Title: Rushing Roulette – how do learners perform routine tasks under time pressure?

Authors: David Topps, Ana Popovic, Teejay Horne, Jean Rawling, Maureen Topps

Institution: University of Calgary, Calgary, AB, Canada

Corresponding Author: David Topps

topps@ucalgary.ca

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1859v1 | CC-BY 4.0 Open Access | rec: 15 Mar 2016, publ: 15 Mar 2016

Abstract

Introduction: Remediating, or preferably, predicting which residents will have difficulty before they need remediating, is a challenging task. Most of us perform better when pumped for an exam. But how do we respond when under routine pressures? Do weaker learners adapt differently, despite coaching?

Methods: Using an adaptation of virtual patient software, we explored how learners cope with handling repetitive yet time-sensitive routine tasks. We emulated the performance of routine tasks within a virtual electronic medical record (EMR) environment, tracking individual learner activity and decision pathways, time to act (with and without enforced pressure from programmed time-outs) and their adaptation trajectories over time with coaching. Learners were assessed using Situational Judgement and modified Script Concordance Testing, with reproducible and granular time constraints introduced into the clinical reasoning process.

Results: Our case designs introduce a number of competing elements: time pressures, competing priorities and instructions, resource availability and unpredictable outcomes. Learner behaviour is assessed using a variety of metrics including time-stamped decision points, decision pathways and internal counter scores. Clinical reasoning pathways, as compared to a reference peer panel, are in turn compared with and without the time pressures.

Conclusions: Predictive analytics have made great promises in diagnosing problems for learners in difficulty but are complex and expensive to deploy widely. Our simpler, rapidly reproducible approach may provide a more practical solution.

Background

Remediation of professional learners, such as medical residents, is a costly and stressful endeavor. An early diagnosis tool is needed to provide earlier warning about those learners more likely to need remediation - many show up too late in their career trajectories.(Audétat et al., 2012) Although nobody wants to hear bad news, most would agree that it is better to provide more gentle correction early on than radical intervention late in residency.

Part of the difficulty here lies in the largely subjective nature of most learner assessments in workplace learning. The predominant modes of formative and summative feedback, such as ITERs, Mini-CEX, are very dependent on the impressions of experienced clinical teachers. However, preceptors hate to be the Judge (David Topps, Evans, Thistlethwaite, Nan Tie, & Ellaway, 2009), (Miller, 2003), (Rees, Knight, & Wilkinson, 2007) but are nonetheless, keen observers.

The biggest problem with which remedial learners struggle is clinical reasoning (Zbieranowski, Takahashi, Verma, & Spadafora, 2013)(Guerrasio, Garrity, & Aagaard, 2014), and yet we employ very few objective techniques that assess this skill. Virtual patients have been shown to be effective tools for this(Gunning & Fors, 2012)(Forsberg, Georg, Ziegert, & Fors, 2011) but they are used predominantly in undergraduate medical education - few residents are exposed to them as an assessment method.

The progression from novice to expert is a developmental trajectory which is not at all consistent across learners. We frequently detect maladaptive developmental patterns in dysfunctional learners but these are sometimes hard to describe precisely.

Environmental stressors can affect clinical reasoning and decision-making both positively and negatively. (Arora et al., 2010)(Farri et al., 2013)(Kamata, 2006) This applies to all of us: novices are affected more than experts,(Hoffman, Aitken, & Duffield, 2009)(Peña, 2010) but might dysfunctional learners/practitioners be impacted further? With sufficient stress, we all face cognitive load breakdown: is this effect more marked in dysfunctional learners, and can it be quantified by measurement of maladaptive behaviors?

We need a more standardized approach to assess such behaviors. We need many objective data points to show not just a learner's answer, but also the reasoning that leads to their answer. We also know that assessment drives behaviour ("Will this be in the test, sir?") - and so we asked how these behaviors change when the activity becomes routine. In particular, we wanted to use Covey quadrant III tasks (Covey, 2004) - those that are urgent but not important – to invoke an environment in which important cues are more likely to be missed. We hypothesized that erroneous decision-making under conditions of excessive cognitive load may represent poorly developed clinical reasoning strategies

Materials & Methods

Research Design

We employed a design-based research approach(Design-Based Research Collective, 2003) because of the highly exploratory pilot nature of this study. We did not have preconceived ideas about how the use of such testing would affect the behaviors captured and the analytics generated from the activities.

"The design-based research methodology is often employed by learning scientists in their inquiries because this methodological framework considers the subject of study to be a complex system involving emergent properties that arise from the interaction of more variables than are initially known to researchers, including variables stemming from the researchers themselves"("Design-based research - Wikipedia," 2015)

Procedure

We used OpenLabyrinth, an open-source educational research platform, that we have previously used extensively for virtual patient authoring and publication.(Ellaway, 2010; David Topps, Ellaway, & Corral, 2015) OpenLabyrinth provided the appropriate web interface and database structure that afforded the creation of several different learning designs. The software has extensive and detailed metrics on how participants respond to questions and cases. It also has exact timestamps on all learner actions, and can provide time-based triggers and navigational pathways.

We created a virtual electronic medical record (EMR) interface, using tools built into OpenLabyrinth so as to provide a data representation with greater verisimilitude. Based on previous projects,(David Topps, 2010) we did not use a real EMR program because it did not provide the flexibility and chronological tracking that we required.

We established a series of 20 case vignettes on our OpenLabyrinth server at http://demo.openlabyrinth.ca/renderLabyrinth/index/578 — the case presentation followed a modified script concordance testing approach.(D Topps, Taenzer, Armson, Carr, & Ellaway, 2014) There is an initial stem, presenting the basic case details, but instead of presenting a pre-formulated hypothesis, we allowed the participants to synthesize their own. They were then presented with additional data in the format of routine findings in the virtual EMR. Normal ranges were available as needed but the participants were required to interpret whether the results presented were normal or abnormal.

In a further modification of the script concordance testing approach, rather than asking whether the new data made their hypothesis more or less likely, we asked them instead what action they would take in this case. There were five possible actions:

- File it
- Let Patient Know (LPK)
- To Come In (TCI)
- Change management
- Skip to next case

We asked them to consider, when making their decision, how important this result was to patient care. They also were asked to evaluate the amount of work required, both of themselves and their virtual clinic staff, to pursue follow-up activities such as making phone calls and arranging appointments. These are typical decisions that are made on a daily basis by a clinician going through their own real investigations and EMR tasks, delegating jobs to real staff.

An automatic lock-out was set at 30 seconds. At this point, if participants had not chosen a course of action they were presented with the next case. We suspended the lock-out clock if participants chose 'Change management', to give them time to briefly describe the change they would recommend. All actions were logged and time-stamped by the OpenLabyrinth database.

The business world calls this an "In-Tray Exercise", which examines worker performance during routine, repetitive tasks.(GILL, 1979) As an additional stressor, we added this test on to a preceding 45-minute computer-based assessment, so that the main participant groups were engaging in this exercise when tired and slightly stressed from the previous exam.

Within the supplementary additional data, we provided some irrelevant normal or subtly abnormal results, similar to those a clinician would encounter when reviewing routine investigations in the workplace. There were 6 minimally relevant or likely irrelevant abnormalities, and 3 more significant abnormalities. Only 1 case had an anomaly significant enough that all the expert panel agreed that the patient should be recalled.

Participants

We had five groups of participants in this pilot program evaluation. Our expert panel of 3 very experienced family physician preceptors performed face validity testing and reviewed the cases for inconsistencies and errors - HA, DT, MT. We recruited a convenient small group of volunteer family medicine residents to perform additional face validity testing and assessment of the virtual EMR interface.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1859v1 | CC-BY 4.0 Open Access | rec: 15 Mar 2016, publ: 15 Mar 2016

We recruited a convenient small group of experienced family medicine faculty to act as a reference panel for initial scoring of the SCT questions and to provide a simple assessment of construct validity.

Our main test group was comprised of IMG participants (n=45) in an orientation period prior to entering Canadian residency training. The participants were aware that their involvement was voluntary.

Our main reference panel for scoring the SCT questions was composed of final-year Canadian medical students (Clerks) preparing for their step one computer-based Canadian medical licensing exam. Participation for this group was also voluntary.

Measures

Previous work (D Topps et al., 2014) demonstrated that clinicians appreciate the avoidance of Single Best Answer questions, as a more realistic representation of some of the subtleties of clinical decision-making. Hence, we initially based our question design on a modified SCT approach, with five reasonable responses. Similarly to SCT, the optimal or most appropriate response related to the proportion of choices made by a reference panel with weighted scoring.

However, it should be noted that, unlike SCT, we did not evenly weight responses around the mean of agreement/disagreement with an action. We scaled these responses, based on some of the inferences of Prospect Theory, wherein decisions under uncertainty yield to Type 1 Thinking and associated heuristics. We considered the trade-offs among immediate effort required, risk of erroneous reasoning, and risk of showing ignorance.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1859v1 | CC-BY 4.0 Open Access | rec: 15 Mar 2016, publ: 15 Mar 2016

Accordingly, we ranked the responses as follows:

File it = result stored in chart with no further action; very little work or impact.

LPK = minimal work for participant or for the virtual staff (simply making a phone call)

TCI = moderate work and risk for candidate; moderate work for virtual staff.

Change management = most work for participant and virtual staff. Short text instruction required.

Skip to next case — Initially, we struggled with how to interpret this response. One might consider that skipping over a decision generated the least amount of work or impact for the participant and virtual staff. However, this response guaranteed a zero mark. Additionally, in the case of a forced lock-out, this conceivably also invoked the most stress for participants due to the admission that they did not know the answer. Ultimately, we determined that this response contributed the most risk (Kahneman & Tversky, 1979) and least predicted value to exam scores.

We also considered that we were dealing with uncertainty and noise in the data to a large extent. In any real set of routine clinical investigations, there is always a certain amount of noise. Experienced clinicians are used to dealing with results containing trivial abnormalities that can be safely disregarded. The trick is to know when such abnormalities are significant or not and this often depends on the context of the case or stem material.

Accordingly, we also examined our responses from the perspective of Signal Detection Theory (Thompson et al., 2008). Based on radio communications, this is the ability to discriminate, under increasingly noisy conditions, between true Signal and Noise (avoidance of

Type II error). We purposely did not employ Item Response Theory because these sophisticated calculations are predicated upon a Single Best Answer.

Results

	Experts (n=8)	Clerks (n=16)	Participants (n=45)
time taken (secs)	575 ± 240.8	541 ± 121.7	604 ± 121.4
nodes visited	41.4 ± 1.3	41.3 ± 2.0	41.2 ± 2.2
Filed	14.1 ± 2.8	2.6 ± 1.9	2.9 ± 1.9
Phoned	6.1 ± 1.1	12.8 ± 2.5	12.3 ± 2.8
Recalled	1 ± 0.0	3.5 ± 3.1	4 ± 3.4
Skipped	0.75	1.8 ± 1.4	1.8 ± 1.6

We noted a higher intervention threshold for learners compared to experts; both learner types were comparable in this regard.

Brief qualitative analysis of feedback and comments from the participants produced one particularly memorable faculty comment “It was dead boring - just like being at work” - which we regarded as being a successful indicator that we were reproducing an environment more akin to routine task handling and not a specific examination situation. Other similar comments were received from our faculty and resident validity testers that it was just like doing tasks on our EMR.

We thought we might see a language effect in the IMGs but there were not enough data points to comment on this.

Using Signal Detection Theory (SDT), we also analyzed our data using Receiver Operator Curves. These tables are quite extensive and we are still working out which scoring approaches provide us with the most valid interpretations. See original data tables stored here(David Topps, Popovic, Horne, Rawling, & Topps, 2015). Assigning different values, based on the effort required according to Prospect Theory, has some interesting effects. For example, with the Skip, it was assigned a value of 0 on the weighting. When put on the effort scale to do the ROC curves, if the correct decision was that nothing had to be done about the case, a skip was considered a correct answer. If the correct decision was to treat the patient in some way, a skip was wrong. ANOVA suggests that there are significant variances between participants that are associated with overall performance.

Discussion

As expected, the degree of intervention demonstrated by learners of both types was much higher than for faculty members. Faculty members were not significantly quicker and there was much larger range of response times than expected, possibly due to concurrent distractions.

With coaching about the expected rates of intervention, and by simply showing learners how their data compared with faculty, we expect to see a drop in intervention levels to more in line with what we see with our faculty members.

The early indications are that this is an acceptable method to learners and faculty. While the exercise was mundane (which was the intent, so as not to prime them for supra-normal performance), it was felt to be quite a realistic simulation of the routine daily task list.

When looking at the SDT data, we can say ROC curves may be an adequate predictor of someone's level of expertise, given from the ANOVA test, since there is a clear difference in the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1859v1 | CC-BY 4.0 Open Access | rec: 15 Mar 2016, publ: 15 Mar 2016

means of the three groups. However, we need to have this data on more learners, in order to follow trends of remedial training, and program withdrawal. We can then potentially zero in on specific trouble nodes, or add/subtract nodes for whatever reason. Given their score on the test, paired with enough evidence from remedial residents, we could calculate a probability that learner X will need remedial training in the future.

There are several possible modifications being explored as we continue with this series: adaptive testing, with variable lock-out intervals; variable complexity of investigation results; timed automated distractors; and variable rates of abnormality prevalence. In the next steps for this ongoing project, we will have a series of tests, built into routine daily workflow, similar to what currently the residents should be doing in clinic.

The 'generated staff workload' may turn out to be a useful metric. Note that for those who are examining the results in detail, there is a compound effect for some actions e.g. The TCI choice generates both a phone call and other actions for the virtual team members. We are considering creating a variable progression rate through the case, forcing the participants to wait a second or two longer for the next data, when they choose a more complex action that generates more virtual work.

The idea of looking at task performance under different levels of stress is by no means new. Even back in 1965, NASA found a Goldilocks effect (Kamata, 2006)- not too little, not too much stress - produced a hump in the performance curve. But there is little work in medical education that examines performance of tasks under routine, repetitive, conditions as seen in the daily grind of work, as distinct from the acute world of the examination.

We generally found that most participants took these cases seriously and treated them as they might their own patients. One participant was clearly simply 'button-pushing' just to finish

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1859v1 | CC-BY 4.0 Open Access | rec: 15 Mar 2016, publ: 15 Mar 2016

the cases as quickly as possible. This individual's data lay so far outside the range of other responses that the scores were removed in this case. 'Button-pushing' might increase as the boredom factor of multiple case runs starts to kick in. We may explore whether to place a 'moral reminder' at start of the exercise. This has the effect of reducing cheating and improving compliance with desired behaviors.(Ariely, 2012)

We have also found that the generation of multiple cases with many, many data points is taking somewhat of a toll on our case authors. We hope to include Answer Data Automatic GEneration (ADAGE)(David Topps, Bennion, Rawling, & Topps, 2016) as a next step, a data table-driven bespoke add-on for OpenLabyrinth that will afford the semi-automatic generation of a much greater number of cases and investigation data points.

We acknowledge the limitations of some of the stats because we are tending to use parametric stats on non-parametric or categorical data. The assumption that the risk judgement is linear and evenly spaced across the SCT scoring is also a weakness of this analysis.

Conclusions

Learning analytics that are predictive of learner performance are much sought after by educational institutions but are difficult to craft and expensive to implement. Few assessments in medical education objectively look at routine work performance outside of the examination setting. This approach was easy to implement and realistic in context. It shows promise for highlighting dysfunctional work patterns earlier in the learner trajectory, possibly affording earlier remediation opportunities. Further trend patterns will be explored as we continue with the series of cases.

Acknowledgements

Rachel Ellaway for help with learning design and data analysis; Michele Cowan for help with case construction; Heather Armson for help with modified SCT designs.

References

Ariely, D. (2012). The honest truth about dishonesty. H. Books, editor. HarperCollins, New York.

Arora, S., Sevdalis, N., Nestel, D., Woloshynowych, M., Darzi, A., & Kneebone, R. (2010). The impact of stress on surgical performance: a systematic review of the literature. *Surgery*, *147*(3), 318–30, 330.e1–6. http://doi.org/10.1016/j.surg.2009.10.007

Audétat, M. C., Dory, V., Nendaz, M., Vanpee, D., Pestiaux, D., Junod Perron, N., & Charlin, B. (2012). What is so difficult about managing clinical reasoning difficulties? *Medical Education*, *46*(2), 216–227. http://doi.org/10.1111/j.1365-2923.2011.04151.x

Covey, S. (2004). *The 7 Habits of Highly Effective People: Powerful Lessons in Personal Change*. Free Press. Retrieved from http://www.amazon.com/The-Habits-Highly-Effective-People/dp/0743269519

Design-based research - Wikipedia. (2015). Retrieved November 12, 2015, from https://en.wikipedia.org/wiki/Design-based_research

Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher*, *32*(1), 5–8.

Ellaway, R. H. (2010). OpenLabyrinth: An abstract pathway-based serious game engine for professional education. In *Digital Information Management (ICDIM), 2010 Fifth International Conference on* (pp. 490–495).

Farri, O., Monsen, K. a., Pakhomov, S. V., Pieczkiewicz, D. S., Speedie, S. M., & Melton, G. B. (2013). Effects of time constraints on clinician-computer interaction: A study on information synthesis from EHR clinical notes. *Journal of Biomedical Informatics*, *46*(6), 1136–1144. http://doi.org/10.1016/j.jbi.2013.08.009

Forsberg, E., Georg, C., Ziegert, K., & Fors, U. (2011). Virtual patients for assessment of clinical reasoning in nursing - A pilot study. *Nurse Education Today*, *31*(8). http://doi.org/10.1016/j.nedt.2010.11.015

GILL, R. W. T. (1979). The in-tray (in-basket) exercise as a measure of management potential. *Journal of Occupational Psychology*, *52*(3), 185–197. http://doi.org/10.1111/j.2044-8325.1979.tb00453.x

Guerrasio, J., Garrity, M. J., & Aagaard, E. M. (2014). Learner deficits and academic outcomes of medical students, residents, fellows, and attending physicians referred to a remediation program, 2006-2012. *Academic Medicine : Journal of the Association of American Medical Colleges*, *89*(2), 352–358. http://doi.org/10.1097/ACM.0000000000000122

Gunning, W. T., & Fors, U. G. H. (2012). Virtual patients for assessment of medical student ability to integrate clinical and laboratory data to develop differential diagnoses: comparison of results of exams with/without time constraints. *Med Teach*, *34*(4), e222–8. http://doi.org/10.3109/0142159X.2012.642830

Hoffman, K. a., Aitken, L. M., & Duffield, C. (2009). A comparison of novice and expert nurses'

cue collection during clinical decision-making: Verbal protocol analysis. *International Journal of Nursing Studies*, *46*(10), 1335–1344. http://doi.org/10.1016/j.ijnurstu.2009.04.001

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk on JSTOR. http://doi.org/10.2307/1914185

Kamata, E. S. (2006). *Influence of Psychological Factors on Product Development: Lessons from Aerospace and other Industries*. Springer Science & Business Media. Retrieved from https://books.google.com/books?id=qRTnBwAAQBAJ&pgis=1

Miller, P. J. (2003). The Effect of Scoring Criteria Specificity on Peer and Self-assessment. *Assessment & Evaluation in Higher Education*, *28*(789284620), 383–394. http://doi.org/10.1080/0260293032000066218

Peña, A. (2010). The Dreyfus model of clinical problem-solving skills acquisition: a critical perspective. *Medical Education Online*, *15*, 1–11. http://doi.org/10.3402/meo.v15i0.4846

Rees, C. E., Knight, L. V., & Wilkinson, C. E. (2007). Doctors being up there and we being down here: A metaphorical analysis of talk about student/doctor-patient relationships. *Social Science and Medicine*, *65*(4), 725–737. http://doi.org/10.1016/j.socscimed.2007.03.044

Thompson, C., Dalgleish, L., Bucknall, T., Estabrooks, C., Hutchinson, A. M., Fraser, K., … Saunders, J. (2008). The effects of time pressure and experience on nurses' risk assessment decisions: a signal detection analysis. *Nursing Research*, *57*(5), 302–311. http://doi.org/10.1097/01.NNR.0000313504.37970.f9

Topps, D. (2010). Notes on an EMR for Learners. Calgary, AB, Canada: ResearchGate. http://doi.org/10.13140/RG.2.1.5064.6484

Topps, D., Bennion, L., Rawling, J., & Topps, M. (2016). ADAGE: Answer Date Automatic GEneration | OpenLabyrinth. Retrieved March 13, 2016, from http://openlabyrinth.ca/adage-answer-date-automatic-generation/

Topps, D., Ellaway, R., & Corral, J. (2015). OpenLabyrinth web site. http://doi.org/http://www.webcitation.org/6a3ZiWGrb

Topps, D., Evans, R. J., Thistlethwaite, J. E., Nan Tie, R., & Ellaway, R. H. (2009). The one minute mentor: a pilot study assessing medical students' and residents' professional behaviours through recordings of clinical preceptors' immediate feedback. *Educ Health (Abingdon)*, *22*(1), 189. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19953438

Topps, D., Popovic, A., Horne, T., Rawling, J., & Topps, M. (2015). Signal Detection Theory data for Roulette cases | OpenLabyrinth. Retrieved March 13, 2016, from http://openlabyrinth.ca/signal-detection-theory-data-for-roulette-cases/

Topps, D., Taenzer, P., Armson, H., Carr, E., & Ellaway, R. (2014). *DynIA: Dynamically Informed Allegories. (final report)*. Calgary, AB, Canada. Retrieved from http://dspace.ucalgary.ca/handle/1880/50360

Zbieranowski, I., Takahashi, S. G., Verma, S., & Spadafora, S. M. (2013). Remediation of

residents in difficulty: a retrospective 10-year review of the experience of a postgraduate board of examiners. *Academic Medicine : Journal of the Association of American Medical Colleges*, *88*(1), 111–6. http://doi.org/10.1097/ACM.0b013e3182764cb6