**8P1 (133079)**

**Designing a programme of assessment: a stepwise approach**

*Angelique Timmerman*, **Maastricht University, Maastricht, Netherlands**

*Joost Dijkstra* (Maastricht University, Maastricht, the Netherlands)

**Background:** Assessment of complex tasks integrating several competencies necessitates a comprehensive evaluation. There has been a shift in thinking about assessment from methods to programs, combining several assessment activities to meet assessment purposes. A programmatic design approach is a next step, for which an overarching framework with 73 interrelated guidelines has been developed. Purpose is to develop a practical approach for programmatic assessment design.

**Summary of Work:** All guidelines have been applied to the assessment program of a GP residency training. Data were gathered through stakeholder interviews and analysis of key (legal) documents on which assessment practices are based. A written interpretative summary was made analysing strengths and weaknesses, resulting in themes important for assessment design, which were input for developing the practical approach.

**Summary of Results:** A stepwise design approach has been devised, covering the areas of the overarching framework. Start with outlining the context in which the program has to function (A. Assessment environment/organisational setting). Second, an initial approach has to be decided on, selecting those guidelines that are relevant (B. Overview and relevance of guidelines). Justification through documenting and underpinning is necessary to evaluate the robustness of design decisions and serve public accountability (C. Justification, expertise and implementation).

**Discussion:** The design and implementation of an assessment programme necessitates a holistic perspective, as all assessment activities are interrelated. The need of acceptance and expertise from stakeholders involved in running assessment activities makes a programmatic design approach always a team effort.

**Conclusion:** It is devised to start with a programmatic approach early in curriculum design, and involve context expertise, while looking at assessment not only as a psychometric design problem, but also as an instructional and organizational design problem.

**Take Home Messages:** Defining what the principal assessment activity is in a specific educational context is starting point for a programmatic design approach and development of a fit-for-purpose assessment programme through iterative cycles of redesign.

---

**#8P1 (135598)**

**Large-scale computerized adaptive progress testing: individual precision estimates, test-retest reliability and developmental validity**

*Carlos Fernando Collares*, **Maastricht University and European Board of Medical Assessors, Maastricht, Netherlands**

*José Juan Góngora Cortés* (Escuela de Medicina y Ciencias de la Salud. Instituto Tecnológico y de Estudios Superiores de Monterrey, Mexico)

*Silvia Lizett Olivares Olvera* (Escuela de Medicina y Ciencias de la Salud. Instituto Tecnológico y de Estudios Superiores de Monterrey, Mexico)

*Jorge Eugenio Valdez García* (Instituto Tecnológico y de Estudios Superiores de Monterrey, Mexico)

*Cees van der Vleuten* (Maastricht University, Maastricht, The Netherlands)

**Background:** Computerized adaptive testing (CAT) consists of a procedure in which an algorithm dynamically selects and administers the most appropriate items depending on previous responses in order to match items’ difficulties to examinees’ ability. Using a CAT approach for progress testing could potentially increase reliability while decreasing test duration.

**Summary of Work:** Two editions of a computerized adaptive progress test, configured to contain 50 items, have been administered in the same institution within a six-month interval (N = 1149 and N = 1217). Individual precision estimates were calculated based on Rasch model individual standard error estimates. Correlation of scores obtained in both test administrations was used to calculate test-retest reliability. A paired t-test and a Cohen’s d coefficient were used to describe the difference of scores for all students who took both tests (N = 882).

**Summary of Results:** Mean reliability of the first test was 0,81 and the mean reliability of the second test was 0,85. None of the individual reliability estimates was below 0,7. Test-retest correlation was 0,67 (p < 0,001). Score increase was significant on the paired t-test (t = 13,505; df = 881; p < 0,001) with a moderate effect size (Cohen’s d = 0,37).

**Discussion:** Individual reliability estimates were reasonably adequate, even for beginners and underachievers. Compared to other studies, the psychometric results were satisfactory, especially considering the reduced test length. A larger test could yield better reliability estimates and provide more comprehensive content validity. Students’ perceptions and interpretation of scores still need further studies.

**Conclusion:** Computerized adaptive progress testing presented adequate estimates of individual precision, a reasonable degree of test-retest reliability and a significant increase in scores after 6 months, with a moderate effect size.

**Take Home Messages:** Computerized adaptive progress testing can be a reliable and valid solution, which is especially attractive to schools that cannot administer a test to all students simultaneously.
A progress test to identify medical students with potential learning difficulties and to predict scores on the Canadian certification exam

Robert Gagnon*, Université de Montréal, Montréal, Canada
Christian Bourdy (Université de Montréal, Montréal, Canada)

Background: A formative progress test has been elaborated at the University of Montreal in 2012. It was expected that this procedure could help students to develop a self-learning approach and help them to keep a broader use of medical knowledge.

Summary of Work: Medical students were invited to stand a 180 items test on 5 occasions during year 3 and 4. All tests were equated on difficulty level. Two cohorts (n = 550) were used to get data on their performance. Data from pre-admission GPA, results on exams (year 1-2), and results on the official Canadian certification exam (Medical Council of Canada part I) were collected. Multiple regression analysis was used.

Summary of Results: A strong association between mean scores on all 5 tests and score on the MCC certification exam were observed (r = 0.68). When combined with all predictors (GPA, year 1 and 2 scores); multiple R is estimated at 0.73 (explained variance = 53 %). Progress test mean score show the largest partial correlation (r = 0.54). Analysis of the profile of students who failed on the certification exam (n = 19) clearly shows that this outcome was predictable.

Discussion: Progress tests were formative but we know all clerks were fully engaged in doing them. Everyone have received an individual feedback after each progress test so they can adjust their preparation before the next one. Those with difficulties in a special discipline also have been met to help them overcome their deficiencies.

Conclusion: The use of progress test is a major asset in the early identification of students who may show difficulties in integrating medical knowledge, which is the essential criteria for medical certification in Canada.

Take Home Messages: Using progress testing is a powerful tool for medical schools for early identification of students who may need special attention. It is also an efficient tool for students as self-learners.

Influence of curricula on medical students’ knowledge acquisition

Dario Cecilio Fernandes*, University of Groningen and University Medical Center Groningen, Center for Education Development and Research in Health Professions (CEDAR), on behalf of the Dutch Progress test Consortium, Groningen, Netherlands
Wouter Kerdijk (University Medical Center Groningen, Groningen, The Netherlands)
Debbie Jaarsma (University of Groningen and University Medical Center Groningen, Center for Education Development and Research in Health Professions (CEDAR), Groningen, Netherlands)
René A. Tio (University of Groningen and University Medical Center Groningen, Center for Education Development and Research in Health Professions (CEDAR), Groningen, Netherlands)

Background: Over the past decades, many medical curricula were changed to improve students abilities to apply their theoretical knowledge to real patient cases. These curriculum changes were expected to affect students’ knowledge acquisition and the way they apply knowledge. Therefore, we compared students’ knowledge development in four undergraduate medical curricula to better understand the impact of curricular differences on students’ knowledge acquisition.

Summary of Work: Participants were 1440 medical students from four medical schools. To measure students’ knowledge, we used their progress test results from 2008 to 2013. The progress test aims to assess knowledge at end of curriculum level and has the advantage that all students take the same tests four times a year. Questions requiring knowledge application to a case were classified as vignette questions. The other questions were classified as simple questions. We compared the growth of simple and vignette questions using mixed models.

Summary of Results: A cubic growth model best fitted our data for both type of questions. Initially, students scored higher on simple than on vignette questions. The initial growth was faster for simple questions whereas the acceleration over time was slower compared to vignette questions. At the end of the growth curve, the acceleration of simple questions decreased faster. Since curricula were significant predictors of both models, we analyzed growth curves per curriculum. The curricula differed in initial score, growth and acceleration but the trend of the curves was similar to the overall model.

Discussion: The way students acquired and applied their knowledge was similar in all curricula, but the growth was different for each curriculum.

Conclusion: Curricula have an important impact on knowledge acquisition and application.

Take Home Messages: Students’ ability to apply knowledge increases over time, independent of the curriculum.
Background: Progress testing is an assessment of all students in different levels of training at regular intervals. These tests are designed to assess the final curricular competences. For each level, participants are expected to obtain increasing scores according to their acquisition of competences. The objective is to evaluate an OSCE for residents of Family Medicine (FM) used as progress test.

Summary of Work: We implemented the OSCE-progress test in the FM residency in a University Hospital in Buenos Aires. The purpose was to give students feedback, analyze the profile of residents in each dimension, define the progress of the acquisition of competences across different years and design possible remedial plans. Reliability, discrimination rates and performance per year per student were calculated. Satisfaction was evaluated by a 5-point Likert scale.

Summary of Results: In 2015, 16 residents were assessed in a 14-station OSCE. Cronbach’s Alpha coefficient was 0.83. The mean score per year of residency were 72.76 (SD 2.53) for fourth year, 65.05 (SD 2.19) for third year, 55.42 (SD 2.53) for second year and 50.89 (SD 1.96) for first year (p<0.05). Feedback was adjusted to the relative performance of each participant regarding their year of residency. Resident satisfaction grade was 4.3.

Discussion: We consider the OSCE-progress test is a powerful guide to the acquisition of competencies and clinical skills during Residency Programme. The provision of feedback based on its results stimulates reflection and deeper learning.

Conclusion: OSCE-progress test is useful to assess the acquisition of competences during the residency and give constructive feedback.

Take Home Messages: OSCE-progress test reflects the increasing performance for residents during their training.

#8P5 (133774)
Objective Structured Clinical Examination (OSCE) used as progress test in a Residency of Family Medicine

Maria Dolores Arceo*, Instituto Universitario del Hospital Italiano de Buenos Aires, Buenos Aires, Argentina
Nadia Musarella
María Nieves Ganiele
Carlos Brailovsky
Eduardo Durant

Background: Clinicians making decisions require the ability to have accurate self-assessed certainty in being correct. Self-assessment and certainty of performance has been measured in various ways, generally showing that those with less experience and lower performance abilities are less accurate. This study investigates measuring certainty using descriptors authentic to practice, as reflection-in-action, for students with a wide range of experience and ability.

Summary of Work: Year 2-5 medical students (n>1100) gave certainty in their responses to two iterations of a progress test during one calendar year. Analyses compared correctness for response certainty, student year cohort and performance level, defined by criterion scores.

Summary of Results: For both iterations of the test, the number of correct responses increased with student certainty and year group. For any given level of certainty those with more experience and better performance gave more correct responses. Student subsets with less experience and lower performance groups still demonstrated an increase in correctness with increasing certainty. All these results were statistically significant (ANOVA for differences and linearity).

Discussion: Unlike previous work we have shown a degree of accurate self-assessment for those with less experience and lower performance.

Conclusion: We postulate that our finding of similar increases in correctness with increasing certainty, for all groups, relates to certainty descriptors being worded in a way that is authentic to clinical practice, reflection-in-action.

Take Home Messages: 1. Reflection-in-action descriptors are authentic to practice as a better way to document student response certainty 2. Even students in lower year groups and lower performance groups who scored fewer correct in total and fewer correct for any given level of certainty, still demonstrated an increase in correctness with increasing certainty. 3. Using reflection in action descriptors, which are authentic for self-assessment in practice, enable those of lower ability to confirm their appropriate certainty.

#8P6 (132542)
The unskilled are aware: insightfulness when reflecting-in-action

Mike Tweed*, University of Otago Wellington, Wellington, New Zealand
Tim Wilkinson (University of Otago, Christchurch, New Zealand)

Background: Clinicians making decisions require the ability to have accurate self-assessed certainty in being correct. Self-assessment and certainty of performance has been measured in various ways, generally showing that those with less experience and lower performance abilities are less accurate. This study investigates measuring certainty using descriptors authentic to practice, as reflection-in-action, for students with a wide range of experience and ability.

Summary of Work: Year 2-5 medical students (n>1100) gave certainty in their responses to two iterations of a progress test during one calendar year. Analyses compared correctness for response certainty, student year cohort and performance level, defined by criterion scores.

Summary of Results: For both iterations of the test, the number of correct responses increased with student certainty and year group. For any given level of certainty those with more experience and better performance gave more correct responses. Student subsets with less experience and lower performance groups still demonstrated an increase in correctness with increasing certainty. All these results were statistically significant (ANOVA for differences and linearity).

Discussion: Unlike previous work we have shown a degree of accurate self-assessment for those with less experience and lower performance.

Conclusion: We postulate that our finding of similar increases in correctness with increasing certainty, for all groups, relates to certainty descriptors being worded in a way that is authentic to clinical practice, reflection-in-action.

Take Home Messages: 1. Reflection-in-action descriptors are authentic to practice as a better way to document student response certainty 2. Even students in lower year groups and lower performance groups who scored fewer correct in total and fewer correct for any given level of certainty, still demonstrated an increase in correctness with increasing certainty. 3. Using reflection in action descriptors, which are authentic for self-assessment in practice, enable those of lower ability to confirm their appropriate certainty.