**Appendix 1**

**Case Study – Applying Modern Validity Framework**

To illustrate how one might demonstrate various sources of validity evidence when developing an assessment of procedural ability, a case study is provided. In this example, a performance-based examination is being developed to assess Internal Medicine residents’ ability to perform specific procedures. A description of ways to gather evidence to support validity is suggested.

**Examination Purpose**

To assess the skills of first year Internal Medicine residents while performing specific procedures.

**Examination Description**

A formative 6-station hybrid performance-based examination utilizing standardized patients, scripted allied health professionals (confederates), and partial task models.

**Content Evidence**

The examination blueprint was based on the ten procedures that graduates of Canadian Internal Medicine programs are expected to be able to perform. Six procedures were chosen by content experts: lumbar puncture, thoracentesis, paracentesis, knee aspiration, central venous catheter insertion, and advanced cardiac life support.

It was decided that both technical and non-technical skills (including getting patient consent, and collaborating with other team members) would be assessed.

Content experts (i.e., people who regularly perform these procedures and/or assess residents) were asked to develop the cases and scoring instruments. The cases and scoring instruments were then pilot-tested in a mock examination.

Checklists were used to assess key steps in the procedure, and rating scales used to assess the various components of the technical (i.e., preparation, flow of procedure, and adherence to sterile technique), and non-technical (i.e., communication, collaboration and professionalism) skills.

**Response Process Evidence**

Faculty of the Department of Medicine were recruited as physician examiners. They were provided with frame-of-reference training to prepare them for their role as an evaluator for this examination. Two examiners were used in each station and each provided independent ratings.

All instructions and prompts were reviewed with raters and examinees in an orientation session the night of the examination and both groups were provided with an opportunity to ask questions. Staff was available throughout the examination to clarify any concerns.

Standardized patients and allied health professionals were instructed in their roles by experienced trainers, using a standardized script.

All residents were assessed using the same partial-task models that had been used to teach them the skills.

Quality assurance measures were implemented to ensure that data were accurately collected and entered. This included having staff review completed examination sheets between examinees to ensure that raters were filling in all relevant boxes, and that examinee identification stickers were properly adhered to the sheets. During data entry, a second staff member randomly double-checked the accuracy of 20% of the entries.

To minimize the risk of cheating, all examinees were examined in one evening, using simultaneously administered equivalent tracks (circuits).

For each station, the checklist and rating scale components were weighted. Using expert judgment, a focus group of faculty educators assigned relative weights to each component of the rating instruments used.

**Internal Structure Evidence**

After the examination, a G study was used to estimate score reliability and to explore various potential sources of measurement error. Facets included: participant, case, rater, and track.

Item-total-correlations (case-total correlations) were calculated to help identify poor-performing items. Any case with a correlation of < 0.2 underwent revision prior to the next examination.

**Relation to Other Variables Evidence**

A sample of PGY-4 residents was included in the examination to use as a comparison group. Scores were compared by level of training, with the expectation that the more senior residents would score higher than the PGY-1s.

Scores were also compared to a non-procedural performance-based examination, and to a written examination, with the expectation that the correlation would be positive, but only moderate (since the examinations are designed to measure different constructs).

**Consequential Evidence**

For standard-setting, a contrasting-groups method was used to identify those who had achieved mastery of the skill.

Trainees who performed below expectations were required to meet with their Program Director and undergo remedial teaching sessions. A re-take examination was offered the following year.

Procedural complications within the Department were tracked before and after implementing this examination.