

## Amsterdam University of Applied Sciences

### Applicability of progress testing in veterinary medical education

Favier, Robert P.; van der Vleuten, Cees P. M.; Ramaekers, Stephan P. J.

**DOI**

[10.3138/jvme.0116-008R](https://doi.org/10.3138/jvme.0116-008R)

**Publication date**

2017

**Document Version**

Proof

**Published in**

Journal of Veterinary Medical Education

[Link to publication](#)

**Citation for published version (APA):**

Favier, R. P., van der Vleuten, C. P. M., & Ramaekers, S. P. J. (2017). Applicability of progress testing in veterinary medical education. *Journal of Veterinary Medical Education*, 44(2), 351-357. <https://doi.org/10.3138/jvme.0116-008R>

**General rights**

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

**Disclaimer/Complaints regulations**

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please contact the library: <https://www.amsterdamuas.com/library/contact/questions>, or send a letter to: University Library (Library of the University of Amsterdam and Amsterdam University of Applied Sciences), Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

# Applicability of Progress Testing in Veterinary Medical Education

Robert P. Favier ■ Cees P.M. van der Vleuten ■ Stephan P.J. Ramaekers

## ABSTRACT

A substantial part of graduate education in veterinary medicine is spent in clinical practice. During the clinical experiential phase, it is difficult to monitor students' actual knowledge development: they build individual records of experiences based on the cases they have to deal with, while mainly focusing on knowledge that is of direct, clinical relevance to them. As a result, students' knowledge bases may differ to such a degree that a single test alone may not be able to provide an adequate reflection of progress made. In these circumstances, progress testing, which is a method of longitudinal assessment independent of the curricular structure, may offer a viable solution. The purpose of this study, therefore, was to determine the extent to which progress tests (PT) can be used to monitor progress in knowledge development at a graduate level in veterinary medical education. With a 6-month interval, we administered two tests to students based on the Maastricht Progress Test format that covered a large variety of veterinary topics. Consequently, we analyzed students' progress in knowledge development. Based on a substantive appraisal of the questions and analysis of the test results, we concluded that the tests met the measurement criteria. They appeared sensitive enough to gauge the progress made and were appreciated by the students. Hence, in spite of the differences within the whole graduate group, the PT format can be used to monitor students' knowledge development.

**Key words:** progress test, veterinary, medical curriculum, validity, reliability, generalizability

## INTRODUCTION

Since the last decade, there has been increasing awareness of the problems students confront in the transition from pre-clinical study to learning and working in a clinical setting.<sup>1,2</sup> To ease this transition from theory to practice, educationalists have advocated increasing the practical components in the pre-clinical program and raising the theoretical components in the clinical phase.<sup>3,4</sup> As a result, the traditional division of medical curricula into a pre-clinical and a clinical phase is increasingly being substituted by a more vertically integrated program that embraces a gradual transition from theory to practice.

In highly integrated courses (clinical or otherwise), however, it is difficult to monitor the development of students' knowledge base. So-called progress tests (PTs) have the advantage of being less dependent on the timing of particular course content, and allow for long-term monitoring of knowledge development.<sup>5</sup>

Progress testing is a longitudinal way of assessing the growth in functional medical knowledge for each student.<sup>5-8</sup> Essentially, it is a repeated assessment based on samples of the knowledge domains that students are expected to have mastered by the time they graduate and enter the veterinary profession. Such a longitudinal assessment approach not only promotes the reliability of test results, it also positively affects student learning behavior, discourages "binge learning," and results in deep learning.<sup>9,10</sup> Moreover, a PT is independent of local


curricula and can be used in a multi-center collaboration.<sup>11</sup> Theoretically, the use of progress testing draws upon constructivist learning theories and notions of self-regulated learning.<sup>12,13</sup> PTs promote the development of long-term functional knowledge, as students have to maintain their knowledge base during the whole course and show cumulative growth.

In spite of over 30 years of experience with progress testing in dental and medical curricula,<sup>5,6,14</sup> the use of PTs in a multi-species curriculum has only started recently.<sup>15</sup> A major difficulty might be that veterinary curricula need to cover many different species. Furthermore, it might be technically demanding to organize PTs within and between differentiated tracks during clinical rotations, and the emphasis on knowledge reproduction could arguably be inconsistent with the educational program's focus on the development of competence in handling real clinical cases.

This study addresses the applicability of PTs in a master's program in veterinary medicine with differentiated outcomes. In the Netherlands, the veterinary entry program is a master's program, equivalent to the education received to earn the DVM/VMD degree in North America. This study is guided by the following three research questions:

1. To what extent can PTs measure progress in knowledge development across time? Are they sensitive enough to capture progress made over a 6-month

**Table 1:** Overview of the master's programs in farm-animal health (FAH), companion-animal health (CAH), and equine health (EH)

		Program	Tracks and number of weeks			
			FAH	CAH	EH	
	<b>Master's year 1</b>	<b>Uniform part (major uniform)</b>				
			Hygiene/microbiological/pathological diagnostics	3	3	3
	Electives		Management and the veterinarian's societal responsibility	4	4	4
			Responsible use of experimental animals	1	1	1
			Research project	12	12	12
			Free academic electives	10	10	10
	<b>Major differentiated</b>		Intramural clerkships and theoretical education	53	50	50
			Extramural clerkships	8	8	8
			Basic rotations (uniform for all students)	12	15	17
	<b>Master's year 3</b>	Electives	Track (clinics, research, or management)	17	17	15
<b>Total</b>			<b>120</b>	<b>120</b>	<b>120</b>	

Adapted and reprinted from the Self-Study Report 2014, Faculty of Veterinary Medicine, Utrecht University (with permission of Faculty of Veterinary Medicine, Utrecht University)

period? To be able to answer this question, we also need to determine the PT's quality (research question 2).

- To what extent do these PTs meet validity and reliability requirements?
- Do students perceive progress testing as beneficial to their development in this phase of their training?

## METHODS

To find answers to these research questions, we developed a PT with two runs and different items, to be administered at a 6-month interval to all students in the master's program.

### Context

In 2007, the Dutch Faculty of Veterinary Medicine at Utrecht University (FVMU) carried through a major curriculum reform. The intended curricular changes included a more gradual change in emphasis from theoretical (pre-clinical) to practical (clinical) education. In line with this, assessments were redesigned and a pilot PT was introduced into the 3-year clinically oriented master's program of the 6-year curriculum (2007 Curriculum, or C2007).<sup>16,17</sup> Launched in September 2010, the new master's program comprised a number of 1- to 7-week clinical rotations in disciplines related to three master's tracks: equine health, companion-animal health, and farm-animal health. Students select one of these tracks and work alongside staff in the clinic, where they engage in a variety of learning activities (Table 1). Formal teaching is aimed at promoting in-depth understanding of topics encountered during clinical work by means of a competence-based approach.<sup>17</sup> Given the fact that the FVMU program has a differentiated outcome (a common core track leading to three different master's tracks), students should gain experience in relation to this differentiated outcome.

## Test Development

Development of the PTs included the following steps:

- Based on data concerning the various topics, subjects, and veterinary problems that are covered in both the bachelor's and master's programs, we developed a two-dimensional blueprint (topics by disciplines) to achieve a representative sample of items.
- A group of experienced clinicians/teachers constructed or selected test items that suited the blueprint. All proposed test items were reviewed and optimized by the first researcher (RF) before they were included in the final versions.
- To determine the face validity of the tests, we invited nine experts (two faculty members and one private practitioner for each of the three master's tracks) to examine the relevance of the items to be used with regard to (a) veterinary practice and (b) the knowledge base required at the time of graduation.
- We composed the final PT versions, each covering 150 test items. Previous studies have indicated that a PT in human medicine requires about 150 to 200 test items to achieve a high level of reliability (Cronbach's  $\alpha > .80$ ) or a high G coefficient ( $> .80$ ).<sup>18,19</sup>

## Test Format

We developed the test according to the format of the Maastricht Progress Test.<sup>20</sup> Each test consisted of 150 multiple-choice questions formulated as single-best-answer items. Items had an "I don't know" option (question mark). The 150 items were related to the core program (90 items) and to the three master's tracks (60 items; 20 items per track). The final score was established as follows: correct answer = 1 point, incorrect answer = -1 point, and question mark = 0 points (formula scoring). This resulted in a range of possible scores from -150 to +150 points. The

**Table 2:** Mean scores for PT 1 and PT 2

	Total items	Correct	Incorrect	? ("I don't know")	Overall score
PT 1 (n = 331)	150	69	41	40	27 (SD = 16.6)
PT 2 (n = 292)	149	68	36	45	32 (SD = 17.1)

90 core- program items were selected from the modules in the veterinary bachelor's curriculum (C2007). The other 60 master's-track items were specifically created for use in these tests.

### Participant Post-Test Questionnaire

In addition to the test, a short questionnaire was administered to students asking for their feedback, in particular about the PT format and the representativeness of the questions. It consisted of 10 questions (Figure 1) to be rated on a 5-point Likert scale (from 1 = *completely disagree* to 5 = *completely agree*).

### Data Analysis

First, we evaluated reliability and validity as follows:

1. Based on the answer key, we calculated the estimated internal consistency reliability, *p* values (difficulty index), point-biserial correlations (item discrimination index), and distractor efficiency (DE). The individual scores of participants were checked to uncover deviant response patterns.
2. If indicated (*p* values  $\leq .1$  or  $\geq .9$  and item-total correlations  $< .25$ ), two senior veterinarians independently reviewed items to reassess their (content and construct) validity.<sup>21</sup>
3. If necessary, we removed items from the test that proved invalid. Based on the final answer key and scoring model, we established the final scores of participants and re-estimated internal consistencies.
4. Generalizability theory provides a method by which to disentangle the contributions of multiple factors (e.g., items, test occasions, raters) and their interactions with the reliability of results.<sup>22</sup> To determine the reproducibility of test results and the effects of repeated use, we conducted a G study (variance component analysis) based on a two-facet fully-crossed design with the items, participants, and the two occasions as facets. Student scores were transformed to *z* scores before the G study was done. D studies were done to establish the number of test occasions and items in each test required to achieve a satisfactory level of reliability.

Finally, to evaluate test sensitivity, we compared the individual results obtained in both tests to disclose whether this PT format measured a significant change in knowledge base.

### Test Conduct and Participants

This test was conducted during the master's phase of the veterinary medical curriculum. Students can enter this phase twice a year. This means that at any given time there are five or six different student cohorts in the 3-year master's program. Two formative tests were administered with an interval of 6 months (the first in December 2011, the second in June 2012). At the time the tests were taken, the master's program accommodated students from both the 2001 curriculum (C2001) and the 2007 curriculum (C2007). Participation in the tests was only mandatory for the C2007 students; students from C2001 were encouraged to participate in both tests. The test duration was 3 hours. All students from the three master's tracks (equine health, companion-animal health, and farm-animal health) participated in the same test. The results from these two tests were neither included in the course assessment program nor revealed to the teaching staff. The students received individual feedback about their scores and guidance in the interpretation of results.

### Confidentiality

Before students participated in the PT, we informed them that the test results would be used anonymously for evaluation and research purposes. The Netherlands Association for Medical Education approved this study with regards to ethical considerations.

### RESULTS

In total, 331 students participated in the first test, 292 participated in the second, and 247 participated in both. This disproportionate ratio between students who took both tests and those who took only one can mainly be explained by the fact that the last students who entered the master's program and those students who completed the program after the first run had no opportunity to take both. Table 2 presents the average end scores for both tests, the number of correct and incorrect answers, and the number of times students opted for the "I don't know" option.

### Reliability and Content Validity

The internal consistency (Cronbach's  $\alpha$ ) of the first and the second test were .86 and .88, respectively. The internal consistency for the combined scores of those students who participated in both tests was .81.

**Table 3:** Reliability and content validity of PT 1 and PT 2

	PT 1 (n = 331)	PT 2 (n = 292)
Cronbach's $\alpha$	.86	.88
mean $p$	.46 $\pm$ .24	.46 $\pm$ .23
mean RIR	.19 $\pm$ .09	.20 $\pm$ .09
% items of average difficulty	86	71
% items with a good or excellent RIR	62	69
% optimal items ( $p = .30$ to $.70$ ; RIR $> .20$ )	69	98

**PT 1**

The mean  $p$  value and item-rest correlation (RIR) were .46  $\pm$  .24 and .19  $\pm$  .09, respectively. Sixty-eight percent (102) of items were of average (recommended) difficulty (mean  $p = .53 \pm .18$ ) and had DE = 80.4% (Table 3). Sixty-two percent (92) of items had a good or excellent RIR (.27  $\pm$  .05) with DE = 82.6%. Combining the two indices, 69 items (46%) could be called "optimal" ( $p = .30$  to  $.70$ ; RIR  $> .20$ ) and had DE = 86.1%. Only one item had a non-functioning distractor (NF-D). On the basis of these results, 20 items (13%) were reviewed to identify possible (content or construct) validity problems. No items were removed.

**PT 2**

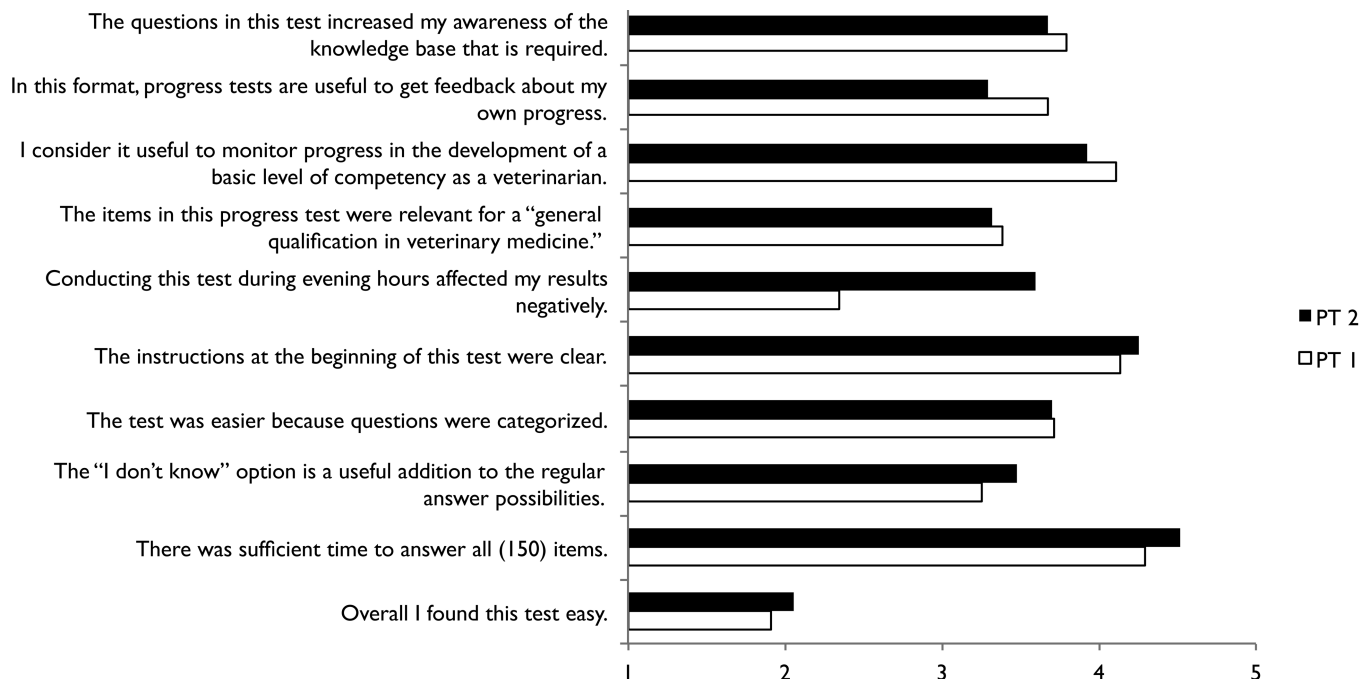
The second test revealed similar results; the mean  $p$  value and RIR were .46  $\pm$  .23 and .20  $\pm$  .09, respectively. Seventy-one percent of items (106) were of average difficulty (mean  $p = .50 \pm .17$ ) and had DE = 79.4% (Table 3).

Sixty-nine percent of items (102) had an excellent RIR (.47  $\pm$  .20) with DE = 81.0%. With the two indices combined, 98 items (66%) could be called "optimal" ( $p = .30$  to  $.70$ ; RIR  $> .20$ ) and had DE = 82.8%. There were four items with one NF-D. With these results, eight items (5.3%) were reviewed to identify (content or construct) validity problems. One item was removed.

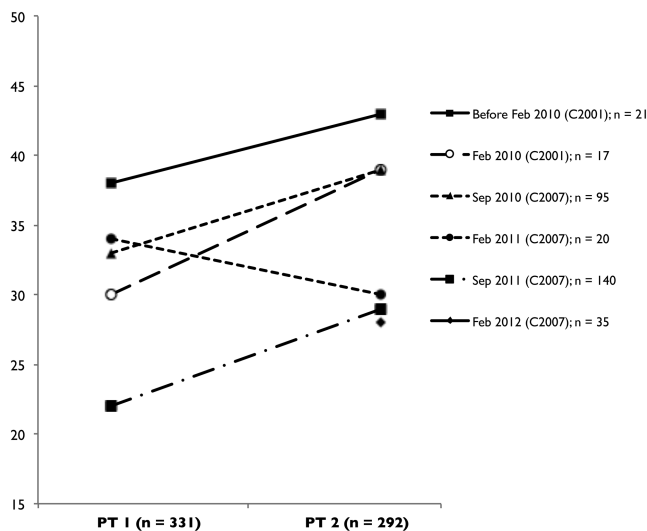
As can be seen from their mean  $p$  values ( $\pm SD$ ), both tests were about equally hard to take (PT 1:  $p = .46 \pm .24$  and PT 2:  $.46 \pm .23$ ).

**Expert Opinions (Validity)**

For both tests, all experts reviewed the 90 items pertaining to the core curriculum to establish their level of complexity and relevance to practice. The remaining 60 track-specific items were only reviewed by the experts of the respective tracks. Among the experts, the average agreement on items was 89% for the first test and 91% for the second. Complete agreement existed regarding

**Figure 1:** Survey outcomes regarding students' perceptions of the progress test, rated on a 5-point Likert scale (1 = completely disagree to 5 = completely agree)

PT 1: n = 326; PT 2: n = 287



**Figure 2:** The overall scores (ranging from -150 to +150 points) for PT 1 and PT 2 of the concurrent cohorts (C2001 and C2007)

The master’s program (C2007) has two starting times per year, resulting in September and February cohorts.

69% of all items. Differences between core-curriculum items and track-specific items were non-significant. Disagreements among experts mainly concerned the relevance of an item for practice; in most of these cases, one or two experts considered the knowledge too advanced or detailed to be part of students’ knowledge base at the time of graduation.

### Student Scores and Progress

Figure 2 presents the overall scores of the concurrent cohorts (from first semester year 4 to second semester year 5 [C2007] and year 5 and year 6 [C2001]). Except for the cohort that started in the master’s phase in February 2010, all cohorts showed progress between the two tests. This progress ranged between 32% (September 2011) and 12% (February 2010), with an average of 16% for all cohorts. The effect size (Cohen’s *d*) for the combined groups is .45. Progress was mainly achieved in the track-specific issues (13% increase) and choice of treatment (18%) categories.

Comparing scores on the core items to the whole test results shows that, on average, students performed better on the core items, but they progressed less in this cate-

gory of items (mean 29% correct versus 18% in PT 1; 25% versus 22% in PT 2).

At an individual level, student scores (correct minus incorrect) improved from the first test ( $M = 26.6$ ,  $SD = 16.6$ ) to the second test ( $M = 32.4$ ;  $SD = 17.1$ ). The improvement is significant ( $t = -7.307$ ,  $df = 246$ ,  $p < .001$ ). Furthermore, students’ individual scores on the first and second test correlate positively ( $r = .62$ ,  $N = 147$ ,  $p < .001$ ). The “I don’t know” option was selected slightly more often (4%) in the second test.

### Generalizability

Table 4 details the results from the generalizability analysis<sup>22</sup> of participant results and the relative contribution of different sources of variance. The G study revealed that 70.2% of the result-to-result variance was caused by realistic differences between participants. Four of the five cohorts exhibited improvement on the test (Figure 1). An additional D study demonstrated that, in order to obtain a reliable measurement of participants’ progress made as indicated by a generalizability coefficient of  $>.8$ , at least four tests should be taken that each contain a minimum of 125 items (Table 4).

### Participant Post-Test Questionnaire

The responses to both questionnaires revealed that students appreciated the PT for its ability to reveal their knowledge progress (ratings of 4.1 and 3.9 on a 5-point Likert scale for PT 1 and 2, respectively), to increase their awareness of the level of knowledge required for graduation (3.8 and 3.7, respectively), and to provide feedback (3.7 and 3.3, respectively). Despite the clear instructions (4.1 and 4.2, respectively) and sufficient time to take the tests (4.3 and 4.5, respectively), the questions were judged to be difficult (not easy) (1.9 and 2.1, respectively). Students were neutral about the value of the “I don’t know” option (3.2 and 3.5, respectively) (Figure 1).

### DISCUSSION AND CONCLUSIONS

The results from this study about the use of progress testing in a veterinary curriculum reveal that the test format met the quality criteria and expectations for assessment.

First, given that the test format is relatively independent of the educational program, the PT appeared to be suitable for use in an integrated curriculum. Although progress in knowledge development is usually most pronounced in the first years of medical education,<sup>6</sup> the first- or second-year master’s students in our study, except for

**Table 4:** G coefficients as a function of the number of items (100, 125, 150, 175, 200) and test occasions (2, 3, 4, 6)

Number of items	Number of test occasions			
	2	3	4	6
100	.641	.709	.775	.834
125	.676	.742	.802	.855
150	.702	.767	.821	.870
175	.722	.782	.835	.881
200	.738	.796	.846	.889

the February 2011 cohort, did achieve moderate progress in the retention of functional knowledge on the pilot PTs. The deviating cohort followed the same educational program during which no incidents were reported. The decrease in test score may be explained by incidental variation between groups. Other PTs organized in medical curricula that were also based on two subsequent test occasions have produced similar results. Nevertheless, these groups also demonstrated progress in the retention of functional knowledge after administration of more tests over a longer period of time.<sup>23,24</sup> Furthermore, a comparison of the individual results for both tests confirms that the PT was sensitive enough to detect differences in student knowledge at different levels of experience. On average, student progress between the two tests proved to be substantial. Other comparative studies based on PTs report similar degrees of progress in this phase of medical training.<sup>25</sup>

Second, both tests proved sufficiently reliable to render them acceptable tools for both formative as well as summative assessment.<sup>18</sup> Despite the variety of species and clinical issues to be covered, and despite the differences between students from the 2001 and 2007 curricula who participated in the test, the internal consistency was high. The reliability of results was confirmed by the high G coefficient from the generalizability analysis. At least three tests a year during the 3-year master's phase would be sufficient to support and monitor students' knowledge development.

Third, participants and experts largely agreed on the validity of the test, both in terms of the relevance of items for veterinary practice and the appropriateness of the items' levels of complexity. In most of the cases in which experts clearly disagreed, the external practitioner considered the item irrelevant, too difficult, or too advanced, whereas FVMU staff members regarded the item as part of the knowledge base students should have acquired by the end of their initial training. At the time the tests were taken, the C2001 participants were in the second or third year of their master's, whereas the C2007 students were in their first or second year. This means that the students from the revised curriculum participated in the PT about a year earlier than the other students did. Still, both groups obtained comparable scores. This confirms that if the PT targets functional knowledge that should be acquired by the time of graduation, it can be used more or less independently of the specific curricular structure or sequence of topics covered.<sup>11</sup>

Fourth, students perceived progress testing as relevant to their future practice. It increased their awareness of the level of knowledge required for graduation and allowed them to monitor their knowledge development. Although students perceived many items in the test as difficult, their opinions of the "I don't know" option remained neutral, albeit with a large *SD*.

In this study, the PT was only administered twice. Developing proper standards for the scores to be expected at various times in the master's program requires a lot more tests. Furthermore, in this study we mainly focused on the psychometric analysis of the test scores and results. The short- or long-term effects of progress testing on the

way students adapt their study behavior to competence development was beyond the scope of this study. Another limitation of the study might be the fact that both pilot PTs were used as formative and not summative assessments. This could have affected the students' choices, for example with regard to their preparation for the test or to the use of the "I don't know" option. Furthermore, at the time both tests were taken, no other knowledge testing was part of the master's program of the C2007 curriculum. C2007 students had no other bench marker for how they were performing and they might have been happy to be tested. This aspect might also have positively influenced the outcome of the student post-test questionnaire.

In conclusion, the PT format used in our study is suitable for incorporation into a veterinary medical curriculum. In order to deal with issues around relevance of items, item complexity, and test format, a review committee seems indispensable. Establishing a systematic framework as described by Wrigley et al.,<sup>19</sup> including a review committee that reviews and controls an item bank, will be an important additional step in introducing progress testing into veterinary education.

## ACKNOWLEDGMENTS

The authors wish to thank all the students, external experts, and FVMU staff members who participated in this study, Dr. Roos Goverde who helped with the layout of figure 1, and Angelique van den Heuvel for improving the English language.

## AUTHOR INFORMATION

**Robert P. Favier**, DVM, PhD, is Assistant Professor in Companion Animal Internal Medicine, Faculty of Veterinary Medicine, Utrecht University, P.O. Box 80154, 3508 TD Utrecht, Netherlands. Email: R.P.Favier@uu.nl. His research interests include assessment programs, work-based assessment, and competency frameworks. He is responsible for the master's program in Companion Animal Health.

**Cees P.M. van der Vleuten**, MA, PhD, is Professor in Education, Department of Educational Development and Research, Faculty of Health, Medicine, and Life Sciences, Maastricht University, P.O. Box 616, 6200 MD Maastricht, Netherlands. Email: C.vanderVleuten@maastrichtuniversity.nl. His research interests include assessment programs, work-based assessment, and quality assurance.

**Stephan P.J. Ramaekers**, PhD, is Associate Professor in Education, currently at the Amsterdam School of Health Professions. His research interests include the development and assessment of expertise, clinical reasoning, and shared decision making.

## REFERENCES

- 1 Prince KJ, Boshuizen HP, van der Vleuten CPM, et al. Students' opinions about their preparation for clinical practice. *Med Educ*. 2005;39(7):704–12. Medline:15960791 <http://dx.doi.org/10.1111/j.1365-2929.2005.02207.x>.

- 2 Gilling ML, Parkinson TJ. The transition from veterinary student to practitioner: a “make or break” period. *J Vet Med Educ.* 2009;36(2):209–15. Medline:19625670 <http://dx.doi.org/10.3138/jvme.36.2.209>.
- 3 Diemers AD, Dolmans DH, Verwijnen MG, et al. Students’ opinions about the effects of preclinical patient contacts on their learning. *Adv Health Sci Educ Theory Pract.* 2008;13(5):633–47. Medline:17629786 <http://dx.doi.org/10.1007/s10459-007-9070-6>.
- 4 Ramaekers SP, van Beukelen P, Kremer WD, et al. An instructional model for training competence in solving clinical problems. *J Vet Med Educ.* 2011;38(4):360–72. Medline:22130412 <http://dx.doi.org/10.3138/jvme.38.4.360>.
- 5 Schuwirth LW, van der Vleuten CPM. The use of progress testing. *Perspect Med Educ.* 2012;1(1):24–30. Medline:23316456 <http://dx.doi.org/10.1007/s40037-012-0007-2>.
- 6 van der Vleuten CPM, Verwijnen GM, Wijnen WHFW. Fifteen years of experience with progress testing in a problem-based learning curriculum. *Med Teach.* 1996;18(2):103–9. <http://dx.doi.org/10.3109/01421599609034142>.
- 7 McHarg J, Bradley P, Chamberlain S, et al. Assessment of progress tests. *Med Educ.* 2005;39(2):221–7. Medline:15679690 <http://dx.doi.org/10.1111/j.1365-2929.2004.02060.x>.
- 8 Rademakers J, Ten Cate TJ, Bär PR. Progress testing with short answer questions. *Med Teach.* 2005;27(7):578–82. Medline:16332547 <http://dx.doi.org/10.1080/01421590500062749>.
- 9 Norman G. Research in medical education: three decades of progress. *BMJ.* 2002;324(7353):1560–2. Medline:12089095 <http://dx.doi.org/10.1136/bmj.324.7353.1560>.
- 10 Norman G, Neville A, Blake JM, et al. Assessment steers learning down the right road: impact of progress testing on licensing examination performance. *Med Teach.* 2010;32(6):496–9. Medline:20515380 <http://dx.doi.org/10.3109/0142159X.2010.486063>.
- 11 Albano MG, Cavallo F, Hoogenboom R, et al. An international comparison of knowledge levels of medical students: the Maastricht Progress Test. *Med Educ.* 1996;30(4):239–45. Medline:8949534 <http://dx.doi.org/10.1111/j.1365-2923.1996.tb00824.x>.
- 12 Bereiter C, Scardamalia M. Learning to work creatively with knowledge. In: De Corte E, Verschaffel L, Entwistle N, et al, editors. *Unravelling basic components and dimensions of powerful learning environments.* Amsterdam: Pergamon; 2003, p. 55–68.
- 13 Zimmerman BJ, Schunk DH. Self-regulated learning and academic achievement: theoretical perspectives. Mahwah, NJ: Lawrence Erlbaum; 2001.
- 14 Bennett J, Freeman A, Coombes L, et al. Adaptation of medical progress testing to a dental setting. *Med Teach.* 2010;32(6):500–2. Medline:20515381 <http://dx.doi.org/10.3109/0142159X.2010.486057>.
- 15 Siegling-Vlitakis C, Birk S, Kröger A, et al. PTT: Progress Test Tiermedizin. Ein individuelles Feedback-Werkzeug für Studierende [Veterinary medicine progress test: an individual feedback tool for students]. *Dtsch Tierärzteblatt.* 2014;8:1076–82.
- 16 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. Medline:22364452 <http://dx.doi.org/10.3109/0142159X.2012.652239>.
- 17 Bok HG, Teunissen PW, Favier RP, et al. Programmatic assessment of competency-based workplace learning: when theory meets practice. *BMC Med Educ.* 2013;11(13):123. Medline:24020944 <http://dx.doi.org/10.1186/1472-6920-13-123>.
- 18 Van der Vleuten CP, Norman GR, De Graaff E. Pitfalls in the pursuit of objectivity: issues of reliability. *Med Educ.* 1991;25(2):110–8. Medline:2023552 <http://dx.doi.org/10.1111/j.1365-2923.1991.tb00036.x>.
- 19 Wrigley W, van der Vleuten CP, Freeman A, et al. A systemic framework for the progress test: strengths, constraints and issues: AMEE Guide No. 71. *Med Teach.* 2012;34(9):683–97. Medline:22905655 <http://dx.doi.org/10.3109/0142159X.2012.704437>.
- 20 van Berkel HJM, Sprooten J, Graff E. An individualised assessment test consisting of 600 items—the development of a progress test for a multi-master programme health sciences curriculum. In: Bouhuijs PAJ, Schmidt HG, van Berkel HJM, editors. *Problem-based learning as an educational strategy.* Maastricht: Network of Community-Oriented Educational Institutions for Health Sciences, Network Publications, 1993, p. 259–69.
- 21 Borsboom D, Mellenbergh GJ, van Heerden J. The concept of validity. *Psychol Rev.* 2004;111(4):1061–71. Medline:15482073 <http://dx.doi.org/10.1037/0033-295X.111.4.1061>.
- 22 Brennan RL. *Generalizability theory.* New York: Springer; 2001. <http://dx.doi.org/10.1007/978-1-4757-3456-0>.
- 23 Tomic ER, Martins MA, Lotufo PA, et al. Progress testing: evaluation of four years of application in the school of medicine, University of São Paulo. *Clinics (Sao Paulo).* 2005;60(5):389–96. Medline:16254675 <http://dx.doi.org/10.1590/S1807-59322005000500007>.
- 24 Interuniversitaire Voortgangstoets Geneeskunde (iVTG). Progress testing [Internet]. iVTG; 2015 [cited 2016 Dec 19]. Available from: <http://ivtg.nl/>.
- 25 Verhoeven BH, Verwijnen GM, Scherpbier AJ, et al. Growth of medical knowledge. *Med Educ.* 2002;36(8):711–7. Medline:12191053 <http://dx.doi.org/10.1046/j.1365-2923.2002.01268.x>.