Inter-Rater Reliability of Grading Undergraduate Portfolios in Veterinary Medical Education

Robert P. Favier ■ Johannes C.M. Vernooij ■ F. Herman Jonker ■ Harold G. J. Bok

ABSTRACT

The reliability of high-stakes assessment of portfolios containing an aggregation of quantitative and qualitative data based on programmatic assessment is under debate, especially when multiple assessors are involved. In this study carried out at the Faculty of Veterinary Medicine, Utrecht University, the Netherlands, two independent assessors graded the portfolios of students in their second year of the 3-year clinical phase. The similarity of grades (i.e., equal grades) and the level of the grades were studied to estimate inter-rater reliability, taking into account the potential effects of the assessor's background (i.e., originating from a clinical or non-clinical department) and student's cohort group, gender, and chosen master track (Companion Animal Health, Equine Health, or Farm Animal/Public Health). Whereas the similarity between the two grades increased from 58% in the first year the grading system was introduced to around 80% afterwards, the grade level was lower over the next 3 years. The assessor's background had a minor effect on the proportion of similar grades, as well as on grading level. The assessor intraclass correlation was low (i.e., all assessors scored with a similar grading pattern [same range of grades]). The grades awarded to female students were higher but more often dissimilar. We conclude that the grading system was well implemented and has a high inter-rater reliability.

Key words: high-stakes assessment, portfolio, inter-rater reliability, veterinary

INTRODUCTION

Educating students to become competent professionals requires ongoing assessment of knowledge, skills, and performance.[1–3] A challenge in the promotion of sustainable assessment strategies and tools includes both enhancing learning and adequately assessing competency development, both formative, low-stake and summative, high-stake.[4] Epstein proposed that "various domains of competence should be assessed in an integrated, coherent, and longitudinal fashion with the use of multiple methods and provision of frequent and constructive feedback."[5(p.394)] Recently, van der Vleuten et al. described a theoretical model for programmatic assessment, built around learning, assessment, and supporting activities.[6] In this model, assessment and learning are combined by making each individual assessment maximally informative for learning. In the end, high-stakes assessment of learning for promotion or licensure is based on data from several individual assessments.[6,7] Besides facilitating students' learning, this model of programmatic assessment should improve the validity and reliability of assessments and the documentation of competence development. O'Brien et al. demonstrated that in undergraduate medical education, a portfolio combining low-stakes and high-stakes assessment is feasible and a worthwhile addition to an assessment system.[8] However, high-stakes assessment is subject to discussion, as it has been argued that it should be based on an adequate number of data points with high internal consistency and reliability.[7,9–11] Reliability is essential for assessment and test quality, and can be subdivided into internal consistency, stability, and inter-rater reliability.[12,13]

Portfolio assessments have been criticized because of the difficulty in reliably assessing them.[7,14] When introducing portfolios, Roberts et al. suggested that people should "take a research-based approach and publish data on validity, feasibility and effects on student learning."[10(p.900)] Roberts et al. demonstrated modest precision in assessing students' achievement by means of a portfolio as part of an assessment program of an integrated clinical placement.[15] Gadbury-Amyot et al. strongly suggested, based on their validity and reliability study in two dental schools using programmatic portfolio assessment, that two raters should independently evaluate each portfolio.[16] Royal and Hecker identified a list of rater errors and concluded that if criterion-referenced assessment of clinical performance of students is to be used, then there should be thorough and repeated training sessions for raters, and the inter-rater reliability should be measured routinely.[13] Inter-rater reliability, the amount of agreement between two or more raters, can be evaluated by statistics, such as Kappa, Pearson's rho, Spearman's rho, and intraclass correlation (ICC).[13] O'Brien et al. quantified the inter-rater reliability of the final evaluation of a portfolio in each of five competency domains. Using a three-grade system, they achieved at least 77% agreement

in each of the domains before reconciliation.[8] It has not yet been investigated whether factors such as student gender, year of cohort, clinical background of assessor, or student master track affect inter-rater reliability. The aims of this study were to assess the inter-rater reliability of portfolio assessment and to evaluate the effects of student gender, student's chosen master track (one of three: Companion Animal Health, Equine Health, or Farm Animal Health/Public Health), year of cohort, and the assessor's clinical background (clinical or non-clinical department) on inter-rater reliability (i.e., similarity between grades for a portfolio).

MATERIALS AND METHODS

Context

The Faculty of Veterinary Medicine of Utrecht University (FVMUU) offers a 6-year Doctor of Veterinary Medicine (DVM) program that consists of 3 years of pre-clinical education followed by 3 years of clinical rotations. The 3-year clinical phase comprises several 1- to 7-week clinical rotations in disciplines related to three different tracks: Equine Health (EH), Companion Animal Health (CAH), and Farm Animal Health/Public Health (FAH/PH). Students select one of these tracks and work alongside staff in the clinic, engaging in a variety of learning activities. Given the fact that the FVMUU program has a differentiated outcome (i.e., a common core track leads up to three different tracks), students gain experience in relation to this differentiated outcome. Formal teaching is aimed at promoting in-depth understanding of topics encountered during clinical work by means of a competency-based approach.[17]

Low-Stakes Learning and Assessment Program at FVMUU

Assessments during the 3-year clinical phase of the curriculum are based on a programmatic assessment approach.[6] In this model, a program of assessment, built around learning, assessment, and supporting activities, should improve the validity and reliability of the assessment and documentation of competence development, and should also enhance and facilitate students' learning. To support the development of skills in all seven competency domains (veterinary expertise, communication, collaboration, scholarship, health and welfare, personal development, and entrepreneurship),[17] students receive regular feedback about their skills and competence.[11] This feedback is stored in a personal electronic portfolio so that the student can refer to it, especially when formulating new learning objectives for an upcoming period (every 6 months). The portfolio systematically documents the feedback, feedforward, and self-reflection on the student as he/she gains competence in the seven competency domains.[11,17] The portfolio contains different feedback instruments (mini-CEX, 360-degree feedback [multisource feedback; MSF], and evidence-based case reports [EBCRs]). Additionally, the student's reflection on the learning process, progress, and learning objectives for the next period documented in the Personal Developing Plan are part of the portfolio. The portfolio therefore enables the student and mentor to monitor the learning process and to adjust if necessary.

High-Stakes Assessment Procedure

In addition to fostering students' development in a formative way as described above, the portfolio also serves as the basis for the high-stakes assessment of the student. The large amount of longitudinally collected feedback about the student's performance enables the Portfolio Evaluation Committee (PEC), consisting of senior faculty assessors, to establish a robust high-stakes judgment about that student's performance. There are two high-stakes assessment events during the 3-year clinical phase: at the end of years 2 and 3 (expected moment of graduation). This high-stakes portfolio assessment approach was introduced in October 2012 and is described below:

1. A student's portfolio is assigned by convenience to two members of the PEC, ensuring an adequate spread of portfolios between assessors.
2. Both assessors formulate qualitative feedback on the development of competency in each domain and scored the development on a 5-point Likert scale (Table 1). Standards/criteria for each domain are described for each year of the 3-year clinical phase. Scores on one domain cannot be used to compensate for scores on other domains.
3. The assessor awards the entire portfolio a final grade (on a 1–10 scale), as described by ten Cate et al.[18] (exceeds expectations [9–10]; meets expectations [6–8]; below expectations [4–5]) (Table 1).
4. If there is disagreement between assessors about the grade, a third assessor is asked to provide the final grade. In complex cases (i.e., where there is a large difference [≥ 2 points] or cases with remarkable content), the portfolio is discussed during a biweekly meeting of the PEC.
5. Students receive their final grade plus narrative feedback on each competency domain from the assessors.

Table 1: Required overall competency scores (1–5)* to pass (threshold) related to the number of years in the 3-year clinical phase

	Competency score, end of year 1	Competency score, end of year 2	Competency score, end of year 3
Exceeds expectations (9–10)	≥ 3	≥ 4	5
Meets expectations (6–8)	2	3	4
Below expectations (4–5)	1	≤ 2	≤ 3

* See ten Cate J, ter Braak EWMT, Frenkel J, et al. The 4-to-10 expected level scale (410VN-schaal) for personal evaluations [De 4-tot-10 verwacht niveau-schaal (410VN-schaal) bij persoonlijke beoordelingen]. Tijdschrift voor medisch onderwijs. 2006;25(4):157–63

Ethical Considerations

The Ethical Review Board of the Netherlands Association for Medical Education (NVMO-ERB) approved this study (NVMO-ERB-nr: 891).

Study Population

This retrospective cohort study included 574 portfolios assessed at the end of year 2, coming from four cohorts: (1) October 2012–September 2013, (2) October 2013–September 2014, (3) October 2014–September 2015, and (4) October 2015. Data collected from the portfolios included the grades of both assessors (i.e., before possible discussion in the PEC or involving a third assessor), assessors' IDs, student's gender, track (EH, CAH, FAH/PH), background of each assessor (originating from a clinical or non-clinical department), and cohort group.

Statistics

A multivariable mixed effect logistic regression model (lme4 package by Bates et al.[19]) was used to analyze the difference (*yes/no*) between the grades awarded by the two assessors, with assessor as the random effect, to take the grading of multiple portfolios by same assessor into account. Cohort group, student's gender, student's chosen track, and assessor's background (originating from a clinical or non-clinical department) were used as explanatory variables. As each portfolio was graded by two assessors, the given grades are dependent (e.g., both grades will be low for a poor portfolio and high for a rich one). To handle this dependence, we constructed a new data set by randomly selecting one of the two assessors and corresponding difference between grades per portfolio and analyzing the data as above. In this case, the model result depended on only one random data set. To accommodate this, we constructed 1,000 data sets (more would not further improve the precision of the estimates) for each randomly picked assessor (with difference [*yes/no*] in grading result) per available portfolio (Figure 1). Each assessor should have been used in each of the data sets at least five times; otherwise the portfolios of this assessor were excluded from the analysis of the specific data set. The above-described statistical model was applied to each of the data sets, resulting in the same number of results. Estimated parameter values and model convergence values were calculated and stored for further processing. To be accepted as valid, random effect estimates (*SD*) for each of the 1,000 model results should be larger than 0.0001, and the maximum gradient of the Hessian matrix should be smaller than 0.001.

To calculate the ICC, the Zeger et al.'s[20] formula was used:

$$\tau^2 / (\tau^2 + (15/16)^2 * \pi^2/3)$$

where τ^2 is the estimated random effect variance.

The estimates obtained are expressed as median values and 95% percentile intervals (95% PI). The summary statistics for the parameter estimates for each cohort by grade, gender, master track, and assessor's clinical background were back-transformed by taking the antilog (e^{β}).

Second, a proportional odds model (ordinal package by Christensen[21]) was applied, using the portfolio grade as outcome. The grades were grouped (≤ 5, 6, 7, 8, and ≥ 9) for estimation reasons. The model approach was similar to the logistic regression model, with random selection of the grade with associated assessor per portfolio. The validity of the models was assessed by the random variance estimate for assessor (> 0.0001) and by the conditional Hessian matrix (< 10,000). All models were applied in R version 3.3.0 (2016.05-03).[22]

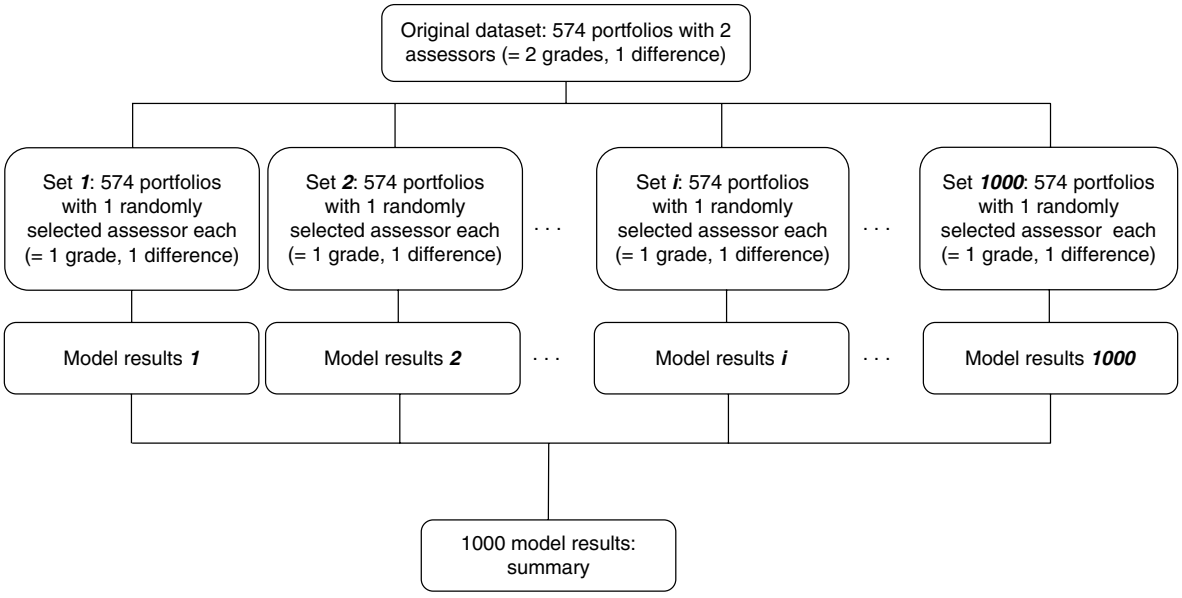

Figure 1: Schematic representation of the modeling procedure (difference = grade1 − grade2)

RESULTS

In total, 574 portfolios were graded by two independent assessors. Table 2 shows how the grades differed for each factor investigated. Difference between grades was more often observed (42%) in the first year that the grading system was used (cohort group 1) than in later cohorts (5%–22%). Grades differed less often for the FAH/PH track (16%) than for the two other tracks (24%–25%).

Table 3 shows the association between the two grades. The proportion of portfolios with identical grades was 0.782 (95% CI = 0.746–0.815). The proportion of portfolios with grades that differed by 1, 2, or 3 points was 0.20, 0.016, and 0.002, respectively. Most portfolios were awarded a grade of 7 or 8. Individual assessors evaluated between 12 and 144 portfolios, and the proportion of portfolios with identical grades varied between 0.38 and 0.97 for the different assessors.

A summary of the parameter estimates is presented in Table 4. Of a total of 1,000 random data sets, 837 models were considered valid. The estimate for correlation of grade difference of different assessors (ICC) was 3.8% (95% PI = 0.3–11.3), meaning the variability (distribution) of grade difference was very similar between assessors. Grade difference of portfolios was less present in cohorts 2–4 than in the first cohort (OR = 0.39, 0.07, and 0.28, respectively), in which the grading system was introduced. Moreover, differences between grades was more often observed for female students than male students. Grade differences for FAH/PH-track portfolios were observed less often (OR = 0.60; 95% PI = 0.57–0.65) than for CAH-track portfolios. In contrast, grade differences for EH-track portfolios were more likely than for CAH-track portfolios (OR = 1.31; 95% PI = 1.24–1.44). Assessors from non-clinical departments more often showed a grade difference compared with

Table 3: Frequency of the grades (highest vs. lowest) awarded for each portfolio

	Lowest grade						
Highest grade	4	5	6	7	8	9	10
4	3	–	–	–	–	–	–
5	0	94	–	–	–	–	–
6	0	15	36	–	–	–	–
7	0	2	29	156	–	–	–
8	0	0	3	57	133	–	–
9	0	0	1	4	12	27	–
10	0	0	0	0	0	2	0

Table 2: Cross tables with number and difference (*yes*/*no*) between the grades awarded by two assessors for 574 portfolios

	Difference between both grades				
	No		Yes		
Variable	Frequency	Proportion	Frequency	Proportion	Total
Cohort group					
(1) Oct. 2012–Sept. 2013	170	0.58	122	0.42	292
(2) Oct. 2013–Sept. 2014	296	0.78	84	0.22	380
(3) Oct. 2014–Sept. 2015	310	0.95	18	0.05	328
(4) Oct. 2015	122	0.82	26	0.18	148
Student gender					
Male	168	0.81	40	0.19	208
Female	730	0.78	210	0.22	940
Track					
CAH	522	0.77	160	0.24	682
FAH/PH	254	0.84	50	0.16	304
EH	122	0.75	40	0.25	162
Clinical department*					
Yes	573	0.80	146	0.20	719
No	325	0.76	104	0.24	429

CAH = Companion Animal Health; FAH/PH = Farm Animal Health/Public Health; EH = Equine Health

Note: Each portfolio was graded by two assessors, resulting in two differences (A-B and B-A) and, in total, 1,148 differences.

* Background of each assessor (clinical or non-clinical department)

Table 4: Summary of the parameter estimates (ICC, baseline odds, and odds ratios) of 837 valid results (from 1,000) of a logistic regression mixed model to analyze a difference (*yes*/*no*) between portfolio grades

Parameter	Estimate*	95%LL†	95%UL†
Intraclass coefficient (%)	3.8	0.3	11.3
Cohort group			
(1) Oct. 2012–Sept. 2013 (Ref)	1.00	–	–
(2) Oct. 2013–Sept. 2014	0.39	0.36	0.42
(3) Oct. 2014–Sept. 2015	0.07	0.06	0.08
(4) Oct. 2015	0.28	0.25	0.30
Student gender			
Male (Ref)	1.00	–	–
Female	1.16	1.09	1.24
Track			
CAH (Ref)	1.00	–	–
FAH/PH	0.60	0.57	0.65
EH	1.31	1.24	1.44
Clinical department‡			
Yes (Ref)	1.00	–	–
No	1.20	0.84	1.73

ICC = intraclass correlation; LL = lower limit; UL = upper limit; Ref = reference class; CAH = Companion Animal Health; FAH/PH = Farm Animal Health/Public Health; EH = Equine Health

* Median value of estimates

† 95% percentile interval of estimates

‡ Background of each assessor (clinical or non-clinical department)

assessors from clinical departments (OR = 1.20; 95% PI = 0.84–1.73).

The distribution of grades awarded is presented in Table 5. Most grades given were 7 and 8 (65%). The variation in grades awarded in the period 2014–2015 and October 2015 was greater than the variation in the first two years of grading. The portfolios of female students were given higher grades than the portfolios of male students, and FAH/PH-track portfolios had higher grades than CAH- or EH-track portfolios. The distribution of grades was independent of the background of the assessor.

Analysis of the random data sets revealed that the odds of a higher grade were lower in the three later cohorts than in the first cohort (mean grades: 7.2, 6.9, 6.9, and 6.6 in cohorts 1–4, respectively) (Table 6). The odds that the portfolios of female students would have a higher grade were higher than the odds for the portfolios of male students to have a higher grade (mean grade: females = 7.0; males = 6.7).

Table 5: Frequencies (*n*) and proportion of both grades for each portfolio per studied factor

	Grade*						
Overall	4	5	6	7	8	9	10
Frequency (*n*)	6	205	120	404	338	73	2
	0.005	*0.18*	*0.11*	*0.35*	*0.29*	*0.06*	*0.002*
Cohort group (*n*)							
(1) Oct. 2012–Sept. 2013 (ref)	0	6	51	136	82	17	0
	0.000	*0.02*	*0.18*	*0.47*	*0.28*	*0.06*	*0.000*
(2) Oct. 2013–Sept. 2014	4	48	57	146	109	16	0
	0.011	*0.13*	*0.15*	*0.38*	*0.29*	*0.04*	*0.000*
(3) Oct. 2014–Sept. 2015	2	91	11	83	110	30	1
	0.006	*0.28*	*0.03*	*0.25*	*0.34*	*0.09*	*0.003*
(4) Oct.2015	0	60	1	39	37	10	1
	0.006	*0.28*	*0.03*	*0.25*	*0.34*	*0.09*	*0.003*
Student gender (*n*)							
Male	0	62	18	59	58	11	0
	0.000	*0.30*	*0.09*	*0.28*	*0.28*	*0.05*	*0.000*
Female	6	143	102	345	280	62	2
	0.006	*0.15*	*0.11*	*0.37*	*0.30*	*0.07*	*0.002*
Track (*n*)							
CAH	2	140	65	253	183	38	1
	0.003	*0.21*	*0.10*	*0.37*	*0.27*	*0.06*	*0.001*
FAH/PH	2	35	38	93	116	20	0
	0.007	*0.12*	*0.13*	*0.31*	*0.38*	*0.07*	*0.000*
EH	2	30	17	58	39	15	1
	0.012	*0.19*	*0.11*	*0.36*	*0.24*	*0.09*	*0.006*
Clinical department (*n*)†							
Yes	4	125	81	247	216	45	1
	0.006	*0.17*	*0.11*	*0.34*	*0.30*	*0.06*	*0.001*
No	2	80	39	157	122	28	1
	0.005	*0.19*	*0.09*	*0.37*	*0.28*	*0.07*	*0.002*

CAH = Companion Animal Health; FAH/PH = Farm Animal Health/Public Health; EH = Equine Health

Notes: In total, 574 portfolios were graded.
Italic numbers indicate proportions.

* Each portfolio was graded by two assessors resulting in two grades and, in total, 1,148 grades

† Background of each assessor (clinical or non-clinical department)

FAH/PH-track portfolios were more likely to have higher grades than CAH-track portfolios; EH-track portfolios had similar grades to those of the CAH-track portfolios (mean grade: FAH/PH = 7.1; EH = 6.93; CAH = 6.87). The background of the assessors did not influence grading.

Table 6: Summary of parameter estimates (ICC, baseline odds, and odds ratios) of 820 valid results (from 1,000 runs) of a proportional odds mixed model to analyze portfolio grades

Parameter	Estimate*	95%LL†	95%UL†
Intraclass correlation (%)	1.1	0.2	2
Cohort group			
(1) Oct. 2012–Sept. 2013 (Ref)	1.00	–	–
(2) Oct. 2013–Sept. 2014	0.70	0.53	0.91
(3) Oct. 2014–Sept. 2015	0.85	0.55	1.34
(4) Oct. 2015	0.42	0.22	0.79
Student gender			
Male (Ref)	1.00	–	–
Female	1.44	0.99	2.09
Track			
CAH (Ref)	1.00	–	–
FAH/PH	1.41	1.08	1.87
EH	0.88	0.56	1.40
Clinical department‡			
Yes (Ref)	1.00	–	–
No	1.05	0.81	1.38

ICC = intraclass correlation; LL = lower limit; UL = upper limit; Ref = reference class; CAH = Companion Animal Health; FAH/PH = Farm Animal Health/Public Health; EH = Equine Health

Note: Each portfolio was graded by two assessors.

* Median value of estimates

† 95% percentile interval of estimates

‡ Background of each assessor (clinical or non-clinical department)

DISCUSSION AND CONCLUSIONS

In this retrospective cohort study, we analyzed the inter-rater reliability of the high-stakes assessment of undergraduate students' portfolios at the end of the second year of clinical rotations, as well as factors that could influence inter-rater reliability. We demonstrated the following. (1) The grading of portfolios during the past 3 years showed an inter-rater reliability of around 80%, following a moderate inter-rater reliability (58%) during the first year that the grading system was implemented. (2) The grades awarded to the portfolios of female students more often showed a difference, and (3) awarded grades to the portfolios of students of the EH-track tended to differ between assessors, whereas (4) there was less often a difference between the grades of the portfolios of FAH/PH-track students. (5) The background of the assessor (non-clinical/clinical department) did not affect inter-rater reliability. Moreover, (6) the portfolios of female students were awarded higher grades than those of male students, and (7) FAH/PH-track portfolios had higher grades than CAH- and EH-track portfolios. (8) The grades of awarded portfolios in the first year the system was used were higher than those of awarded portfolios in later years. Because only 1.2% (7/574 portfolios) of all students appealed against their awarded grade of *meets expectations* (1 graded 6; 6 graded 7) and no students appealed against a *below expectations* grade, we can conclude that students accepted the high-stakes portfolio assessment approach.

The grading system was more reliable after the first year of its introduction, possibly because difficult portfolios (i.e., those with large difference in grades awarded [≥ 2 points] or those missing content) were discussed in the PEC. This discussion facilitated uniform use of the rating method and criteria. These aspects are now highlighted in the training given to new assessors. Another explanation for the increased reliability is that after 1.5 years of grading portfolios, the PEC decided that the two assessors should discuss their grades before awarding the final grade (as opposed to after they had awarded their grades). This resulted in better agreement between the grades awarded. However, this policy was reversed when it came to the grading of the portfolios of students in the fourth cohort, resulting in slightly lower agreement between the grades awarded by the two assessors. The PEC opts for no or a small difference (≤ 1) between the grades of both assessors but understands that subjectivity will always play a role in grading, even when criteria are uniformly used, but should rarely result in larger differences. The inter-rater reliability was comparable to that reported by O'Brien et al.[8] To maximize the reliability of the assessments of the portfolio reviewers, O'Brien et al. used a procedure by which two reviewers independently scored each portfolio with a discussion afterwards, and if no consensus could be reached, a third reviewer became involved.[8] This resulted in an agreement of at least 77% before reconciliation and 98% after reconciliation—comparable to our results. However, we decided to be more careful about reconciliation and chose to opt more often for a third reviewer when both reviewers came up with different grades.

Over time, the assessors awarded lower grades, possibly because they developed judgment expertise. During the period studied, there were 18 assessors from eight different departments, and they were assigned portfolios regardless of the study track of the student. Of these 18 assessors, 13 came from one of the three clinical departments (EH, CAH, FAH/PH); however, the estimated ICC in both models was low, meaning hardly any clustering was present when observing a difference between grades or in the level of grade per assessor.

The grades awarded to FAH/PH-track portfolios were more uniform and higher than those awarded to EH- and CAH-track portfolios. The reason for the higher grades is not clear; the minimal requirements for the portfolios for the three tracks are identical, and it is unlikely that the amount of information in the portfolio explains this difference. The department from which the assessor originated is also not likely to affect the grade level, as 18 assessors from different departments were involved and portfolios were assigned at random. As only five assessors originated from non-clinical departments, and they assessed fewer portfolios,

their estimations might have been less precise and biased. However, all assessors assessed portfolios from all three tracks, and the clustering effect (ICC) for assessors was small. Nevertheless, the explanation for differences in grading between tracks might be due to student characteristics. Most students of the FAH track were selected at the start of the Bachelor of Veterinary Science program on the basis of an interview rather than a lottery, which is the case for the other students. The literature, mainly focusing on postgraduate programs, is conflicted regarding the predictive value of application interviews on performance.[23–26] Some indications show that students of the FAH/PH track were supervised more closely than students of the other tracks, possibly resulting in higher grades for their portfolios. The possible explanations warrant further research.

Only 104 of the students were males (18%, about 25 per cohort), which might have caused bias in the portfolios of females being more often graded differently. Voyer and Voyer, as we did, found that females had higher grades than males.[27] This gender difference was found to have a stable advantage in school marks in all courses but was largest in language courses and smallest in math courses. The gender composition, with more females than males, was also in favor of higher grade levels for females. Voyer and Voyer speculate about possible causes for this gender difference—for example, parents' involvement, stereotype threat, differences in learning styles, and biological influences—but emphasize the complexity of this issue and state that much research is still required.[27] In our situation, additional investigation regarding this finding is necessary.

Limitations

Data analyses were not straightforward due to the correlation of both grades within a portfolio and a possible correlation between awarded grades within assessors (i.e., assessors might tend to have their own level and range of awarded grades). The high number of portfolios with identical grades, especially in the most recent cohorts, the low number of male students, and the limited number of assessors from non-clinical departments make it difficult to estimate the effect of gender and clinical background because of the lack of variability. Nevertheless, the high proportion of identical grades, combined with the fact that only 1.2% of the grades were appealed, suggests that the grading system works effectively.

The content of the courses, the feedback by fellow students (via mini-CEX and MSF), the low- and high-stakes assessments, the teachers involved, and the portfolio requirements are all unique to FVMUU, and this should be taken into account when extrapolating the results of this study to another context.

Future Research

Recently, it was demonstrated that the portfolios of intrinsically motivated students have more content (more low-stake assessments than required, e.g., mini-CEX) than portfolios of more extrinsically motivated students.[28] This might increase the likelihood that the portfolios are awarded the same grades by different assessors. That portfolio content might be an indicator of motivation and a possible explanation for identical grading requires further investigation.

REFERENCES

1 Lurie SJ, Mooney CJ, Lyness JM. Measurement of the general competencies of the accreditation council for graduate medical education: a systematic review. Acad Med. 2009;84(3):301–9. https://doi.org/10.1097/ACM.0b013e3181971f08. Medline:19240434

2 van der Vleuten CP, Schuwirth LW. Assessing professional competence: from methods to programmes. Med Educ. 2005;39(3):309–17. https://doi.org/10.1111/j.1365-2929.2005.02094.x. Medline:15733167

3 Whitehead CR, Kuper A, Hodges B, et al. Conceptual and practical challenges in the assessment of physician competencies. Med Teach. 2015;37(3):245–51. https://doi.org/10.3109/0142159X.2014.993599. Medline:25523113

4 Boud D, Falchikov N. Aligning assessment with long-term learning. Assess Eval High Educ. 2006;31(4):399–413. https://doi.org/10.1080/02602930600679050.

5 Epstein RM. Assessment in medical education. N Engl J Med. 2007;356(4):387–96. https://doi.org/10.1056/NEJMra054784. Medline:17251535

6 van der Vleuten CP, Schuwirth LW, Driessen EW, et al. A model for programmatic assessment fit for purpose. Med Teach. 2012;34(3):205–14. https://doi.org/10.3109/0142159X.2012.652239. Medline:22364452

7 Driessen E. Do portfolios have a future? Adv Health Sci Educ Theory Pract. 2017;22(1):221–8. https://doi.org/10.1007/s10459-016-9679-4. Medline:27025510

8 O'Brien CL, Sanguino SM, Thomas JX, et al. Feasibility and outcomes of implementing a portfolio assessment system alongside a traditional grading system. Acad Med. 2016;91(11):1554–60. https://doi.org/10.1097/ACM.0000000000001168. Medline:27028027

9 Hecker KG, Norris J, Coe JB. Workplace-based assessment in a primary-care setting. J Vet Med Educ. 2012;39(3):229–40. https://doi.org/10.3138/jvme.0612.054R. Medline:22951458

10 Roberts C, Newble DI, O'Rourke AJ. Portfolio-based assessments in medical education: are they valid and reliable for summative purposes? Med Educ. 2002;36(10):899–900. https://doi.org/10.1046/j.1365-2923.2002.01288.x. Medline:12390455

11 Bok HG, Teunissen PW, Favier RP, et al. Programmatic assessment of competency-based workplace learning: when theory meets practice. BMC Med Educ. 2013;13:123. https://doi.org/10.1186/1472-6920-13-123. Medline:24020944

12 Royal KD, Hecker KG. Understanding reliability: a review for veterinary educators. J Vet Med Educ. 2016;43(1):1–4. https://doi.org/10.3138/jvme.0315-030R. Medline:26560547

13 Royal KD, Hecker KG. Rater errors in clinical performance assessments. J Vet Med Educ. 2016;43(1):5–8. https://doi.org/10.3138/jvme.0715-112R. Medline:26560550

14 McGill DA, van der Vleuten CP, Clarke MJ. Construct validation of judgement-based assessments of medical trainees' competency in the workplace using a "Kanesian" approach to validation. BMC Med Educ. 2015;15:237. https://doi.org/10.1186/s12909-015-0520-1. Medline:26715145

15 Roberts C, Shadbolt N, Clark T, et al. The reliability and validity of a portfolio designed as a programmatic assessment of performance in an integrated clinical placement. BMC Med Educ. 2014;14(1):197. https://doi.org/10.1186/1472-6920-14-197. Medline:25240385
16 Gadbury-Amyot CC, McCracken MS, Woldt JL, et al. Validity and reliability of portfolio assessment of student competence in two dental school populations: a four-year study. J Dent Educ. 2014;78(5):657–67. Medline:24789826
17 Bok HG, Jaarsma DA, Teunissen PW, et al. Development and validation of a competency framework for veterinarians. J Vet Med Educ. 2011;38(3):262–9. https://doi.org/10.3138/jvme.38.3.262. Medline:22023978
18 ten Cate J, ter Braak EWMT, Frenkel J, et al. The 4-to-10 expected level scale (410VN-schaal) for personal evaluations [De 4-tot-10 verwacht niveau-schaal (410VN-schaal) bij persoonlijke beoordelingen]. Tijdschrift voor medisch onderwijs. 2006;25(4):157–63. https://doi.org/10.1007/BF03056737.
19 Bates D, Mächler M, Bolker B, et al. Fitting linear mixed-effects models Using lme4. J Stat Softw. 2015;67(1):1–48. https://doi.org/10.18637/jss.v067.i01.
20 Zeger SL, Liang KY, Albert PS. Models for longitudinal data: a generalized estimating equation approach. Biometrics. 1988;44(4):1049–60. https://doi.org/10.2307/2531734. Medline:3233245
21 Christensen RHB. Ordinal—regression models for ordinal data. R package version 2015.6-28 [Internet]. Available from: https://cran.r-project.org/package=ordinal.
22 R Core Team. The R project for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2016 [cited 2019 Jan 19]. Available from: https://www.R-project.org/.
23 Timer JE, Clauson MI. The use of selective admissions tools to predict students' success in an advanced standing baccalaureate nursing program. Nurse Educ Today. 2011;31(6):601–6. https://doi.org/10.1016/j.nedt.2010.10.015. Medline:21056921
24 Dubovsky SL, Gendel MH, Dubovsky AN, et al. Can admissions interviews predict performance in residency? Acad Psychiatry. 2008;32(6):498–503. https://doi.org/10.1176/appi.ap.32.6.498. Medline:19190295
25 Handelman SL, Iranpour B, Brunette PM, et al. Evaluation of common predictors for selection of postdoctoral dental students. J Dent Educ. 1983;47(3):155–9. Medline:6572207
26 Chen F, Arora H, Martinelli SM, et al. The predictive value of pre-recruitment achievement on resident performance in anesthesiology. J Clin Anesth. 2017;39:139–44. https://doi.org/10.1016/j.jclinane.2017.03.052. Medline:28494890
27 Voyer D, Voyer SD. Gender differences in scholastic achievement: a meta-analysis. Psychol Bull. 2014;140(4):1174–204. https://doi.org/10.1037/a0036620. Medline:24773502
28 de Jong LH, Favier RP, van der Vleuten CPM, et al. Students' motivation toward feedback-seeking in the clinical workplace. Med Teach. 2017;39(9):954–8. Medline:28521573

AUTHOR INFORMATION

Robert P. Favier, DVM, PhD, is Assistant Professor, Department of Clinical Sciences of Companion Animals, Faculty of Veterinary Medicine, Utrecht University, 3508 TC Utrecht, the Netherlands; and is currently working as the Program Director of the Evidensia Academy Netherlands. Email: robertfavier14101973@gmail.com. His interests are in work-place based assessment and programmatic assessment.

Johannes C.M. Vernooij, MSc, (https://orcid.org/0000-0002-2646-9216) is Assistant Professor, Biostatistician, and Teacher in Methodology and Statistics, Department of Farm Animal Health, Faculty of Veterinary Medicine, Utrecht University, 3508 TC Utrecht, the Netherlands. His interest is in reliability of measurements.

F. Herman Jonker, DVM, PhD, is Assistant Professor, Chair of the Portfolio Evaluation Committee, and Teacher in Reproduction, Department of Farm Animal Health, Faculty of Veterinary Medicine, 3508 TC Utrecht University, Utrecht, the Netherlands.

Harold G.J. Bok, DVM, PhD, (https://orcid.org/0000-0002-6435-0240) is Assistant Professor, Centre for Quality Improvement in Veterinary Education, Faculty of Veterinary Medicine, 3508 TC Utrecht University, Utrecht, the Netherlands. His interests are in work-based learning and assessment, programmatic assessment, feedback, and expertise development.

Robert P. Favier and Johannes C.M. Vernooij contributed equally to this article.