1. Title: Red, Yellow and Green; What does it mean? How the Progress Test informs and supports student progress

2. **Authors:** Given, K., Hannigan, A., McGrath, D.

3. **Institution research conducted:** University of Limerick Graduate-Entry Medical School

4. **Corresponding Author:** Karen Given, 67 Inis Cealtra, Ballina, Co. Tipperary, Ireland. Tel: 00353 876417550 Fax: 061 233778 Email address: karengiven16@yahoo.co.uk

Abstract

Objectives: Most medical schools using progress tests (PTs) provide feedback by identifying poorly-performing students utilizing a traffic-light system of green (satisfactory), yellow (borderline) and red (unsatisfactory) categories. There is little research assessing students' perceptions or usage of this feedback. This study proposed to determine the effectiveness of formative PTs at informing and supporting student progress.

Methods: A mixed methods study was performed, involving a retrospective analysis of a results database to establish the predictive validity of PT categories and semi-structured interviews to explore students' perceptions of PT feedback.

Results: Quantitative analysis revealed that students who always scored green performed better in their summative exams and graduated with a higher final degree than those who received a yellow or red category at least once. Qualitative analysis revealed that most students scoring well perceived the PT as having informed their progress with all students scoring poorly perceiving that it didn't inform their progress. Most students agreed that the current feedback is insufficient and doesn't guide their on-going learning.

Conclusion: While this study demonstrated that the PT is a useful predictive tool for informing student progress, in its current format it's not fulfilling a truly formative role and thus supporting student progress sufficiently.

Introduction

Progress Tests

Progress tests (PTs) are longitudinal, repeated assessments measuring knowledge acquisition over time, set at the level of a newly-qualified doctor, but completed by all students throughout their training (Wrigley et al., 2012). Although progress-testing is widely used internationally, its content and implementation in different medical schools varies (Wrigley et al., 2012, Freeman et al., 2010). Furthermore while most schools identify poorly-performing students, by utilizing satisfactory, borderline and unsatisfactory categories, the amount of additional feedback is also variable across institutions (Blake et al., 1996, Wade et al., 2012).

Formative role of progress-testing

PTs are used either for formative or summative assessment (where they may also provide a formative role) (Wrigley et al., 2012). Koh summarised the benefits of formative assessment on student learning as: (1) ‘development of deep thinking’, (2) ‘maintenance of motivation and self-esteem’ and (3) ‘encouragement of self-regulated learning’ (Koh, 2008 p224-225). While formative progress-testing aims to reap these benefits, its success in the latter two areas has not yet been fully investigated.

In 1996, van der Vleuten proposed that the utility of an assessment depends on its validity, reliability, educational impact, acceptability and costs (van der Vleuten, 1996). These aspects are weighted differently depending on the purpose of the assessment. Where the purpose is formative, as is often the case with progress-testing, educational impact and acceptability should be considered high priority.

Studies have indicated that students may not take purely formative PTs seriously (Wrigley et al., 2012, Nouns and Georg, 2010). A survey of 1500 students found that students were motivated to take formative PTs seriously when they understood the concept of formative testing and used their results to assist future learning (Nouns and Georg, 2010). However, how students used their results to assist learning was not investigated as part of this study.

Feedback from progress-testing

Feedback is a key feature of formative assessment (Rolfe and McPherson, 1995, Rushton, 2005, Ende, 1983). The effectiveness of feedback depends on the content and method by which it's given, as well as the mind-set of the recipient (Archer, 2010). Negative feedback can provoke an adverse emotional response which can affect how the feedback is accepted and subsequently used, and may result in missed learning opportunities (Sargeant et al., 2008, Baron, 1988).

A recent quantitative study compared student perceptions of the PT in two UK universities, where feedback is provided in different ways (Wade et al., 2012). School-A delivered feedback by giving students their total score and a norm-referenced band. School-B students received an electronic log with their score (total and by subject area) and a norm-referenced band as well as key learning points for each question, and individual feedback from an academic tutor biannually. Despite School-B's seemingly comprehensive feedback, students still perceived it as insufficient, although less so than at School A (Wade et al., 2012). A qualitative component which was not undertaken may have helped understand why the students regarded the feedback as insufficient and explored how they used this feedback.

In Maastricht, students undertake the PT every three-months and may review the questions and answers afterwards. Their scores are aggregated to different sub-scores by organ systems or disciplines. Interestingly most students have a low level of interest in checking answers and reviewing their sub-scores (van der Vleuten and Verwijnen, 1996). If students tend not to look at the detailed feedback received, the question arises as to whether it's worth investing resources to provide this type of feedback? Alternatively, should we be encouraging more active engagement with feedback?

Predictive validity of progress-testing

Where a test is used summatively, it naturally predicts progression. When used only formatively, it's beneficial to know if it predicts performance in subsequent summative exams, hence informing student progress. Staff can then implement remediation and students can direct their study accordingly.

Although it's assumed that collated longitudinal PT data is a better predictor of future performance than one-off measurements (Schuwirth and van der Vleuten, 2012), institutions vary in their use of individual test or aggregated results. For those using aggregated results, many different statistical methods of collating results exist (Blake et al., 1996, Schuwirth and van der Vleuten, 2012, McHarg et al., 2005, van der Vleuten and Verwijnen, 1996, Ricketts and Moyeed, 2011). The predictive validity of PTs therefore is likely to vary depending on the collation method used and on what it is trying to predict. While some research has already been undertaken on the reliability and predictive validity of individual or aggregated raw scores, there have been no studies investigating the predictive validity of the satisfactory, borderline and unsatisfactory categories commonly given to students with each individual test result.

Aim of study

How students value and trust progress-testing is important when assessing its educational impact and acceptability. From the limited existing research, which lacks a detailed qualitative component, it’s unclear how students perceive and use PT feedback, particularly as to whether it guides their on-going learning. It’s also uncertain if students’ result categories in individual PTs over time are a reliable predictor of progression.

The aim of this study therefore was to explore the effectiveness of the formative PT in informing and supporting student progress at University of Limerick Graduate-Entry Medical School (UL-GEMS) looking in particular at:

a) PT category as a predictor of student performance (categories defined as green/satisfactory, yellow/borderline and red/unsatisfactory).

b) Students’ perception of the PT and whether the corresponding feedback informs and supports their progress.

Methods

Research paradigm and strategy

Pragmatism, which aims to solve practical problems resulting in useful consequences provided the philosophical basis for this study (Feilzer, 2010). A convergent parallel mixed methods design was selected, fitting well within the pragmatism paradigm, which draws on both positivism and interpretive epistemology (Creswell and Plano Clark, 2011).

Context

This study was conducted at UL-GEMS, which delivers a four-year graduate-entry medical programme. Their progress test is sourced from McMaster University in Canada, where it was introduced in 1992 (Blake et al., 1996). Students undertake this compulsory formative test biannually. The following feedback is provided to students online:

- Raw and corrected score (items correct – (0.25 X items incorrect)) as percentages
- Last four scores to see progression
- % attempted – of this % correct and incorrect
- Breakdown into three sub-categories: Biology, Behaviour and Population
- Class average and standard deviations
- Category - red, yellow or green reflecting their performance in relation to their year-group. Their category is assigned based on their Z-score, derived when their raw scores are norm-referenced within each student cohort using the equation Z-score = (student score - mean score)/standard deviation. Red represents unsatisfactory ($Z <-2$), yellow borderline ($Z \geq -2; <-1.5$) and green satisfactory ($Z \geq -1.5$).(Finucane et al., 2010)

All students in the red category attend a compulsory meeting with a senior faculty member where they receive more detailed verbal feedback regarding their progress. Students scoring yellow are invited to meet their academic tutor for further feedback.

Summative knowledge exams (SKE) in each module are scored from 0 to 100%. Degrees are awarded to students based on a quality cumulative average (QCA) score over modules in the final two years of the programme (Table 1).

Ethical approval was granted from the University’s Faculty of Education and Health Sciences Research Ethics Committee.

Quantitative phase

Methodology

A retrospective analysis of an existing PT results database linked to a summative results database was undertaken.

Participant selection

The population studied included all four cohorts of students who have graduated from UL-GEMS since it opened in 2007 (n=285).

Data analysis

Due to small numbers of students allocated to red or yellow categories, these were combined into a new category, “combined flags”. Therefore for the purpose of this paper, a flag refers to both the yellow and red categories. To establish the predictive validity of PT categories, a new variable was created based on each student’s eight PT results over their four years of study to divide students into either “all green” (scoring green in all tests) and “≥ 1 flag”(scoring either yellow or red in one or more tests) categories. The graduating QCA score for these two categories was then compared across all students. In addition, the “all green” and “≥1 flag” variables were calculated for each individual year and related to the SKE result for the corresponding year. Numeric variables were tested for normality and summarised using mean (standard deviation). The differences between mean QCA and mean SKE results for the two PT categories (“all green” and “≥1 flag”) were explored using independent samples t tests. A 5% level of significance was used for all statistical tests. Statistical analysis was carried out using SPSS Version 21 for Windows.

Qualitative phase

Methodology

A pragmatic qualitative research (PQR) approach was selected because it simply seeks to explore and understand the viewpoints of the participants involved. In keeping with the paradigm of pragmatism, PQR utilizes the most practical methods available to answer the research question (Savin-Baden and Howell Major, 2013, Caelli et al., 2003).

Participant selection and recruitment

The sampling frame was the current second to fourth-year students and 2013 graduates currently completing their internship in Ireland (n=428). An independent gatekeeper emailed students an invitation letter and information leaflet. To decide which of the 18 respondents to interview, the ‘maximum variation’ type of purposive sampling was used (Marshall, 1996). Table 2 outlines the rationale for inclusion of each sampling parameter.

Data collection

Eleven one-on-one, face-to-face, semi-structured interviews were conducted by the researcher using an interview protocol developed from the literature review, discussion with students and pilot interview. It was adapted as themes developed through the iterative process of data collection and analysis (Savin-Baden and Howell Major, 2013). Interviews were audio-recorded with consent and subsequently professionally transcribed verbatim. They lasted on average 58 minutes (range 42-76 minutes). Data collection continued until saturation was reached.

Data analysis

Interview data was analysed using thematic analysis (Braun and Clarke, 2006). The process of thematic analysis involved 5 phases of coding, with the computer software package, NVivo version-10 being used to assist the analysis. Various strategies were used to optimise the rigour of the analysis including member checking with all participants reviewing their transcripts, peer debriefing with research supervisor, completing an audit trail and employing a reflexive approach throughout (Houghton et al., 2013, Lincoln and Guba, 1985).

Results

Quantitative results

PT category ("all green" or "≥ 1 flag") as a predictor of student performance:

a) Relationship between PT category and graduating QCA

A graduating QCA was available for 272 students of the 285 analysed. Of these 272 students, 208 (76%) scored green in all eight PT results. The mean graduating QCA of this group on a scale of 0 to 4 was 2.76 (SD 0.28) compared to a mean of 2.54 (SD 0.21) for the group who received one or more flags (n=64; 24%).

The difference of 0.22 in mean QCA between the two groups ("all green" versus "≥ 1 flag") was statistically significant (95% CI for the difference 0.16-0.29, $p<0.001$). No-one with more than one flag (i.e. ≥ 1 flag) received a 1st class honours degree and only 1 student with ≥ 1 flag achieved a 2.1 degree.

b) Relationship between PT category and SKE results

Students receiving a flag in each individual year (comprising two PTs) got lower results, on average, in the corresponding SKE compared to those who received green in both tests. The differences in mean SKE between these two groups for each year were statistically significant ($p<0.001$, Table 3).

Qualitative results

Table 4 outlines the demographics of participants. Figure 1 illustrates the four inter-relating themes generated from the qualitative analysis with feedback as an integral component of all themes.

Theme 1: Informing progress

Students' perceptions of how the PT informs their progress

Students were almost equally divided with the majority of students scoring green believing it was a relatively accurate reflection of their progress, while all students receiving a flag did not believe it was an accurate reflection.

> *"I don't think there will be a huge correlation between that (PT) and the end-of-year results." (P6)*

Some students explained that the PT may not be an accurate reflection of progress for certain colleagues due to lack of effort.

Theme 2: Feedback

Students' perceptions of online written feedback

Overall, most students find this feedback insufficient, particularly those scoring green:

> *"You can't really use it to your advantage because you've no feedback from it really, like feedback that's useful." (P10)*

Most students like getting their last 4 scores so they can see their progression as well as the class average and standard deviations so they had a better idea of where they were in the class. None of the students found the breakdown into 3 sub-categories useful as they felt the headings were too broad. While students like getting feedback from the PT and want more feedback, some students do not look at, understand or use the feedback currently provided:

> *"You get this load of numbers comparing each class and you know adjusted and corrected and all this, I don't fully understand it." (P3)*

Theme 3: Educational impact

The section below focuses on aspects, other than informing progress and feedback that affect the educational impact of the PT.

Students' understanding of the purpose of the PT

Most students believed the purpose of the PT was to allow UL-GEMS to compare itself to other schools (Figure 2). The international students in particular valued this comparison. Approximately half the participants did appreciate the formative nature of the PT but some displayed a lack of understanding of formative assessment:

> *"People would be fine with it (PT) if it was part of the assessment, but because it's not, that's where the negativity comes in because they feel it's not relevant" (P4)*

PT as a learning instrument

An overwhelming theme from the interviews was that students did not find the PT a useful learning instrument:

> *"The benefit of learning comes from seeing your mistakes and being able to like learn from them and we don't get that from the PT." (P10)*

The main way students thought the PT could become a learning instrument was by improving the feedback. The overwhelming student request was to see the PT answers afterwards. A few participants thought seeing the questions would be better, recognising the benefits of active learning. They felt that getting a list of specific topics of questions they got wrong would be useful or at least a more detailed breakdown into specialities and sub-specialities.

> *"Then it's very much turned it into an individual exam because they've given you these sort of pointers" (P9)*

Effect of the PT on motivation and self-esteem

Although students didn't change what or how they studied, the PT did provide most students in the green category with motivation to study. Of the students receiving a flag, one found the PT a de-motivator,

> *"I found it like a de-motivator and something that made me question my ability" (P5)*

while the other two claimed it had no effect on their motivation levels. No students prepared specifically for the PT (as intended); however the mid-term timing of the PT encouraged them to increase their usual self-directed learning earlier than they might otherwise be inclined to do. How the PT affected students' confidence levels varied depending on their result category (Figure 3).

Theme 4: Acceptability

While the above three themes all influence the acceptability of the PT to students, this section specifically focuses on students' perceptions of the PT which also affect its acceptability.

Students' perceptions of the PT

Participants sensed that the general consensus among students was negative towards the PT:

> *"Most students impression of it is it's time wasting because we're not gaining anything from it" (P1)*

Despite this apparent negative general consensus, only a minority of the interviewees had an overall negative opinion of the PT, with the majority being neutral or positive. Participant 11 felt students' results in the PT may affect their opinion:

> *"I think people dismissed it who probably didn't do well in it"*

This was supported by the data which showed that those who were always in the green category were more likely to have a positive opinion than those who received a flag (Figure 4).

Despite a previous study demonstrating that students in later years were more positive towards the PT (Wade et al., 2012), no relationship between year-group and opinion was evident in this data. However non-EU students who performed marginally better than EU students in a previous study conducted by UL-GEMS were more favourable towards the PT than EU students (Finucane et al., 2013).

Discussion

This study investigated the effectiveness of the PT in informing and supporting student progress in UL-GEMS.

Quantitative analysis demonstrated that the PT categories have predictive validity for subsequent summative exams. While it was clear from the qualitative analysis that students who performed poorly and a small number of those performing well did not appreciate this predictive value, it's hoped that the quantitative findings in this study will improve the credibility of PT categories as an informer of student progress and thus increase the likelihood of PT feedback being trusted and accepted (Archer, 2010).

It should be acknowledged, however, that not all students receiving a flag performed poorly in subsequent summative exams, although it was not within the scope of this study to determine whether getting a flag led these students to improve their study and thus perform better in summative exams. However some students did comment on the potential of scoring green providing a false sense of security. Further studies of the raw scores of students in the green category who subsequently perform poorly in summative assessments is therefore necessary to establish if those in this category who subsequently underperform can be identified by the PT.

Previous studies have shown that using aggregated scores reduces the effect of measurement errors and should more accurately rank order students (Ricketts and Moyeed, 2011, Muijtjens et al., 2010). Increasingly universities are now using aggregated results but they differ in their methods of aggregation, weighting of tests and the optimal number of tests included (McHarg et al., 2005, Schuwirth and van der Vleuten, 2012). While the use of aggregated scores is not as important when PTs are used formatively, one could argue that it would also more accurately identify students in need of remediation and its use should therefore be considered in this setting also.

With respect to the qualitative analysis, feedback was integral to all themes generated and was a necessary component for informing and supporting student progress. It was clear that UL-GEMS students wanted more feedback, which is not unique to this particular cohort (Ende, 1983). The qualitative results also suggest that students do not fully understand or reflect on the feedback given, in keeping with previous research both on feedback in general and in relation to the PT (Wade et al., 2012, van der Vleuten and Verwijnen, 1996).

It was also clear from this study that the majority of students did not view the PT as a learning instrument but regarded its chief purpose as a benchmarking exercise for the School. This may reflect the current UL-GEMS philosophy whereby the PT is used as a test of knowledge acquisition and for benchmarking purposes, rather than in the true meaning of a formative assessment instrument, where the focus is on providing feedback (Wrigley et al., 2012, Rushton, 2005).

The findings also suggested that a minority of students may not take the PT seriously, as is known to occur with formative tests and which may skew the results (Nouns and Georg, 2010, Wrigley et al., 2012). It has been shown that students are more likely to take PTs seriously when they understand the concept of formative assessment (Nouns and Georg, 2010). Therefore, focusing on increasing learning from progress-testing and on improving student understanding of its formative role could potentially increase the proportion of students taking the test seriously which in-turn might better inform student progress and improve learning.

Other benefits of formative assessment include the provision of motivation and reassurance, leading to increased self-esteem (Koh, 2008). It has also been acknowledged in the literature that negative feedback may reduce self-esteem and elicit other negative emotional responses that can hinder the use of feedback (Archer, 2010, Baron, 1988, Sargeant et al., 2008). In the case of students scoring green in this study, the PT did indeed increase motivation and confidence. However, confidence was reduced and negative emotions were revealed for two of three interviewees who performed poorly. It is imperative therefore that all efforts should be made to deliver feedback in the most effective way, providing appropriate support to the recipient (Archer, 2010, Veloski et al., 2006, Sargeant et al., 2008).

While this was a small study conducted in one institution and as different medical schools administer progress tests, calculate results and give feedback in different ways, one might question the generalizability and transferability of the findings. This study however has led to the following recommendations being defined which are applicable to all institutions who implement progress testing:

- PT categories as well as raw and aggregated scores are a useful predictor of student progress.

- The use of aggregated PT results in the formative setting may further assist in identifying students in need of remediation.
- When embarking on progress testing institutions should contemplate the function of the PT and implement it accordingly.
- Improving student understanding of the purpose and benefits of the PT is critical to informing progress and improving learning
- Students should be encouraged to reflect on all feedback provided rather than only the category.
- Feedback should be delivered in the most effective way, providing appropriate support to the recipient.

This study has also identified opportunities for future research, in particular looking at:

- The raw scores of students in the green category who subsequently underperform in summative assessments to establish if these students can be identified by the PT and receive early feedback and remediation.
- How students actually use the more comprehensive feedback given in some institutions and on whether usage of this feedback leads to improved performance.
- Opinions of poorly-performing students for whom the feedback may have negative effects and who have most potential to benefit from such feedback.

Conclusion

This study presents the first in-depth insight into students' perceptions of the PT and its corresponding feedback. While this study has demonstrated that simple feedback categories given with PT results inform students how they are progressing and predict subsequent performance, it has also revealed that in its current format, the PT is not fulfilling a truly formative role. It is hoped that greater emphasis on the content and delivery of resulting feedback may lead to improved educational impact and acceptability of the PT for students, without compromising its role in quality assurance.

Practice Points

- Progress test categories are a useful predictor of student progress.
- Institutions using progress tests formatively to guide on-going learning must provide sufficient feedback with appropriate support to the recipient.
- Students should be encouraged to reflect on all feedback provided rather than only the category.
- Improving student understanding of the purpose and benefits of progress testing is critical to enhancing learning.

Notes on contributors

KAREN GIVEN, MB BCh BAO MICGP MSc DWH DCH, PBL and Clinical Skills Tutor, University of Limerick Graduate-Entry Medical School.

AILISH HANNIGAN, BSc PhD, Associate Professor of Biomedical Statistics, University of Limerick Graduate-Entry Medical School.

DEIRDRE MCGRATH, MD FRCP FRCPI MMEd, Director of Education, University of Limerick Graduate-Entry Medical School.

Acknowledgements

We would like to acknowledge and thank the medical students who kindly gave their time to participate in interviews.

Declarations of Interest

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the article.

References:

ARCHER, J. C. 2010. State of the science in health professional education: effective feedback. *Medical Education,* 44**,** 101-108.

BARON, R. A. 1988. Negative effects of destructive criticism: impact on conflict, self-efficacy, and task performance. *The Journal of applied psychology,* 73**,** 199-207.

BLAKE, J. M., NORMAN, G. R., KEANE, D. R., MUELLER, C. B., CUNNINGTON, J. & DIDYK, N. 1996. Introducing progress testing in McMaster University's problem-based medical curriculum: psychometric properties and effect on learning. *Academic Medicine: Journal Of The Association Of American Medical Colleges,* 71**,** 1002-1007.

BRAUN, V. & CLARKE, V. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology,* 3**,** 77-101.

CAELLI, K., RAY, L. & MILL, J. 2003. 'Clear as Mud': Toward Greater Clarity in Generic Qualitative Research. *International Journal of Qualitative Methods,* 2.

CRESWELL, J. & PLANO CLARK, V. 2011. *Designing and Conducting Mixed Methods Research,* Thousand Oaks, California, SAGE.

ENDE, J. 1983. Feedback in clinical medical education. *Journal of the American Medical Association,* 250**,** 777-781.

FEILZER, M. Y. 2010. Doing Mixed Methods Research Pragmatically: Implications for the Rediscovery of Pragmatism as a Research Paradigm. *Journal of Mixed Methods Research,* 4**,** 6-16.

FINUCANE, P., FLANNERY, D., KEANE, D. & NORMAN, G. 2010. Cross-institutional progress testing: feasibility and value to a new medical school. *Medical Education,* 44**,** 184-186.

FINUCANE, P., FLANNERY, D., MCGRATH, D. & SAUNDERS, J. 2013. Demographic attributes and knowledge acquisition among graduate-entry medical students. *Medical Teacher,* 35**,** 134-138.

FREEMAN, A., NOUNS, Z. & RICKETTS, C. 2010. Progress testing internationally. *Medical Teacher,* 32**,** 451-455.

HOUGHTON, C., CASEY, D., SHAW, D. & MURPHY, K. 2013. Rigour in qualitative case-study research. *Nurse Researcher,* 20**,** 12-17.

KOH, L. C. 2008. Refocusing formative feedback to enhance learning in pre-registration nurse education. *Nurse Education in Practice,* 8**,** 223-230.

LINCOLN, Y. & GUBA, E. 1985. *Naturalistic inquiry,* California, Sage.

MARSHALL, M., N 1996. Sampling for qualitative research. *Family Practice,* 13**,** 522-525.

MCHARG, J., BRADLEY, P., CHAMBERLAIN, S., RICKETTS, C., SEARLE, J. & MCLACHLAN, J. C. 2005. Assessment of progress tests. *Medical Education,* 39**,** 221-227.

MUIJTJENS, A. M. M., TIMMERMANS, I., DONKERS, J., PEPERKAMP, R., MEDEMA, H., COHEN-SCHOTANUS, J., THOBEN, A., WENINK, A. C. G. & VAN DER VLEUTEN, C. P. M. 2010. Flexible electronic feedback using the virtues of progress testing. *Medical Teacher,* 32**,** 491-495.

NOUNS, Z. M. & GEORG, W. 2010. Progress testing in German speaking countries. *Medical Teacher,* 32**,** 467-470.

RICKETTS, C. & MOYEED, R. 2011. Improving progress test score estimation using bayesian statistics. *Medical Education,* 45**,** 570-577.

ROLFE, I. & MCPHERSON, J. 1995. Formative assessment: how am I doing? *The Lancet,* 345**,** 837-839.

RUSHTON, A. 2005. Formative assessment: a key to deep learning? *Medical teacher,* 27**,** 509-513.

SARGEANT, J., MANN, K., SINCLAIR, D., VAN DER VLEUTEN, C. & METSEMAKERS, J. 2008. Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use. *Advances in Health Sciences Education,* 13**,** 275-288.

SAVIN-BADEN, M. & HOWELL MAJOR, C. 2013. *Qualitative Research: The essential guide to theory and practice,* London, Routledge.

SCHUWIRTH, L. W. & VAN DER VLEUTEN, C. P. 2012. The use of progress testing. *Perspect Med Educ,* 1**,** 24-30.

VAN DER VLEUTEN, C. P. M. 1996. The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education,* 1**,** 41-67.

VAN DER VLEUTEN, C. P. M. & VERWIJNEN, G. M. 1996. Fifteen years of experience with progress testing in a problem-based learning curriculum. *Medical Teacher,* 18**,** 103.

VELOSKI, J., BOEX, J. R., GRASBERGER, M. J., EVANS, A. & WOLFSON, D. B. 2006. Systematic review of the literature on assessment, feedback and physicians' clinical performance*: BEME Guide No. 7. *Medical Teacher,* 28**,** 117-128.

WADE, L., HARRISON, C., HOLLANDS, J., MATTICK, K., RICKETTS, C. & WASS, V. 2012. Student perceptions of the progress test in two settings and the implications for test deployment. *Advances in Health Sciences Education,* 17**,** 573-583.

WRIGLEY, W., VAN DER VLEUTEN, C. P. M., FREEMAN, A. & MUIJTJENS, A. 2012. A systemic framework for the progress test: strengths, constraints and issues: AMEE Guide No. 71. *Medical Teacher,* 34**,** 683-697.

Table 1: Interpretation of QCA scores

QCA score	Equivalent % Score	Equivalent degree
3.4-4.0	>80	First class honours
3.0-3.39	>70	Second class honours grade 1 (2.1)
2.6-2.99	>60	Second class honours grade 2 (2.2)
2.0-2.59	>50	Third class honours

Table 2: The rationale for each sampling parameter

Sampling Parameter	Rationale
Category in PT	The PT should inform and support the progress of all students.
Year-group	A previous quantitative survey showed students in year 4 were more positive about PTs than those in earlier years (Wade et al., 2012).
Nationality – European-Union(EU) & non-EU	A previous UL-GEMS study revealed international students performed marginally better in the PT than Irish students (Finucane et al., 2013). Comparing EU and non-EU perceptions may therefore be useful.

Table 3: Year 1-4 SKE results (%) for students always in the green category and those receiving ≥1 flag in each year

	A) All green	B) ≥1 flag	Mean difference between A) and B) (95% CI)	p-value, independent samples t test
Mean SKE Year 1 (SD) (n=284)	66.0 (7.23) (n=264)	59.8 (8.36) (n=20)	6.2 (2.90-9.58)	<0.001
Mean SKE Year 2 (SD) (n=274)	68.7 (7.22) (n=253)	62.2 (6.57) (n=21)	6.5 (3.12-9.94)	<0.001
Mean SKE Year 3 (SD) (n=272)	70.0 (6.25) (n=243)	64.1 (5.30) (n=29)	5.9 (3.54-8.30)	<0.001
Mean SKE Year 4 (SD) (n=268)	70.6 (5.96) (n=240)	63.7 (5.09) (n=28)	6.9 (4.61-9.23)	<0.001

Table 4: Demographics of participants

Year	EU Status	Category in PTs
2 Interns 3 Fourth-years 4 Third-years 2 Second-years	8 EU 3 Non-EU	8 All Green 3 Both Red and Yellow

Figure 1: Thematic diagram

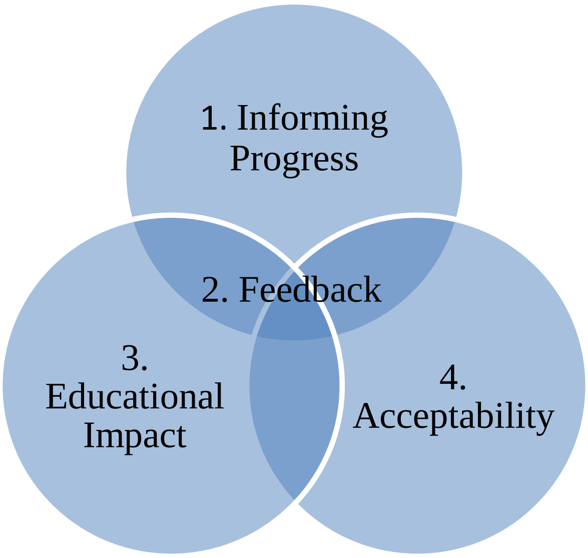


Figure 2: Purpose of the PT

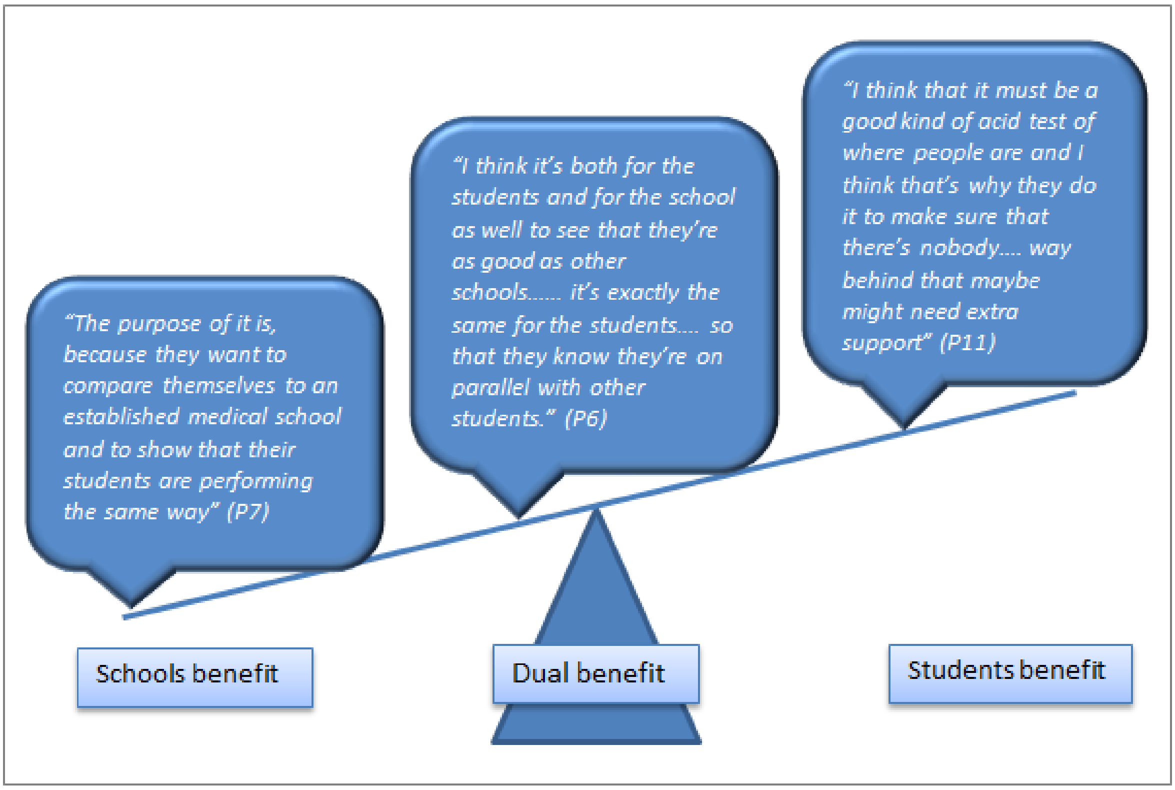

"The purpose of it is, because they want to compare themselves to an established medical school and to show that their students are performing the same way" (P7)
"I think it's both for the students and for the school as well to see that they're as good as other schools...... it's exactly the same for the students.... so that they know they're on parallel with other students." (P6)
"I think that it must be a good kind of acid test of where people are and I think that's why they do it to make sure that there's nobody.... way behind that maybe might need extra support" (P11)
Schools benefit
Dual benefit
Students benefit

Figure 3:

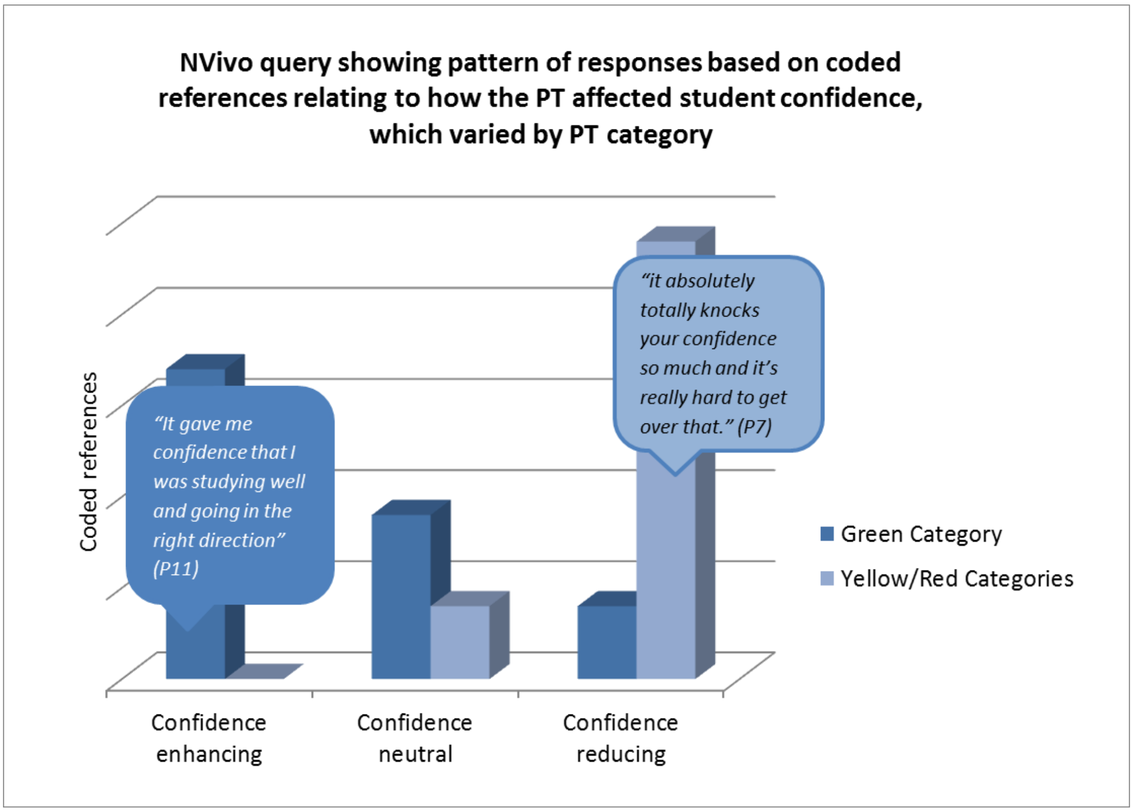
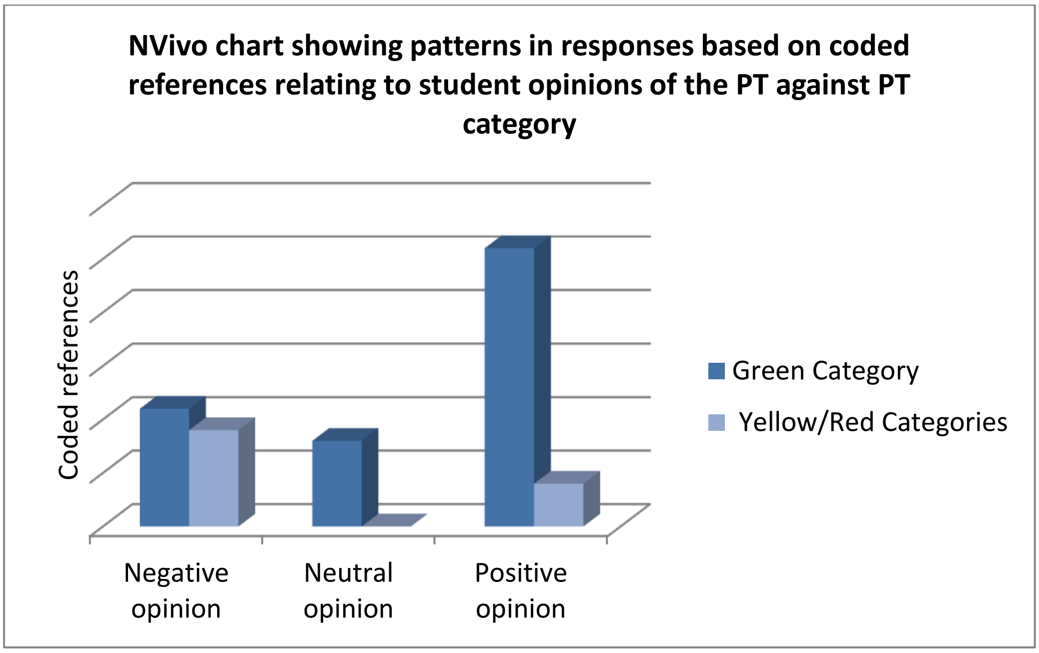

NVivo query showing pattern of responses based on coded references relating to how the PT affected student confidence, which varied by PT category
Coded references
"It gave me confidence that I was studying well and going in the right direction" (P11)
"it absolutely totally knocks your confidence so much and it's really hard to get over that." (P7)
Green Category
Yellow/Red Categories
Confidence enhancing
Confidence neutral
Confidence reducing

Figure 4:


NVivo chart showing patterns in responses based on coded references relating to student opinions of the PT against PT category
Coded references
Negative opinion
Neutral opinion
Positive opinion
Green Category
Yellow/Red Categories