EMBRY-RIDDLE Aeronautical University SCHOLARLY COMMONS

Journal of Aviation/Aerospace Education & Research

Volume 25
Number 1 *JAAER Fall 2015*

Article 3

Fall 2015

A Study of How Flight Instructors Assess Flight Maneuvers and Give Grades: Inter-rater Reliability of Instructor Assessments

Beth M. Beaudin-Seiler
Western Michigan University, beth.seiler@wmich.edu

Ryan Seiler
Western Michigan University, ryan.seiler@wmich.edu

Background

Collegiate flight training is expensive, both in time and money. Flight training costs for a student wishing to pursue a professional pilot degree can increase the expense of a traditional four-year degree by $60,000. It may also add a year or more of school. Flight training, therefore, should be complete and comprehensive without undue repetition or delays in the progress of training. Examining assessment practices and performance outcomes is important to any flight training program to contain costs and to graduate quality aviation professionals. But it can be difficult to fit the technical aspects of flight training into a traditional model of academic assessment for graduation requirements.

Progress in flight training is similar to the educational and training path for nurses in that the proficiency and competency must be demonstrated. Nursing, like flying, has a hands-on component where students are expected to demonstrate competency in clinical areas as a condition for graduation (Bondy, 1983; Eymard, Davis, & Lyons, 2013). A student cannot pass theoretical course work with As and Bs and then demonstrate an unsatisfactory performance in their skill area and expect to graduate in either career path. Education research, however, reveals that a student's education depends to a large degree on the quality of the instructor, and that instructors are least effective in the early stages of their teaching careers (Weisberg, Sexton, Mulhern, & Keeling, 2009).

To pull together traditional academic assessment of coursework and demonstration of technical proficiency in flying for an overall grade can be difficult and confusing. In nursing, research has shown that novice clinical teachers are unsure about their role (Scanlan, Care, & Gessler, 2001). They are reluctant to fail students in clinical practice because they are not confident in their judgments and the ultimate decision about their students' abilities (Scanlan,

Care, & Gessler, 2001). Parallels to this can be drawn to newly certificated flight instructors in a collegiate setting. In the program at Western Michigan University, newly certificated flight instructors are often fellow classmates of their students and may not be confident in their ability to properly evaluate them.

Interviews conducted with our program flight instructors indicated it is difficult to fail a student when you are going to sit next to him or her in a class (Beaudin-Seiler, unpublished, 2014). Furthermore, very few understand the impact that assigning grades can have on students. As one instructor stated, "It's really hard to tell a student they failed a lesson because you don't know if they are going to get mad and yell at you, or start to cry, so I just pass them and note which maneuvers we need to continue to work on in the next lessons" (Beaudin-Seiler, unpublished, 2014. There may be merit to this thinking. Several research studies conclude that for some people a common reaction to negative performance feedback is to increase aggression towards the source of the feedback (Barry, Chaplin, & Grafeman, 2006; Bushman et al., 2009; Jones & Paulhus, 2010; Konrath, Bushman, & Campbell, 2006; Stucke & Sporer, 2002; Vaillancourt, 2013).

Research on college students has shown that there are few events that are as important as receiving grades for their course work (Vaillancourt, 2013; Crocker, Karpinski, Quinn, & Chase, 2003). Studies have shown that a student's self-worth is strongly linked to their academic performance, and moreover, poor grades significantly impact self-esteem (Vaillancourt, 2013; Crocker et al., 2003). Research also shows that grades affect student interest, confidence, self-efficacy, motivation, and future performance (Carey & Carifio, 2012; Docan, 2006).

For college students, research suggests that they are much more focused on grades than the reasons provided from the instructor on why they earned that particular grade. Feedback

from the instructor is all but ignored and the focus remains on the poor grade (Vaillancourt, 2013). While instructors may look at grades as a measure of mastery and/or a motivational tool, students view grades as a key to a future job or scholarship thereby creating a conflict between instructor and student (Edgar, Johnson, Graham, & Dixon, 2014).

Research on the evolution of grades, by Rojstaczer and Healy (2012), indicate that collegiate grades have increased since 1960. The most common grade given in an academic setting is an A, suggesting that instructors have gradually lowered their standards over time. The researchers argue that without regulation, or at least strong guidance, grades at American colleges and universities will likely continue to have less and less meaning (Rojstaczer & Healy, 2012).

Data from our program indicated that flight maneuvers, when not assessed to a standard, led to repeated flight lessons, additional cost, and needless additional one-on-one ground instruction time (Beaudin-Seiler, unpublished, 2014). However, when instructors have confidence in their ability to assess maneuvers properly, students with satisfactory performances will progress and only those needing remediation will need to repeat. Grades will become reflective of true performance. Educational research suggests that simple, straightforward, and easily understood grading systems which are consistent, and result in predictable, fair and accurate assessment of student performance are the best systems for students (Carey & Carifio, 2012).

Project Description

In the spring of 2013, the authors undertook a project to determine how reliable our certificated flight instructor staff was in assessing maneuvers and providing grades for flight lessons through an inter-rater reliability study. The flight program has made it a practice to

follow the progress of its flight students, and to identify areas both in theoretical courses and flight training where students may struggle and need to repeat lessons (Beaudin-Seiler, 2013). However, an examination of the role that the instructor plays in helping the flight students to determine what constitutes a proficient maneuver and a passing grade for lessons had never been undertaken.

Our College of Aviation ("College") has over 800 students pursuing Bachelor of Science degrees in three distinct programs: flight, maintenance, and management. Over 400 students are pursuing a professional flight degree. The flight program utilizes the Cirrus SR 20 with an Avidyne R9 avionics package for primary flight training and Piper Arrows and Seminoles for commercial and multi-engine training. The program also utilizes advanced simulation configured to Cirrus, Seminole and a CRJ-200 (www.wmich.edu/aviation).

The College of Aviation has a building block approach to its curriculum. Computer based training is utilized first, followed by simulation, which is then followed by aircraft training. The College employs six full time flight instructors in faculty tenure-track positions to oversee the quality of primary instruction. Faculty also conduct phase and stage checks and provide guidance and instruction to struggling students.

There are four full time staff flight instructors which include the Chief Flight Instructor, Director of Standards, and two Program Managers. The people serving in these positions conduct phase and stage checks and provide guidance and instruction to students who are underperforming in their courses. The primary flight instruction load is conducted by approximately 40 temporary flight instructors. Each instructor has 4-6 students assigned to him or her and flies three to four times per week with each student. The more experienced flight

instructors are assigned to the multi-engine curriculum, while those with less experience teach private pilot students.

The purpose of this research project was to determine how closely aligned our certified flight instructors were in the assessment of maneuvers and lesson grades to an agreed upon "gold standard". The gold standard was determined by independent ratings assigned by flight faculty in the program. Each maneuver conducted by the student was assessed and graded. The research was also undertaken in order to provide opportunities for practice and training at assessing maneuvers and grading lessons; and to re-test the flight instructors to determine if the training had an impact on how closely aligned their assessments and grades were to the "gold standard".

Method

Development of Scenarios

A small expert committee consisting of a research associate, one program manager, and one flight faculty member was created to evaluate certain maneuvers in the private pilot curriculum. The committee determined that there were certain maneuvers, such as slow flight, power-on stalls, power-off stalls and steep turns that could be performed in the simulator for use in the project.

The program manager and flight faculty member took turns flying each maneuver in the simulator. The flight exercises were recorded (both audio and video) in order to display the instrumentation as well as provide a visual of the maneuvers. They flew different proficiency levels of each maneuver to create a library of videos that could be selected in scenario development. Once the various levels of maneuvers had been flown, the program manager and flight faculty member independently scored each maneuver using the College's required grading

format on a scale ranging from 1 (*low performance*) to 4 (*high performance*). When they finished their scoring, the committee met to review and discuss any disagreements. After further discussion and review, the program manager and flight faculty member agreed upon a final score for each maneuver. This established the "gold standard" for each maneuver.

Four scenarios were developed using the maneuvers recorded in the simulator. Each scenario required all four maneuvers to be graded by the instructor. Two of the scenarios replicated lessons given to students early in the curriculum where proficiency of the maneuver would not necessarily be required. The other two scenarios replicated lessons late in the curriculum where mastery of the maneuvers would be required.

In the College's private pilot curriculum there are certain lessons which required students to fly with a different instructor. With those lessons in mind, instructors received limited information about the student, just as would be expected if it was a student they had not flown with before. Background information on the student provided to the instructor included the preparedness of the student in previous lessons, any current difficulties, performance on all clearing turns and pre-maneuver checks, and a reminder of which lesson they were completing.

The backgrounds of the students did not necessarily match the proficiency level of the maneuver shown. For example, one student's background indicated the student had been struggling with their previous instructor, had met with program managers, and had been put on a plan of action which included chair flying and other activities. Yet the video selections for this scenario showed exemplar maneuvers. This allowed the committee to better understand if flight instructors were grading based on history or observed performance.

After full scenarios were developed, the program manager and flight faculty member independently reviewed the events, graded the maneuvers, and provided an overall grade for the

lesson using the College's overall grading format (the traditional academic format of A, B, C, D or E). Upon the completion of the scoring and grading of each scenario, the committee met to discuss the assessments, ultimately agreeing on a "gold standard" for the grading of each lesson.

Initial Assessment Exercises

All flight instructors were required to complete the exercises. Two were completed in February 2013 and two were completed in March 2013. Flight instructors were given a packet of information, a thumb drive with the video maneuvers, and headphones. The only identifiable information asked of the instructors was their level of experience in giving flight instruction and whether they were approved as a check instructor. The packet included detailed instructions, background information of the "student" and the lesson they were about to complete, and a grading sheet for the lesson. The grading sheets were pre-graded for maneuvers that had been completed, only required the instructor to provide maneuver grades for the four video maneuvers watched, and an overall lesson grade based on all the information provided to them.

The thumb drive held the video maneuvers to each of the scenarios. They were instructed to make sure they were scoring and watching the same lesson, to watch each maneuver only once, and to grade only those line items highlighted on the grading sheet. After grading the maneuvers, the evaluation panel was asked to look at the lesson as a whole and provide an overall grade for the exercise based on what they saw in the maneuvers and what the other line items reflected for grades. They were also asked to provide comments about the students. Lesson standards were identified by graying out the box. An example of the scenario packet given to the flight instructors is included in the Appendix.

Initial Assessment Findings

Data from the initial assessment exercises were collected. No identifying information other than length of flight instruction experience and whether the respondent was a check instructor was asked. We were able to categorize responses by experience of 1 year or less, 13 months to 2 years, and more than 2 years. Each maneuver as well as each scenario had a gold standard that was achieved by expert raters. To assess levels of agreement between instructors, the Cohen's Kappa Coefficients were assessed using SPSS statistical software. Operationally, the level of agreement was defined using a scale from Landis & Koch (1977) as: *almost perfect* (.81 to 1.0); *substantial* (.61 to .80); *moderate* (.41 to .60); *fair* (.21 to .40); *slight* (0 to .2); and *poor* (< 0). Table 1 is the average Cohen's Kappa Coefficient for each group.

The data revealed to us what research in other fields had already indicated. The less experienced instructors had less agreement with the gold standard than more experienced instructors. This was similar to the nursing field, which has found that less experienced clinical instructors have difficulty evaluating students because they lack the proper level of preparation (Luhanga, Yonge, & Myrick, 2008).

1 Year or Less of Flight Instruction Experience

Frequency analyses on each maneuver for each lesson were conducted to examine another level of agreement. Table 2 depicts data from the 1 year or less of flight instructor experience for each maneuver and each lesson using modes and percentage of agreement with the gold standard. A grade of 1 equals *low performance* and a grade of 4 equals *high performance* for the maneuvers. Letter grades assigned to the overall lesson are the traditional A, B, C, D, E scale.

This data reveals a number of things about our less experienced flight instructors. First, it seems as though they may be assessing grades according to what is required to earn a passing grade for a lesson. In the private pilot curriculum the maneuvers would need to be demonstrated to a level 2 in order for the students to pass. Most instructors in this category merely gave the maneuver a 2, even when an exemplar video clip was shown to them. To assess the maneuver a 2 rating does not assist a student in better understanding the difference between merely passing or excelling. Second, it is much easier for novice instructors to assess to a scale that is objective and has clear parameters. For example, in the latter two lessons (#43 and #48), the maneuvers must be demonstrated to a level 3 which in our curriculum is the same as the Federal Aviation Administration's (FAA) Practical Test Standards (PTS). The FAA PTS is easy to understand, clearly defined, and linked to definitive outcomes such as altitude cannot deviate more than +/- 100 feet in the maneuver.

This category of instructor had the most agreement with the gold standard when they knew what the grading scale was and could clearly see the level demonstrated in the video. When those clear, delineated parameters were not set, this category of flight instructors struggled more to assess maneuvers to the gold standard. Again, a finding not uncommon in nursing research, which shows that assessment tools that use a more discriminatory grading system and have clear descriptors are welcomed by instructors that have to grade technical competency (Heaslip & Scammell, 2012; Bondy, 1983).

Finally, this category of flight instructors seemed to have a difficult time failing the overall lesson. This is also common in other technical performance-based assessment programs. Research indicates that there are many examples of instructors being reluctant to give a failing grade to a marginal or even unsafe student in a nursing program (Luhanga et al., 2008).

13 Months to 2 Years of Flight Instruction Experience

Table 3 depicts data from 13 months to 2 years of flight instructor experience for each maneuver and each lesson using modes and percentage of agreement with the gold standard. The data reveals a number of issues regarding grading from the more experienced instructor base. First, and as we saw in the novice data collected, it appears instructors are merely grading to whatever the lesson standard calls for rather than to what was actually observed. This is supported by the responses in the early lessons (#27 and #28) where the maneuvers need only be to a level 2, yet the video shown was exemplar at a level 4, and no instructor graded it a level 4. Conversely, Lesson 43 – power-off stalls – shows a very poor maneuver, yet most instructors merely gave the rating a 2, which is less than what the lesson standard would require but higher than the gold standard. Inexperienced instructors may be unaware as to what constitutes levels 1, 2 or 4 in a maneuver. More experienced instructors, however, may have a different and perhaps ineffective philosophy in how they go about assessing student performance. This is supported by qualitative feedback from an instructor in this group who stated, "I don't give 4's".

There is more agreement with the gold standard on grading of maneuvers as well as overall lesson grades, indicating that the more experience the instructor has, the more confident they are in assigning grades. The nursing literature again supports this finding showing that research suggests new faculty have a more difficult time with clinical evaluation than senior faculty (Seurynck, Buch, Ferrari, & Murphy, 2014; O'Connor, 2001).

2+ Years of Flight Instruction Experience

Table 4 depicts data from the 2+ years of flight instructor experience for each maneuver and each lesson using modes and percentage of agreement with the gold standard.

These findings again support that notion that instructors may be merely checking to see that the elements of the maneuver are above what the lesson standard requires, and not actually grading the observed maneuver. The researchers adopted the term "pencil-whipping" for this experience. The instructors seem to merely check off the maneuver as being above standard, not really making an assessment of it. In this more experienced group, some instructors may be philosophically opposed to assigning high grades. The higher levels of agreement in failing of a student for a particular lesson shows that instructors are confident in knowing when the lesson should be failed.

Practice Exercises and New Hire Curriculum

The data revealed that the cadre of certificated flight instructors in our program would probably benefit from practice and discussions regarding:

- Assessing the observed maneuver itself, not against the lesson standard
- Allowing exemplar performances to be assessed as exemplar
- Knowing when (and that it is ok) to give a failing overall grade

A two tiered approach was taken, one that addressed the current flight instructor group and one that addressed newer flight instructors. First, a new hire training course was designed by committee members for all new flight instructors beginning in the fall of 2013. This curriculum utilized various maneuvers from the video-taped library. The new hires were shown the video and asked to assess a grade for the maneuver based on the College's 1-4 scale. After the new hires had an opportunity to consider a grade for each video, a discussion was led by the flight faculty member on the committee. They discussed what they saw, what the gold standard for that maneuver was, and whether they would have changed their grade after having discussion.

This debriefing allowed new hire instructors to ask questions, clarify assumptions, and to posit scenarios such as "what if the student did this" to the flight faculty in a friendly environment.

During one of the new hire sessions, an instructor spoke about the steep turn maneuvers he had just watched (or reviewed). When asked why he failed the maneuver, he said, "I thought that the student has to go directly in to the left turn followed by the right turn with no break in between, so that's why I failed it". This justification would be correct for a commercial level student, but not a private pilot student. Improving an instructor's ability to better assess a student would result in less failed maneuvers and save time and money.

Three new hire groups were put through this curriculum in the fall of 2013. Discussions were fruitful. Not only were the groups able to listen to the rationale of their peers, but the flight faculty member was able to dispel any untruths or assumptions the new hires had on how to assess grades.

The second phase of this research involved the development of practice sessions for current flight instructors. The College holds mandatory flight instructor meetings once per month. The researchers were allowed to come in to these meetings, every other month, to review a selected video, have the group score it and then discuss why they scored it the way they did. Again, knowledge was gained during the discussion where more experience instructors were able to say, "This is what I saw and why I graded it a certain way". Younger, more inexperienced instructors were able to ask questions and benefit from the experienced instructors. The facilitator of the discussion was again the flight faculty member on the committee, who was able to explain the gold standard and why it was selected. Practice sessions with all flight instructors occurred September 2013, November 2013, and February 2014. This led to the post-test data collection phase where all instructors were once again given the exact same package of videos

and instructions to provide maneuver grades and overall lesson grades. Post-test data collection occurred in March and April 2014.

Post-Test Assessment Findings

Just as before, data from the post-test assessment exercises were collected with no identifying information other than length of flight instruction experience and whether the respondent was a check instructor. Cohen's Kappa Coefficients were completed for each category of responses to understand the level of agreement between each group and the gold standard. Table 5 is the average Cohen's Kappa Coefficient for each group in the post-test phase, as well as the initial assessment phase.

1 Year or Less of Flight Instruction Experience

Table 6 depicts data for flight instructors with no more than one year of experience for each maneuver and lesson using modes and percentage of agreement with the gold standard from the post-test assessment phase.

These findings indicate that this group of instructors has made progress in rating and assessing maneuvers and overall grades of lessons. While they are not issuing as many top level ratings (4), they seem to have become more understanding of the need to grade the lesson accurately and issue failing grades (Es), when appropriate. This group increased their percentage of agreement on 15 out of 23 line items from the initial assessment to the post-test assessment.

13 Months to 2 Years of Flight Instruction Experience

Table 7 depicts data from the 13 months to 2 years of flight instructor experience for each maneuver and each lesson using modes and percentage of agreement with the gold standard from the post-test assessment phase.

This group of instructors had lower agreements in the post-test assessment as compared to the initial assessment data. However, this makes sense. During the initial testing phase, this group of instructors would have had one year or less of experience. A year later, that same cadre had at least 13 months and as much as 2 years' worth of experience.

This group of instructors did not have the benefit of the new hire curriculum and only had the practice sessions that were held every other month during flight instructor meetings. The average Cohen's Kappa Coefficient for this cadre of instructors in the initial assessment testing was .197 and in the post-test assessment testing this cadre of instructors was .270. It's clear that the results show progress but it is not to the level as those instructors who were exposed to the curriculum. Further examination of the data reveals that 35% of the line items (8 out of 23) increased in agreement from the initial assessment. However, when looking at this group of flight instructors as a cohort (as the 1 year or less experienced group from the previous year) we find that 12 out of the 23 line items (52%) increased in agreement.

2+ Years of Flight Instruction Experience

Table 8 depicts data for the group of flight instructors with two or more years of experience for each maneuver and each lesson using modes and percentage of agreement with the gold standard from the post-test assessment phase.

This group of instructors had slightly lower agreements from the initial assessment phase compared to the post-test phase. However, this group was not subject to the new hire curriculum, but only had the opportunity to practice rating and grading maneuvers every other month in flight instructor meetings. Again, this group would have been the 13 month to two years of experience group from the initial assessment phase.

The average Cohen's Kappa Coefficient from this cadre of flight instructors in the initial phase was .325 and now in the post-test assessment phase is .344, a slight improvement. This group tends to be able to recognize high performance by rating maneuvers 4s better than the other groups, and recognizing when failing a lesson is appropriate by giving Es more than the other groups. In looking at the details of the data from this group, 11 of the 23 line items (48%) increased in agreement in this experience level of instructors. This group increased the most in recognition and assessment of high performance, issuing ratings of 4s more than any other group.

Discussion and Future Work

The art of assessing performance and issuing grades is difficult in any environment. Grades are increasingly more important to students as they are the basis for admissions, scholarships and self-worth (Randall & Engelhard, 2010; Crocker et al., 2003; Docan, 2006). Empirical research suggests that even experienced teachers, skilled in both teaching and assessing grades, often use other characteristics such as behavior and motivation to determine what the final grade will be for a student, causing angst and difficulty for the teacher (Randall & Engelhard, 2010). For collegiate flight training programs, it is important that certified flight instructors understand and exhibit confidence in assessing flight maneuvers. However, flight instructors are trained in how to fly, not necessarily how to assess it. It is reasonable to conclude they may find themselves struggling to determine performance as well.

The results of this study indicate there is still much work to do to calibrate the rating of maneuvers and grading of lessons between instructors and the expected gold standard. The scores from the Cohen's Kappa Coefficient are not at an acceptable level, but there is movement in the right direction. The results of the post-test assessment data show that the development and addition of assessment practice in the new hire curriculum and the practice sessions at flight

instructor meetings has helped the least experienced flight instructors improve rating maneuvers and grading lessons. Post-test assessment results show that the least experienced group of instructors has more agreement with the expected gold standard than instructors with 13 months to 2 years of experience instructing.

Practice sessions have also helped the other groups by increasing their understanding as to when it is appropriate to give both high marks and failing ones. The post-test assessment data show that we still need to work on the "pencil-whipping" problem. Flight instructors may be just giving the rating that matches the lesson standard, rather than critically analyzing which rating should be given for the maneuver. This is evident in the relatively few high marks in maneuver ratings. While some still struggle with providing failing grades, we have seen the most progress in this area as all groups except for two lessons from the experienced group (two or more years of experience) have increased their agreement to the gold standard.

While nothing is a substitute for actual experience, the development of this new hire curriculum coupled with practice sessions allow instructors to view maneuvers, rate and discuss their thoughts has proven to generate some level of agreement among those instructors who lack experience. Future work will involve the development of practice video sessions at flight instructor meetings as well as increasing the amount of open discussion between flight instructors and flight faculty on what the gold standard is and how to recognize it. Annual assessment will continue to provide an understanding on how closely aligned our instructors are in evaluating students. This ongoing process will help provide information on when new techniques for training may be needed.

References

Barry, C. T., Chaplin, W., & Grafeman, S. (2006). Aggression following performance feedback: The influences of narcissism, feedback valence, and comparative standard. *Personality and Individual Differences, 41*(1), 177-187. http://dx.doi.org/10.1016/j.paid.2006.01.008

Beaudin-Seiler, B. (2013). Private pilot progress: Where do we fall down? *The Journal of Aviation/Aerospace Education and Research, 22*(2), 31-39.

Beaudin-Seiler, B. (2014). *Cost and time of private pilot.* Unpublished manuscript.

Bondy, K. (1983). Criterion-referenced definitions for rating scales in clinical evaluation. *Journal of Nursing Education, 22*(9), 376-382.

Bushman, B., Baumeister, R., Thomaes, S., Ryu, E., Begeer, S., & West, S. (2009). Looking again, and harder, for a link between low self-esteem and aggression. *Journal of Personality, 77*(2), 427-446. http://dx.doi.org/10.1111/j.1467-6494.2008.00553.x

Carey, T., & Carifio, J. (2012). The minimum grading controversy: Results of a quantitative study of seven years of grading data from an urban high school. *Educational Researcher, 41*(6), 201-208. http://dx.doi.org/10.3102/0013189X12453309

Crocker, J., Karpinski, A., Quinn, D. M., & Chase, S. K. (2003). When grades determine self-worth: Consequences of contingent self-worth for male and female engineering and psychology majors. *Journal of Personality and Social Psychology, 85*(3), 507-516. http://dx.doi.org/10.1037/0022-3514.85.3.507

Docan, T. N. (2006). Positive and negative incentives in the classroom: An analysis of grading systems and student motivation. *Journal of Scholarship of Teaching and Learning, 6*(2), 21-40.

Edgar, L. D., Johnson, D. M., Graham, D. L., & Dixon, B. L. (2014). Student and faculty perceptions of plus/minus grading and its effect on course grade point averages. *College Student Journal*, *48*(1), 184+. Retrieved from http://go.galegroup.com/ps/i.do?id=GALE%7CA372252081&v=2.1&u=lom_wmichu&it=r&p=AONE&sw=w&asid=1505ddf17559bb74ce3d973c10bd293a

Eymard, A., Davis, A., & Lyons, R. (2013). Progressive clinical evaluation tools based on the quality and safety education in nursing competencies. *Journal of Nursing Education and Practice*, *4*(2), 116-122. http://dx.doi.org/10.5430/jnep.v4n2p116

Heaslip, V., & Scammell, J. (2012). Failing underperforming students: The role of grading in practice assessment. *Nurse Education in Practice, 12*(2), 95-100. http://dx.doi.org/10.1016/j.nepr.2011.08.003

Jones, D., & Paulhus, D. (2010). Different provocations trigger aggression in narcissists and psychopaths. *Social Psychological & Personality Science, 1*(1), 12-18. http://dx.doi.org/10.1177/1948550609347591

Konrath, S, Bushman, B. J., & Campbell, K. (2006). Attenuating the link between threatened egotism and aggression. *Psychological Science, 17*(11), 995-1001. Retrieved from http://dx.doi.org/10.1111/j.1467-9280.2006.01818.x

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. http://dx.doi.org/10.2307/2529310

Luhanga, F., Yonge, O., & Myrick, F. (2008). "Failure to assign failing grades": Issues with grading the unsafe student. *International Journal of Nursing Education Scholarship, 5*(1), 1-14. http://dx.doi.org/10.2202/1548-923X.1366

O'Connor, A. (2001). *Clinical instruction and evaluation: A teaching resource.* Sudbury, MA: Jones and Bartlett.

Randall, J., & Engelhard, G. (2010). Examining the grading practices of teachers. *Teaching and Teacher Education, 26*(7), 1372-1380. http://dx.doi.org/10.1016/j.tate.2010.03.008

Rojstaczer, S., & Healy, C. (2012). Where A is ordinary: The evolution of American colleges and university grading, 1940-2009. *Teachers College Record, 114*(7), 1-23.

Scanlan, J. M., Care, W. D, & Gessler, S. (2001). Dealing with the unsafe student in clinical practice. *Nursing Education, 26*(1), 23-27. http://dx.doi.org/10.1097/00006223-200101000-00013

Seurynck, K., Buch, C., Ferrari, M., & Murphy, S. (2014). Comparison of nurse mentor and instructor evaluation of clinical performance. *Nursing Education Perspectives, 35*(3), 195-196. http://dx.doi.org/10.5480/1536-5026-35.3.195

Stucke, T., & Sporer, S. (2002). When a grandiose self-image is threatened: Narcissism and self-concept clarity as predictors of negative emotions and aggression following ego-threat. *Journal of Personality, 70*(4), 509-532. http://dx.doi.org/10.1111/1467-6494.05015

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). The widget effect our national failure to acknowledge and act on differences in teacher effectiveness. *The Education Digest, 75*(2), 31-35.

Vaillancourt, T. (2013). Students aggress against professors in reaction to receiving poor grades: An effect moderated by student narcissism and self-esteem. *Aggressive Behavior, 39*(1), 71-84. http://dx.doi.org/10.1002/ab.21450

Appendix

Flight Scenario

Your private pilot student has just flown Lesson 27 in the PPL course. To this point the student has shown normal progress; they have had 2 repeats of lessons pre solo. Student is routinely prepared for lessons, shows up early and completes outside assignments on time.

For Lesson 27, you have already evaluated certain line items and only need to evaluate a few more. You may review the student's performance thus far on the next page. Then watch the video (once only) and evaluate the remaining line items.

Notes: All clearing turns and all pre-maneuver checks have been completed.

After providing scores for the remaining line items, please issue an overall grade for the lesson.

PTS standards are highlighted and already evaluated items are circled.

Line Item Group		**Line Item Description**		**Line Item Grade**								**Comment**
Pre L Req		Preparation for Pilotage to the practice area (Ev)		S	U	I	N A					
X01 Group Divider		Pre-Lesson assignment complete (Ev)		S	U	I	N A					
X111 Preflt Prep		Basic Weather Brief (Re)		S	U	I	N A					
6.5 ATC Airspace		Radio communication phraseology and techniques (Pr)		4	3	2	1	0	I	N A		
x113 TO/LD/GA		Short Field Takeoff/Climb (Ev)		4	3	2	1	0	I	N A		
x103 Clmb-Dsent		Climbing at Best Angle (Vx)(Pr)		4	3	2	1	0	I	N A		

x103 Clmb-Dsent		Climbing at Best Rate (Vy)(Pr)		4	3	2	1	0	I	NA		
Pr Pl Sht Crs		Departure/Arrival Procedures (Pr)		4	3	2	1	0	I	NA		
6.5 ATC Airspace		WMU Gate holding procedures (Pr)		4	3	2	1	0	I	NA		
PR Pl Sht Crs		GPS Usage (Pr)		4	3	2	1	0	I	NA		
x115 Navigation		Pilotage - Visual Navigation (Pr)		4	3	2	1	0	I	NA		
Pr Pl Sht Crs		**Slow Flight (Pr)**		**4**	**3**	**2**	**1**	**0**	**I**	**NA**		
6.7 Trn Manuever		**Power On stall (Pr)**		**4**	**3**	**2**	**1**	**0**	**I**	**NA**		
6.7 Trn Manuever		**Power Off stall (Pr)**		**4**	**3**	**2**	**1**	**0**	**I**	**NA**		
x116 Slw Fl/Stll		**Symptoms of the stall in the turn and recovery (Pr)**		**4**	**3**	**2**	**1**	**0**	**I**	**NA**		
x114a Perfrm Mn		**Steep Turns - Entry (Pr)**		**4**	**3**	**2**	**1**	**0**	**I**	**NA**		
x144a Perfrm Mn		**Steep Turns - Maintenance (Pr)**		**4**	**3**	**2**	**1**	**0**	**I**	**NA**		
x144a Perfrm Mn		**Steep Turns - Rolling Out (Pr)**		**4**	**3**	**2**	**1**	**0**	**I**	**NA**		
x118a Emer Ops		Spiral Diver Recovery (De)		D 2	D 1	D	0	I				

6.7 Trn Manuever		Turns around a point (Pr)		4	3	2	1	0	I	NA		
x118a Emer Ops		Emergency Approach and Landing - Engine Out (Simulated)(Pr)		4	3	2	1	0	I	NA		
x118a Emer Ops		Emergency Approach and Landing (Simulated) - with Power (Pr)		4	3	2	1	0	I	NA		
x112b Airprt Ops		Traffic Patterns - Normal (Pr)		4	3	2	1	0	I	NA		
Pr Pl Sht Crs		Short Field Approach and Landing (Pr)		4	3	2	1	0	I	NA		
x113 TO/LD/GA		Short Field flare and landing - performance application (Pr)		4	3	2	1	0	I	NA		
x600 Cirrus STD		Soft Field Landings (Pr)		4	3	2	1	0	I	NA		
x113 TO/LD/GA		Short Field - Control after landing (Pr)		4	3	2	1	0	I	NA		
				Overall Lesson Grade: _____								

Table 1

Initial Assessment Exercises – Cohen's Kappa Coefficients by Group

Group	Avg. Cohen's Kappa Coefficient
1 year or less experience as flight instructor (N = 12)	0.197
13 months to 2 years' experience as flight instructor (N = 19)	0.325
2 years' or more experience as a flight instructor (N = 13)	0.357
Full time flight faculty (N = 9)	0.340

Table 2

Initial Assessment Exercises – Frequency Data 1 Year or Less Experience – 1 = Low Performance; 4 = High Performance

Group	Description	Gold Standard Grade	Respondents' Mode Grade	Percentage of Agreement with Gold Standard
1 year or less experience as flight instructor (N = 12)	Lesson 27 – Slow Flight	2	2	67%
	Lesson 27 – Steep Turn Entry	4	2	0%
	Lesson 27 – Steep Turn Maintenance	3	2	34%
	Lesson 27 – Steep Turn Roll out	3	2	17%
	Lesson 27 – Power Off Stalls	1	1	50%
	Lesson 27 – Power On Stalls	1	1	67%
	Lesson 27 – Symptoms of Stalls	1	1	42%
	Lesson 27 – Overall Grade	E	C	8%
	Lesson 28 – Slow Flight	4	2	0%
	Lesson 28 – Steep Turns	3	2	34%
	Lesson 28 – Power Off Stall	4	2	0%
	Lesson 28 – Power On Stall	4	2	0%
	Lesson 28 – Overall Grade	E	C	20%
	Lesson 43 – Slow Flight	2	3	42%
	Lesson 43 – Steep Turns	3	3	92%
	Lesson 43 – Power Off Stall	1	2	17%
	Lesson 43 – Power On Stall	3	3	67%
	Lesson 43 – Overall Grade	E	C	17%
	Lesson 48 – Slow Flight	4	3	0%
	Lesson 48 – Steep Turns	2	2	75%
	Lesson 48 – Power Off Stall	2	2	55%
	Lesson 48 – Power On Stall	3	3	90%
	Lesson 48 – Overall Grade	E	B	20%

Table 3

Initial Assessment Exercises – Frequency Data – 13 Months to 2 Years' Experience – 1 = Low Performance; 4 = High Performance

Group	Description	Gold Standard Grade	Respondents' Mode Grade	Percentage of Agreement with Gold Standard
13 mths to 2 yrs' experience as flight instructor (N = 19)	Lesson 27 – Slow Flight	2	2	79%
	Lesson 27 – Steep Turns Entry	4	2	0%
	Lesson 27 – Steep Turns Maintenance	3	3	64%
	Lesson 27 – Steep Turns Roll Out	3	2	37%
	Lesson 27 – Power Off Stalls	1	1	69%
	Lesson 27 - Power On Stalls	1	1	56%
	Lesson 27 – Symptoms of Stall	1	1	50%
	Lesson 27 – Overall Grade	E	E	57%
	Lesson 28 – Slow Flight	4	2	0%
	Lesson 28 – Steep Turns	3	2	34%
	Lesson 28 – Power Off Stalls	4	2	0%
	Lesson 28 – Power On Stalls	4	2	0%
	Lesson 28 – Overall Grade	E	E	79%
	Lesson 43 – Slow Flight	2	2 and 3	50%
	Lesson 43 – Steep Turns	3	3	95%
	Lesson 43 – Power Off Stalls	1	2	23%
	Lesson 43 – Power On Stalls	3	2 and 3	50%
	Lesson 43 – Overall Grade	E	E	67%
	Lesson 48 – Slow Flight	4	3	6%
	Lesson 48 – Steep Turns	2	2	95%
	Lesson 48 – Power Off Stalls	2	2	56%
	Lesson 48 – Power On Stalls	3	3	82%
	Lesson 48 – Overall Grade	E	E	61%

Table 4

Initial Assessment Exercises – Frequency Data – 2+ Years' Experience – 1 = Low Performance; 4 = High Performance

Group	Description	Gold Standard Grade	Respondents' Mode Grade	Percentage of Agreement with Gold Standard
2+ yrs' experience as flight instructor (N = 13)	Lesson 27 – Slow Flight	2	2	77%
	Lesson 27 – Steep Turns Entry	4	2	0%
	Lesson 27 – Steep Turns Maintenance	3	2	47%
	Lesson 27 – Steep Turns Roll Out	3	2	39%
	Lesson 27 – Power Off Stalls	1	1	70%
	Lesson 27 - Power On Stalls	1	1	62%
	Lesson 27 – Symptoms of Stall	1	1	62%
	Lesson 27 – Overall Grade	E	E	61%
	Lesson 28 – Slow Flight	4	2	0%
	Lesson 28 – Steep Turns	3	2	39%
	Lesson 28 – Power Off Stalls	4	2	0%
	Lesson 28 – Power On Stalls	4	2	0%
	Lesson 28 – Overall Grade	E	E	70%
	Lesson 43 – Slow Flight	2	2	62%
	Lesson 43 – Steep Turns	3	3	100%
	Lesson 43 – Power Off Stalls	1	2	54%
	Lesson 43 – Power On Stalls	3	3	54%
	Lesson 43 – Overall Grade	E	E	62%
	Lesson 48 – Slow Flight	4	3	0%
	Lesson 48 – Steep Turns	2	2	85%
	Lesson 48 – Power Off Stalls	2	2	85%
	Lesson 48 – Power On Stalls	3	3	77%
	Lesson 48 – Overall Grade	E	E	69%

Table 5

Post-Test and Initial Assessment Exercises – Cohen's Kappa Coefficients by Group

Group	Avg. Cohen's Kappa Coefficient Initial Assessment	Avg. Cohen's Kappa Coefficient Post-test Assessment
1 year or less experience as flight instructor (N = 12)	0.197	.339
13 months to 2 years' experience as flight instructor (N = 19)	0.325	.270
2 years or more experience as a flight instructor (N = 13)	0.357	.344

Table 6

Post-Test Assessment Exercises – Frequency Data 1 Year or Less Experience – 1 = Low Performance; 4 = High Performance

Group	Description	Gold Standard Grade	Respondents' Mode Grade	Percentage of Agreement with Gold Standard
1 year or less experience as flight instructor (N = 12)	Lesson 27 – Slow Flight	2	2	92%
	Lesson 27 – Steep Turn Entry	4	2	8%
	Lesson 27 – Steep Turn Maintenance	3	3	69%
	Lesson 27 – Steep Turn Roll out	3	3	62%
	Lesson 27 – Power Off Stalls	1	1	69%
	Lesson 27 – Power On Stalls	1	1	62%
	Lesson 27 – Symptoms of Stalls	1	1	62%
	Lesson 27 – Overall Grade	E	C	46%
	Lesson 28 – Slow Flight	4	2	0%
	Lesson 28 – Steep Turns	3	3	54%
	Lesson 28 – Power Off Stall	4	2	0%
	Lesson 28 – Power On Stall	4	2	0%
	Lesson 28 – Overall Grade	E	E	54%
	Lesson 43 – Slow Flight	2	3	31%
	Lesson 43 – Steep Turns	3	3	85%
	Lesson 43 – Power Off Stall	1	2	23%
	Lesson 43 – Power On Stall	3	2	46%
	Lesson 43 – Overall Grade	E	E	50%
	Lesson 48 – Slow Flight	4	3	8%
	Lesson 48 – Steep Turns	2	2	92%
	Lesson 48 – Power Off Stall	2	2	62%
	Lesson 48 – Power On Stall	3	3	62%
	Lesson 48 – Overall Grade	E	E	50%

Table 7

Post-Test Assessment Exercises – Frequency Data 13 Months to 2 Years' Experience – 1 = Low Performance; 4 = High Performance

Group	Description	Gold Standard Grade	Respondents' Mode Grade	Percentage of Agreement with Gold Standard
13 mths to 2 yrs' experience as flight instructor (N = 19)	Lesson 27 – Slow Flight	2	2	67%
	Lesson 27 – Steep Turns Entry	4	3	0%
	Lesson 27 – Steep Turns Maintenance	3	3	56%
	Lesson 27 – Steep Turns Roll Out	3	2	44%
	Lesson 27 – Power Off Stalls	1	1	78%
	Lesson 27 - Power On Stalls	1	2	33%
	Lesson 27 – Symptoms of Stall	1	1	63%
	Lesson 27 – Overall Grade	E	D	13%
	Lesson 28 – Slow Flight	4	2	0%
	Lesson 28 – Steep Turns	3	2	44%
	Lesson 28 – Power Off Stalls	4	2	0%
	Lesson 28 – Power On Stalls	4	2	0%
	Lesson 28 – Overall Grade	E	C	25%
	Lesson 43 – Slow Flight	2	2	44%
	Lesson 43 – Steep Turns	3	3	100%
	Lesson 43 – Power Off Stalls	1	2	33%
	Lesson 43 – Power On Stalls	3	3	56%
	Lesson 43 – Overall Grade	E	C	38%
	Lesson 48 – Slow Flight	4	3	0%
	Lesson 48 – Steep Turns	2	2	89%
	Lesson 48 – Power Off Stalls	2	2	78%
	Lesson 48 – Power On Stalls	3	3	67%
	Lesson 48 – Overall Grade	E	D	25%

Table 8

Post-Test Assessment Exercises – Frequency Data 2+ Years' Experience – 1 = Low Performance; 4 = High Performance

Group	Description	Gold Standard Grade	Respondents' Mode Grade	Percentage of Agreement with Gold Standard
2+ yrs' experience as flight instructor (N = 13)	Lesson 27 – Slow Flight	2	2	59%
	Lesson 27 – Steep Turns Entry	4	2 and 3	12%
	Lesson 27 – Steep Turns Maintenance	3	3	59%
	Lesson 27 – Steep Turns Roll Out	3	3	69%
	Lesson 27 – Power Off Stalls	1	2	41%
	Lesson 27 - Power On Stalls	1	2	41%
	Lesson 27 – Symptoms of Stall	1	2	47%
	Lesson 27 – Overall Grade	E	E	47%
	Lesson 28 – Slow Flight	4	2	18%
	Lesson 28 – Steep Turns	3	3	60%
	Lesson 28 – Power Off Stalls	4	2	12%
	Lesson 28 – Power On Stalls	4	2	12%
	Lesson 28 – Overall Grade	E	E	81%
	Lesson 43 – Slow Flight	2	2	59%
	Lesson 43 – Steep Turns	3	3	81%
	Lesson 43 – Power Off Stalls	1	2	29%
	Lesson 43 – Power On Stalls	3	3	53%
	Lesson 43 – Overall Grade	E	E	65%
	Lesson 48 – Slow Flight	4	3	24%
	Lesson 48 – Steep Turns	2	2	88%
	Lesson 48 – Power Off Stalls	2	3	47%
	Lesson 48 – Power On Stalls	3	3	53%
	Lesson 48 – Overall Grade	E	E	65%