical Education: 610-617.

Grant, A. J., et al., (2007). Exploring students' perceptions on the use of significant event analysis, as part of a portfolio assessment process in general practice, as a tool for learning how to use reflection in learning. BMC Medical Education,7:5.

Harris, I.B., (2002). Qualitative Methods. International Handbook of Research in Medical Education, edited by Norman, G.R., Chapter 2: 45-95.

Jaeger, R. M., (1982). An Iterative Structured Judgment Process for Establishing Standards on Competency Tests: Theory and Application. Educational Evaluation and Policy Analysis, 4(4): 461-475.

Jenkins, L., et. al., (2012). Development of a portfolio of learning for postgraduate family medicine training in South Africa: a Delphi study. BMC Family Practice, 13:11: 1-10.

Karlowicz, K. A., (2010). Development and testing of a Portfolio Evaluation Scoring Tool. Journal of Nursing Education, 49(2): 78-86.

Lincoln, Y. S., & Guba, E. S., (1985). Naturalistic Inquiry, Newbury Park, CA: Sage Publications.

McCready, T., (2006). Portfolios for the assessment of competence in nursing: A literature review. International Journal of Nursing Studies, Available online April 24, 2006.

McNeill, H., et. al., (2010). First year specialist trainee's engagement with reflective practice in the e-portfolio. Advances in Health Science Education 15: 547-558.

Melville, C., et al., (2004). Portfolios for assessment of paediatric specialist registrars. Medical Education, 38: 1117-1125.

Monroe, Harriet, (1896). John Wellborn Root: A Study of His Life and Work. Chapter IV: 66.

Norcini, J. J., et al., (1997). The Credibility and Comparability of Standards. Applied Measurement In Education, 10(1): 39-59.

Norcini, J., et al., (2002). Combining Tests and Standard Setting. International Handbookof Research in Medical Education, edited by Norman, G.R., Chapter 25: 811-834.

O'Sullivan, P. S., et al., (2004). Demonstration of Portfolios to Assess Competency of Residents. Advances in Health Sciences Education, 9: 309-323.

Pinsky, L. E., et al., (2004). Diving for PERLS, Working and Performance Portfolios for Evaluation Reflection on Learning. Journal of General Internal Medicine, 19: 582-587.

Pitts, J., et al., (2002). Enhancing reliability in portfolio assessment: discussions between assessors. Medical Teacher, 24: 600-609.

Rees, C. E., et al., (2004). The reliability of assessment criteria for undergraduate medical students' communication skills portfolios: the Nottingham experience. Medical Education, 38: 138-144.

Rees, C., (2005). The use (and abuse) of the term 'portfolio'. Medical Education, 39: 436.

Roberts, C., et al., (2002). Portfolio-based assessments in medical education: are they valid and reliable for summative purposes? Medical Education, 36: 899-900.

Ryan, M., (2011). Evaluating portfolio use as a tool for assessment and professional development in graduate nursing education. Journal of Professional Nursing 27 (2): 84-91.

Tate, P., et al., (1999). Assessing Physicians' Interpersonal Skills via Videotaped Encounters: A New Approach for the Royal College of General Practitioners Membership Examination. Journal of Health Communication, 4: 143-152.

Tochel, C., et al., (2009). The effectiveness of portfolios for post-graduate assessment and education: BEME Guide No 12. Medical Teacher,31: 299-318.

Usherwood, T., et al., (1992). Profile-based assessment of student project reports. Medical Teacher, 14 2/3: 189-196.

Webb, C., et al., (2003). Evaluating portfolio assessment systems: what are the appropriate criteria. Nurse Education Today, 23: 600-609.

Webb, T. P., and Merkley, T. R., (2012). An evaluation of the success of a surgical resident learning portfolio. Journal of Surgical Education, 69 (1): 1-7.

Wilkinson, T. J., et al., (2002). The use of portfolios for assessment of the competence and performance of doctors in practice. Medical Education, 36: 918-924.

VITA

NAME: Kevin Eugene Van Kanegan

EDUCATION: B.S., Biology, University of Illinois, Urbana, Illinois, 1989

B.S., Dentistry, University of Illinois at Chicago-College of Dentistry, Chicago, Illinois, 1991

D.D.S., University of Illinois at Chicago-College of Dentistry, Chicago, Illinois, 1993

TEACHING EXPERIENCE: Department of Oral Medicine, University of Illinois-College of Dentistry, Chicago, Illinois,1997-2002

Department of Restorative Dentistry, University of Illinois-College of Dentistry, Chicago, Illinois, 2002-2011

College of Dental Medicine-Illinois, Midwestern University, Downers Grove, Illinois, 2011-present

HONORS: Omicron Kappa Upsilon-National Honorary Dental Fraternity

University of Illinois at Chicago, Alumni Loyalty Award

PROFESSIONAL MEMBERSHIP: American Dental Education Association
American Dental Association
Illinois State Dental Society
Chicago Dental Society

PUBLICATIONS: Raja, S., Rajagopalan, C.F., Patel, J. and Van Kanegan, K., (2014) Teaching Dental Students About Patient Communication Following an Adverse Event: A Pilot Educational Model. Journal of Dental Education, 78 (5): 667-672.