Clinical Trainee Performance on Task-Based AR/VR-Guided Surgical Simulation is Correlated with their 3D Image Spatial Reasoning Scores

Roy Eagleson,[a]* Denis Kikinov,[a] Liam Bilbie,[a] & Sandrine de Ribaupierre[b]

[a]*Software Engineering, University of Western Ontario, London, Canada;* [b]*Clinical Neurological Sciences Neurosurgery, University of Western Ontario, London, Canada*

eagleson@uwo.ca

Provide short biographical notes on all contributors here if the journal requires them.

Clinical Trainee Performance on Task-Based AR/VR-Guided Surgical Simulation is Correlated with their 3D Image Spatial Reasoning Scores

This paper describes a methodology for the assessment of simulator-based trainee skills on an AR/VR-guided procedure making use of CT axial slice views for a neurosurgical procedure: External Ventricular Drain (EVD) placement. The task requires that trainees scroll through the stack of axial slices, and form a mental representation of the anatomical structures in order to subsequently target the ventricles to insert an EVD. The process of observing the 2D CT image slices in order to build a mental representation of the 3D anatomical structures is the perceptual aspect of skill being taught, and as a result the cognitive control of the subsequent targeting, by planned motor actions, of the EVD tip to the ventricular system to drain cerebrospinal fluid (CSF). We establish convergence towards the validity of our assessment methodology by examining two objective measures of spatial reasoning, along with one subjective expert ranking methodology, and comparing these to AR/VR guidance. These measures have two components, both speed and accuracy of the targeting, which are used derive the performance metric. Results of these correlations are presented for a population of PGY1 residents attending the Canadian Neurosurgical 'Rookie Bootcamp' in 2019.

Keywords: surgical simulation; AR/VR task performance evaluation; neurosurgery resident training

Introduction

At one level, this paper describes experiences and lessons learned while conducting a Neurosurgical Training `bootcamp' (see fig.1) at CSTAR, in Canada in 2019 and 2022. This paper also includes a description of the overall course objectives and descriptions of the kinds of training techniques used at each of the AR/VR-based training stations that were scheduled for the 53 trainees. This paper also describes the kinds of simulations, with particular emphasis on the methodologies used for gathering data, the systematic evaluative procedures in place at each station, and the results analysed from

these data; gathered from 53 PGY1 resident trainees (fig 1).

Moreover, at another more philosophical level (more interesting for the AECAI audience) this paper carries a narrative and sets up an intial salvo that opens opportunities for discussion at the workshop about what should be methodologies for evaluating the effectiveness of simulation-based training involving Augmented Environments for Computer-Assisted Interventions.

Figure 1. 25 PGY1 residents attended the Canadian Neurosurgical `Rookie Bootcamp' in 2019 (and 28 in 2022), hosted at the Canadian Surgical Technologies and Advanced Robotics Centre (CSTAR).

Canadian Neurosurgery PGY1 Bootcamp Training

The Canadian Neurosurgery Society has sponsored, in collaboration with the Neurosurgery Specialty Committee of the Royal College of Physicians and Surgeons, since 2013 (cf. Haji, Clarke, O'Kelly; 2013, 2015), an intensive training camp for first

year post-graduate residents (PGY1). The ratio of instructors to trainees is approximately 1:2, and the trainees can expect to experience a full gamut of training experiences ranging from didactic to pragmatic; from lectures to hands-on low-fidelity box simulators.

In this paper, we will consider a subset of these training exercises -- ones which involve AR/VR-based simulator training scenarios. In particular, these will involve Extra-Ventricular Drain (EVD) placement; or alternately, Endoscopic Third Ventriculostomy (ETV). Both of these procedures have a main surgical procedure phase which involves the targeting of brain ventricles.

Figure 2. Surgical Simulators can have varied 'realism', to be sure. This one, in particular, is extremely realistic, but has no intrinsic curriculum

The Red Herring: Training Simulator Realism and Face Validity

When considering the evaluation of a surgical simulator, one often considers the amount of `realism' that can be designed for the simulation. This paper will argue that this is a *red herring*, and in fact, it is generally confused with `Face Validity' (cf. Dahlstrom, 2008). For example, if you present a realistic wound to a trainee, without a task, and therefore without a test of performance -- then its ``face validity'' is zero. Face validity means: `on the face of it,' (i.e. ***prima facia***): would an expert consider the *testing evaluation* to be a reasonable one?

Consider an AR/VR-based simulator, which displays realistic views of a simulated neurosurgical case. Without a test, `realism' has nothing to do with the `face validity' of that simulator (cf. Nguyen, Eagleson, Boulton, de Ribaupierre, 2014). Accordingly, with a `fully realistic' VR environment, a trainee might just as well be facing a *cadaver head* without any instructions, if they have no curriculum. Furthermore, `face validity' estimates are based on e*valuation metrics* for explicit tasks chosen from the curriculum. To make the point sharper: Do the measures of performance on the simulator seem to an expert to be reasonable measures of their performance, as related to the real task?

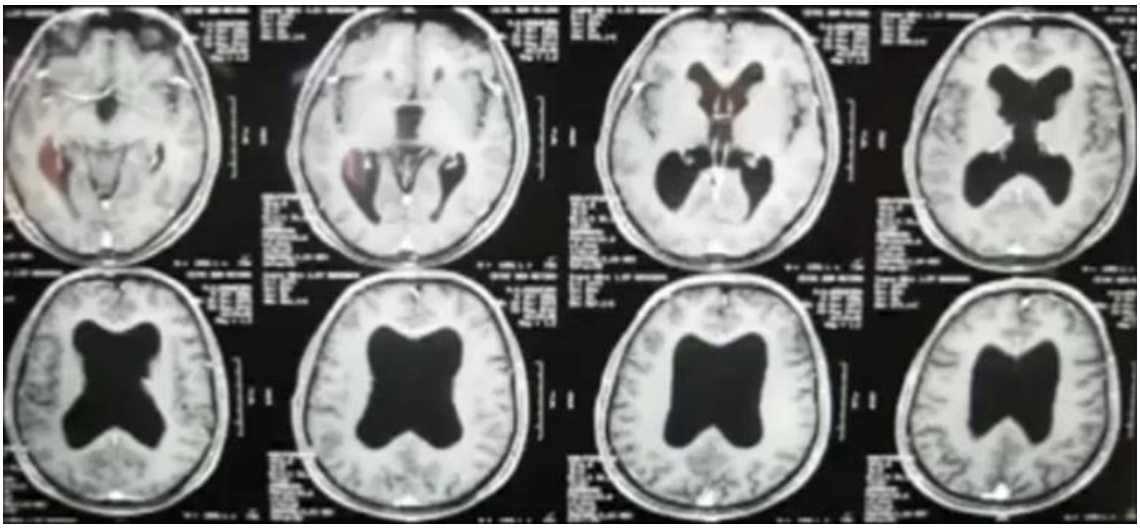

Figure 3. The skilled task being trained is to view a set of CT slices, and then on the basis of reasoning about those views, to plan a point of entry on the skull and then introduce the EVD into the ventricular system.

The Real Problem: How can we move towards integrating the surgical 'process model' with the 'evaluation methodology'

We will be examining a set of procedures which involve targeting neuroanatomical structures within the brain. The trainees will be performing this after viewing the patients preoperative scans (Figure 3). After reviewing them, the must form a mental representation of the target within the patient. The trainees will be required to learn how to transform that 2D information into 3D, in order to visualize a location for entry (on the surface of the skull), along with a 3D direction vector and a depth of entry into the brain.

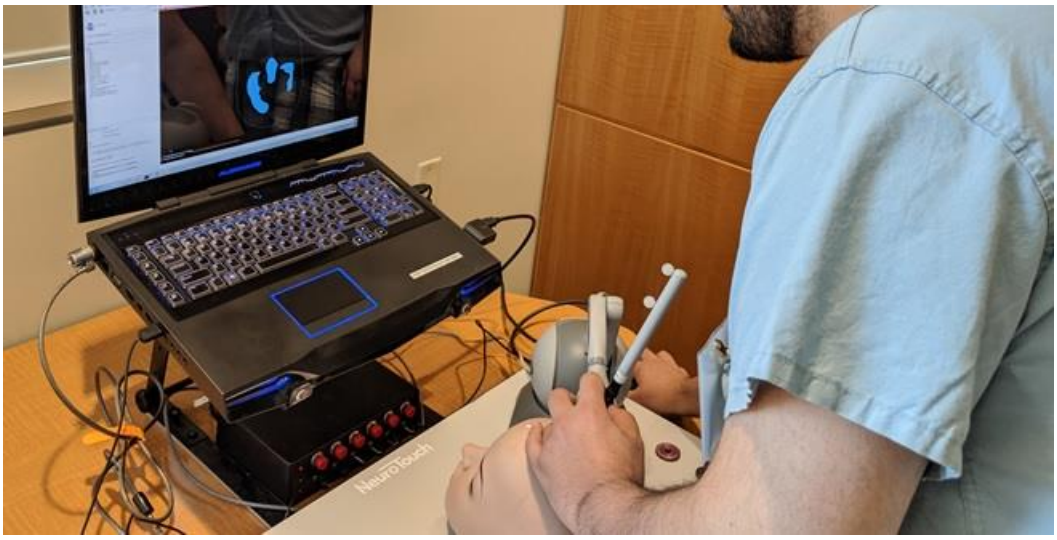

Figure 4. Once the trainee has reasoned about the point of entry and the location of the target, they must move the surgical tool through their estimated 3D entry position, along an estimated 3D orientation and depth.

This central phase of the surgical procedure is quite well-posed, and so measures of the 3D position, the 3D orientation, and the depth of entry -- are all well specified quantities which can be gathered on time-stamped log file entries. These log files can be examined line-by-line in order to form assessments which facilitate inter- and intra-trainee evaluations. Individuals can be compared across their training group, and can be compared to data gathered from expert neurosurgeons.

Estimates of Performance in 3D Localization

Consider figure 5 which shows a number of measures proposed from the literature which can be used to estimate the accuracy of ventricle targeting: (a) 'Engagement', (b) MDM (distance between point and closest ventricle wall), (c) angulation in the sagittal plane, (d) angulation in the coronal plane. We take a moment here to justify why "Engagement' may be a very appropriate measure for EVD placement performance, and relate it to the classical Euclidean distance metric.

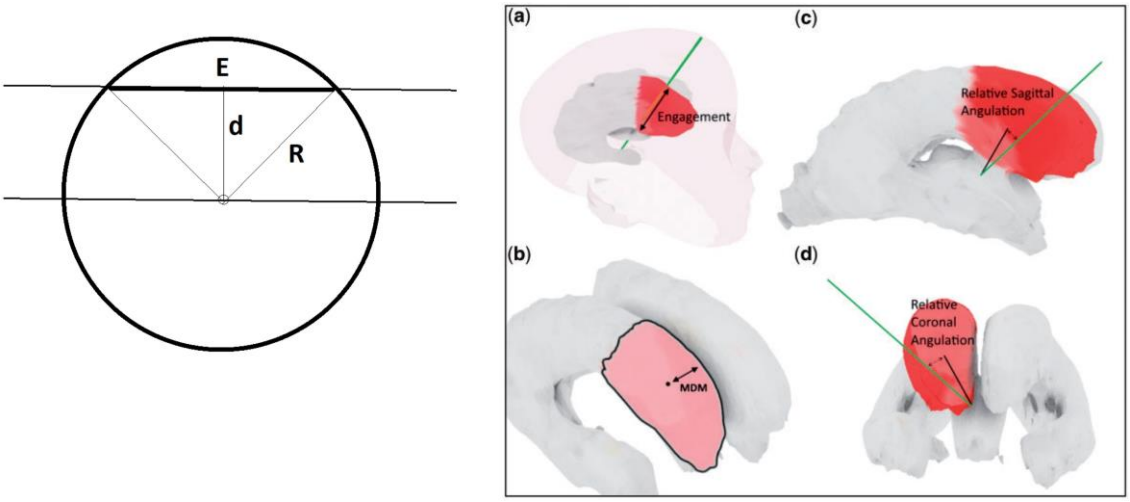

Figure 5. In the literature, a number of measures have been proposed as objective measures of accuracy in targeting a ventricle. We convert the 'Engagement' measure into a distance metric.

In principle, we would like to construct a metric which is such that if the surgeon places the catheter anywhere inside the ventricle, that the task has been performed with 'utility' (ie. in a binary go/no-go sense) – yet questions can still arise about a quantitative score to be attributed to the procedure. Consider the case where the ventricle has been missed: heuristically, being far away from the ventricle should be associated with an error measure, and so the Euclidean distance metric (distance from the centre of the ventricle) is an appropriate measure. However; inside the ventricle, the metric should be 'close to zero' anywhere in the centre of the ventricle, but smoothly increasing as approaching the ventricle wall. We will now extend this heuristic to mathematical formalism as follows: Consider 'E' to be the Engagement, and so if the target volume was a sphere of radius R, we could use E to form a very well-formed distance metric. Through simple geometry, since in the left diagram in figure 5,

$R^2 = (E/2)^2 + d^2$, we have the following expression $D = R - \sqrt{R^2 - d^2}$

This distance metric has very interesting and satisfying properties. By observation, when the trajectory is through the centre, giving maximal 'engagement', then D=0 (perfect accuracy). As d increases by a small amount, D increases by a much smaller amount, corresponding to the interpretation, ``if the user has targeted in the vicinity of the central part of the volume, then the distance D is very close to zero". As d is increased more and more, towards R, then the distance metric D begins to increase rapidly towards R (in fact, 1-D/R is similar to the Lorentz transformation).

When d=R, then D=R, but then subsequently, as the targeting `just misses' the ventricle, but is just outside the ventricle wall, then the distance metric D switches to an imaginary number, which serves as a 'flag' to indicate that the targeting is *outside* the ventricle, at a distance greater than R from the centre. Mathematically, just outside the ventricle wall, D= iR. Then, as we consider what happens as d increases to a `wide miss' of the ventricular system, so that d is much larger than R, then the distance metric D approaches that value, as an imaginary number: $D \rightarrow id$. The absolute value of D increases monotonically (which is a necessary property of a distance metric).

To summarize, D is close to zero anywhere in the vicinity of the central region of the volume, D is most sensitive to changes of position at the edge of the volume, whereupon D=R when d=R, and then, D is approximately d, far away from the volume, and in all of these cases, D increases monotonically as the classical Euclidean distance. Accordingly, one can see that this is an objective and well-posed distance error metric.

Relations between Objective Metrics and Expert Rankings

In this section, we will expose the numerical relationship between the rankings assigned by clinical experts (Neurosurgery Consultants) who are training PGY1

residents and the mathematically well-posed distance metric. A consultant Neurosurgeon was asked to rank the ventricle targeting on a scale of 1 to 6, in the same spirit as the mathematical distance metric derived purely geometrically, a subjective ranking of targeting performance was assigned to each targeting trial by a consulting neurosurgeon.

The subjective scores were assigned such that, `if the targeting was placed well within the ventricle, with good approach angle', then the score given was 1. If the targeting was in the ventricle but at a poor angle, or inside but close to the ventricle wall, then the score was 2. If the targeting was slightly outside the ventricle, then the score was 3. If the targeting was outside the ventricle but with a poor approach angle, the score was 4. And finally, if the targeting was a wide miss, the score was 5, or 6 depending on the whether the approach angle was good or not.

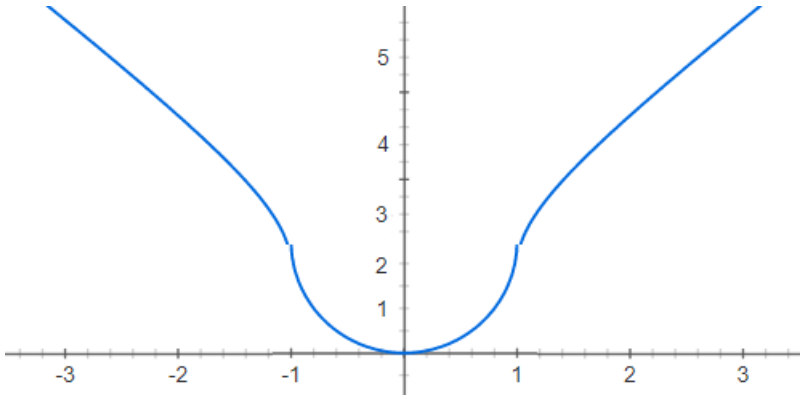

Figure 6. Functional relationship between normalized objective Distance metric D (domain), and Expert Ranking ‘Index of Accuracy’ (range), illustrating the central Lorentz transformation within the sphere of radius=1.

Just as reading a length measure from a ruler can be argued as being either ‘subjective’ or ‘objective’, depending on how much trust is placed on the observer, then so can an expert’s ranking be considered as objective, so long as there can be a principled

functional relationship from their scores, to a metric function. A set of sample expert-assigned subjective scores are illustrated in fig. 7, and full results and analysis are in the Results section.

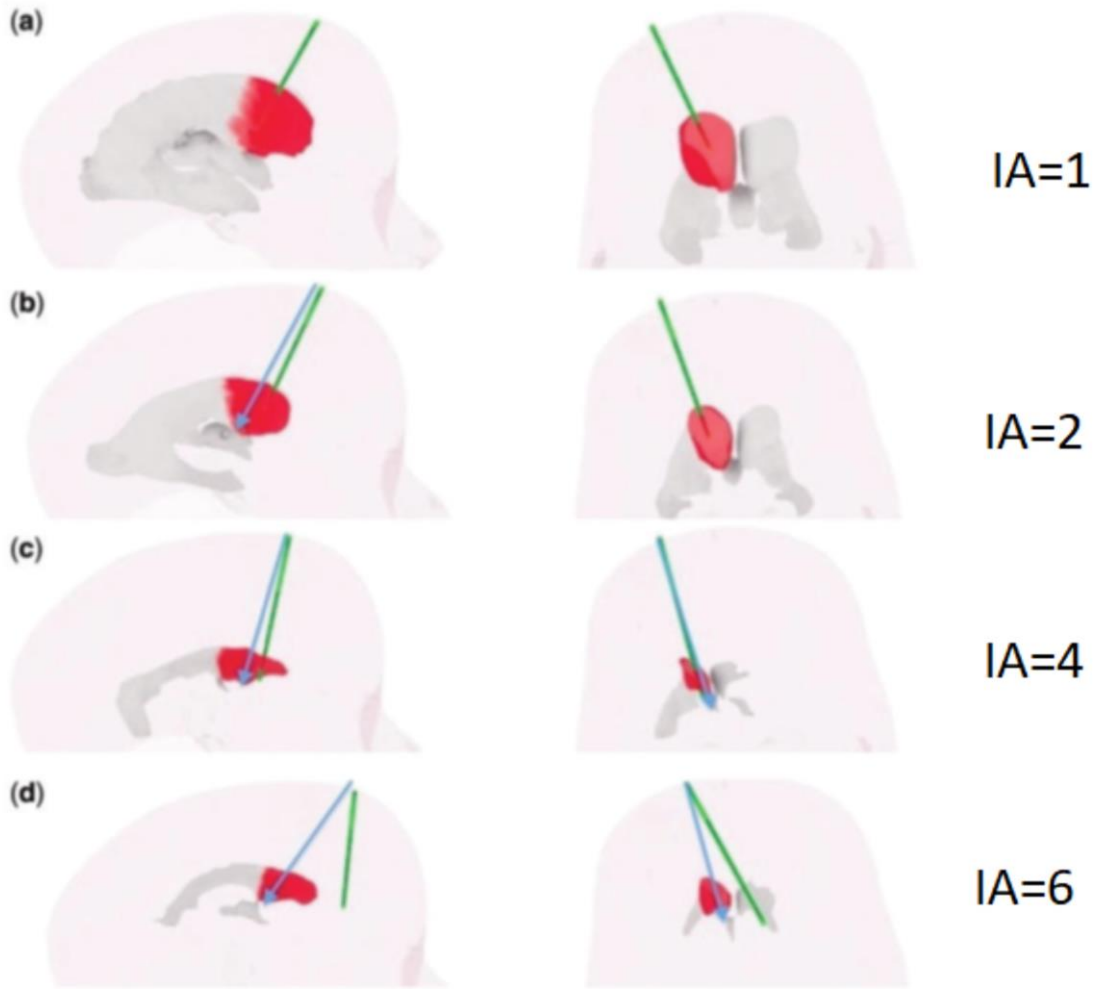

Figure 7. After each trial, for each participant, the targeting of the randomly presented ventricular system is displayed as a green line as feedback to the participant (with blue line showing an ideal trajectory). A screenshot of each trial is saved, and subsequently ranked by an expert neurosurgeon. Sample rankings of the Index of Accuracy (IA) assigned by the neurosurgical expert are shown.

Purely geometrical analysis can lead to the implementation of purely obective metrics. However, when considering patient-specific anatomical structures, objective metrics can be infeasible to formulate or compute. Fortunately, subjective metrics,

which can be assigned by experts, can be articulated in a way which maintains consistency with the well-formed geometrical objective metrics. This consideration will then lead us to a broader and more over-arching discussion about the parallels between Objective Metrics and Subjective Scores as we consider the unavoidable trade-off between ``Internal Validity vs External Validity".

Evaluation of Baseline Spatial Reasoning Skills of Participants

Our over-arching commentary in this paper is intended to prompt discussion about whether the evaluation of the performance of a participant on a simulator should follow the model of expert rankings and subjective assessments -- or whether it should be modeled more on the paradigms of Experimental Psychology and Cognitive Science. In the sequel, we will argue that ``construct validity" can only be attained asymptotically after effortful iterative convergence between both approaches. To begin, we examine a classical methodology for establishing a baseline of user performance for Spatial Reasoning.

3D Spatial Reasoning from clinical 2D Images

It is not controversial to state that the task of observing 2D slices of neuroanatomical structures, and developing a 3D internal respresentation of those structures, forms the foundation for the planning phase of this procedure. The internal representation allows the surgeon to form a plan for burr hole placement, and to move in a way which allows them to introduce the EVD from that location on the skull, in an angle that will allow the tool to intersect the ventricular system.

What is controversial, however, is the nature of this skill: From an information-processing perspective, how are the input images being transformed into a motor plan and an action. It is this skill that is the *construct* we are trying to measure. Accordingly,

this consideration leads to the question: Is this form of spatial reasoning a bottom-up process? (learnable by trial and error, analogous to some deep-learning prescription) or is it a top-down cognitive process that can be learned through instruction and debriefing? To address this question, we make use of our Evaluation methodology to explore the nature of spatial reasoning, within the context of what some would call “mental rotation”.

Objective Methodology for evaluating 3D Spatial Reasoning from 2D Images

We have implemented a computer-based test that makes use of the same set of stimuli introduced by Roger Shepard and Jacqueline Metzler in 1971. For each trial, participants view a pair of randomly-selected images which are 2D presentation of 3D objects, which are simple block shapes. The pairs are balanced so that half of the time the items are "Same" and the other half of the trials are "Different". The participants press a key 's' or 'd', and the response is recorded in addition to the response time.

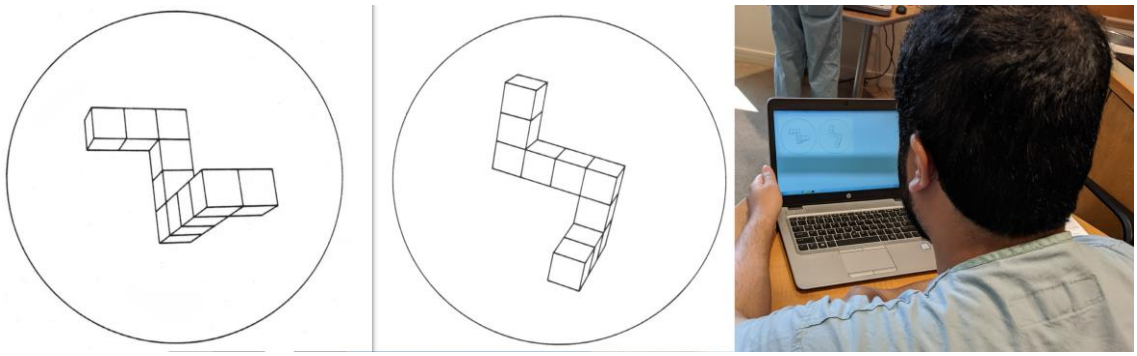

Figure 8. After each session on the EVD simulator, participants were tested using the stimulii from Shepard and Metzler's classical task.

Results and Analysis

We begin with the analysis of the results of the test of baseline 3D spatial reasoning from 2D images. The table on the left of figure 8 show the number of errors made for each participant, along with the mean time to report 'Same' correctly, and the mean time

to report 'Different' correctly. Across all trials, some participants are very accurate, but tend to have longer task time, while some participants are faster (shorter task time), but tend to have more errors.

# errs	Tsame	Tdiff
3	6763	9696.615
6	9381.846	16486.5
8	6833.111	5271.9
9	7786	9689
3	10676.67	13252.75
1	21611.85	27442
3	11417.18	19908.69
2	7761.5	13318
1	10116.46	11864.69
2	7670	8125.857
0	18050.62	22858.43
2	6807	15003.92
3	5231.833	9329.25
12	4222.429	6542.25
2	7831.385	6801.25
3	17216	17190.82
3	4925.9	8168.071
2	6414.538	14173.08
4	3422.167	4495.909
3	5323.083	8182.583
1	5724.333	6681.357
5	9343.333	12197.9
1	3902.077	5230.077
2	9468	11163.86
0	10414.46	11783.36
1	8657.667	15901.79

Figure 9. By comparing the raw Response Time with the number of error trials, we can demonstrate a classical speed-accuracy tradeoff across the population of participants.

We can also make use of this classical paradigm to test a hypothesis about whether the task of performing a 3D task on the basis of 2D images is cognitive; involving top-down spatial reasoning (an alternative hypothesis to what Shepard and Metzler has originally proposed). Many researchers propose that the brain has a mysterious functionality that allows 2D/3D `analogical' representations to be rotated 'in the head'. The postulated theory says something to the effect of seeing two objects, and somehow `rotating' them so that they overlap, and then checking to see if they are same or different. One problem with this theory is that no suggestion has been made as to *which direction*

should the 3D representation be rotated, in order to check (and you would first need to know in which way they were different, in order to know which way to rotate them, to check. “so, as they say… Catch-22”). Aside from that, this analogical theory would predict that once the check has been made, the response at that time would either be same or different, and so the amount of time needed to say `same' or `different' would be identical.

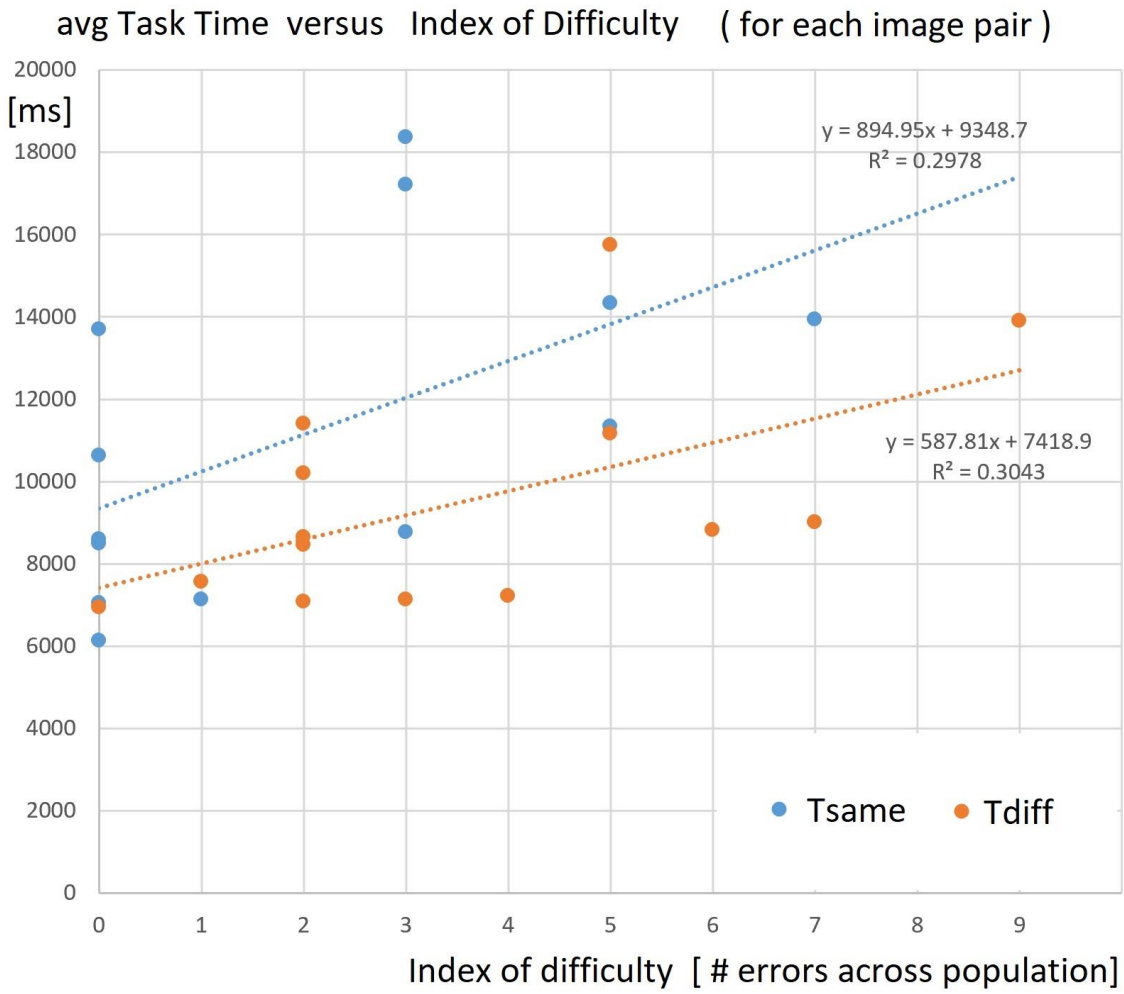

Figure 10. As a function of the Index of Difficulty of the stimulus pair, there is a significant difference between `same' and `different' trials, which is inconsistent with the classical `Mental Rotation' interpretation; a top-down Cognitive Spatial Reasoning interpretation is implicated.

Conversely, for the top-down spatial reasoning account, participants would examine the two side-by-side images that represent 3D structures. On this basis, they would deductively (top-down) check parts of the structures to see if there is any reason that they are different. Notice, that once there is any evidence that the shapes are different, then the participant can respond `different,' whereas if the two objects are the same, then the participant must terminate the checking process on their own. In other words, at some point they would realize that with no evidence for a difference that the objects must be `same.' Accordingly, this alternate hypothesis would predict that the reaction time to say `same' would be longer than `different'.

The results in figure 10 show the reaction time, averaged across all participants, as a function of a raw measure of the Index of Difficulty for each trial pair. (The raw measure is simply the number of errors made across the population of participants -- some image pairs were never mistaken by any of the participants, and some image pairs are so difficult that as many as 9 participants judged them as same/different incorrectly). This figure shows that the reaction time is indeed a function of the Index of Difficulty, as is to be expected. But in addition, we see that the correlation is different for same versus different pairs! Saying `Different' is on average at least 2 seconds different from saying `Same'. One observation afforded by this analysis is the response bias among participants: 53 times they say `same' but the pairs are different, and 29 times they say `different' when the pairs are same. In other words, participants are biased to say, "same", even though the stimulus pairs are balanced evenly. Nevertheless, the difference shown here is problematic for an account based on a putative `Mental Rotation' capacity, and it provides more converging evidence for a top-down Cognitive

account of Spatial Reasoning skills. This will have implications for the way that we recommend training surgical skills for PGY1 residents.

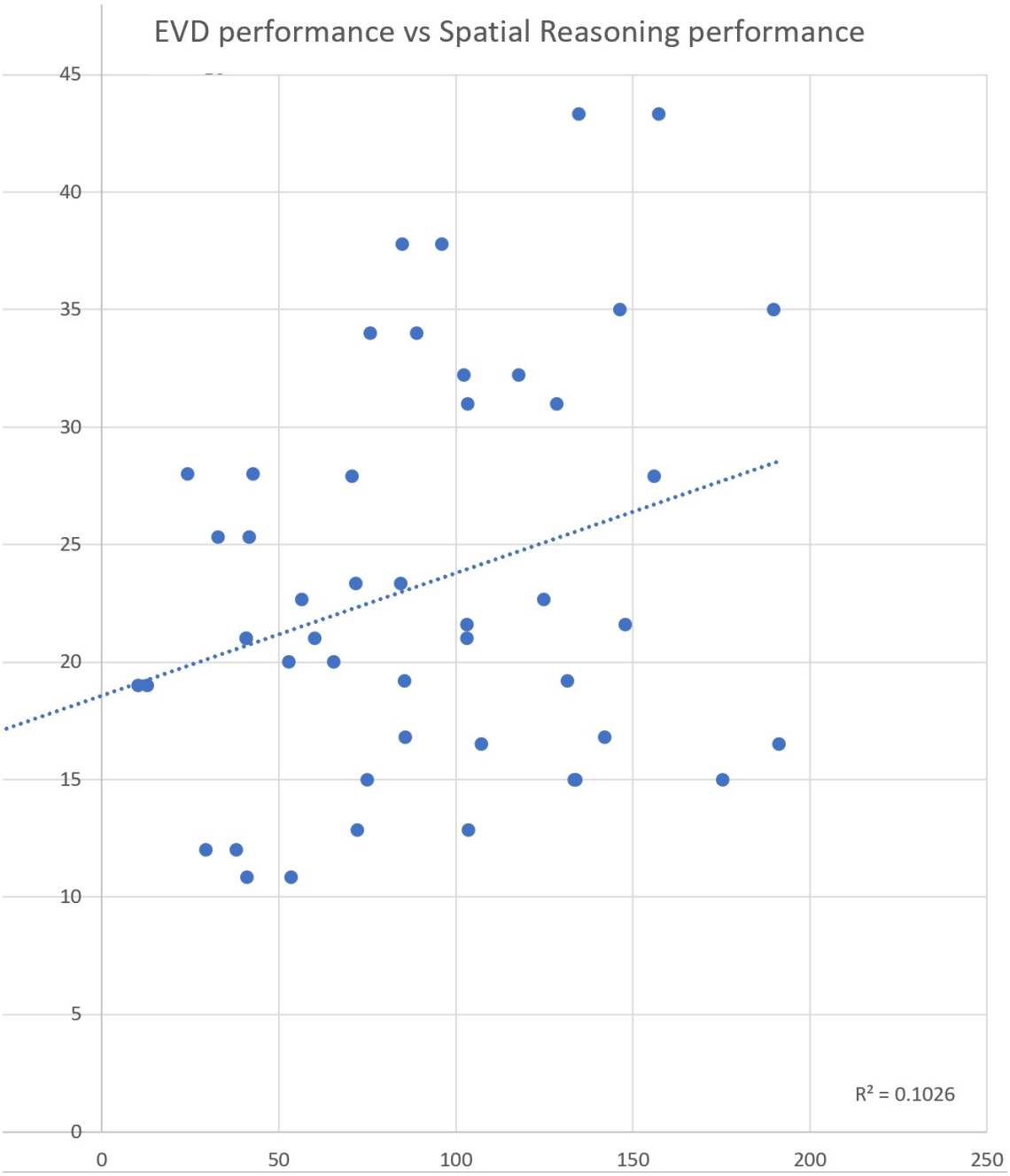

Figure 11. The scores on EVD performance are correlated with the participants' objective scores on the Spatial Reasoning task.

One motivation for studying user behaviour from a psychophysical perspective is that it can allow a focus on the modes of teaching and learning that are most effective. For

example, if from a cognitive science perspective, the behaviour of the user is modeled in a top-down fashion. Accordingly, the teaching and learning mode should be one of didactic training, de-briefing, and mentorship (rather than low-level bottom-up repetative training and perceptual-motor adaptation).

Summary of this Training Simulator Construct Tested: `EVD Trajectory Planning Skill'

There is no broader division in our literature, for approaches to clinical skills training and assessment, than the divide between approaches which prescribe the use of subjective scores, versus objective metrics. Furthermore, because earlier sections of this paper have raised a discussion point about the use of the terms `face validity', `content validity,' and the *elusive* `construct validity,' we offer a controversial point which might help to refocus efforts; at least for discussion amongst participants in this AECAI workshop.

First, on the one hand, `objective metrics' seem to be desirable because, if well-formulated, remove the ubiquitous cognitive biases which are typically manifest for both the experimenter, and the subject of the experiment. On the other hand, in order to formulate an objective metric, many ideological assumptions inevitably must be asserted -- so many it would seem, that the desire to control the parameters of the simulated environment lead to an explosion of experimental conditions that make their presentation infeasible (this is the burden that Psychophysical investigations bear, in the classical literature of Experimental Psychology and Cognitive Science).

In contrast, when trying to assess performance of skill on a reasonably realistic variety of possible scenarios, then the assessment of performance is vulnerable to the ill-posed nature of such scenarios -- in the sense that there may be *several* possible

reasonable actions. In such cases, the only recourse for evaluation is to appeal to the subjective scoring provided by a domain expert. (cf. Neumuth, Loebe, Jannin, 2012).

It doesn't seem to stretch the imagination of a reasonable researcher, to propose that the distinction between 'objective metrics' and `subjective scores' is *exactly* the distinction between the irreconcilable difference in the Validity literature (ie. Cronbach and Meehl, 1955) between `Internal Validity' and `External Validity'. As such, we must face an inevitable trade-off, which has always been acknowledged in Experimental Psychology and Cognitive Science: The more Internal Validity `baked' into your experimental paradignm, the more objective can be your measures. And, on the other hand, the more `External Validity' attributed to your experimental paradigm, the more inevitability will be the need to recourse to subjective scores. Accordingly, we have an inescapable trade-off that needs to be acknowledged face-on. *You cannot have both!* -- your experimental paradigm cannot feasibly have both `high internal validity' and `high external validity'. Your AR-VR-based simulator cannot be ‘anatomically faithful’ and at the same time control for the index of difficulty induced by ‘anatomical variations’ across your training set.

In order to try to extablish ``Construct Validity" there must be iterations between the two paradigmatic styles (not to mention, the often missed step of first identifying what you mean by your `construct' before you try to validate whether your measures are able to provide evidence of such). Iterations between these two styles of experimental paradigm are necessary if they are to establish converging evidence that they are valid measures of the elusive `construct'. Only through effortful convergence and iterative design and evaluation cycles can the holy grail of `construct validity' be established. (you certainly can't establish `construct validity' by showing that your scores corrlate

with the PGY rankings of your participant pool. That is merely `Criterion Validity', and should be acknowledged as such: an initial indicator that shows a rough (but necessary) sensitivity of your measures).

Conclusions and Discussion

Our main intent for this paper has been to foster a renewed discussion about the importance of the domain of the ``Evaluation of Clinical Simulators for Procedure Training``. Our first comment on this topic, which will not be controversial, is that the evaluations should be as objective as possible -- and therefore metrics of performance need to be formulated according to effortful and descriptive information-processing models of the task or procedure that is being trained within a clinical curriculum. While that may seem trite and obvious, many reports in the literature ignore this. Furthermore, by adopting this as a principle, one can still be lead to some difficult questions: For example, if your trainees are performing a procedure in which the outcome does not depend on `path length', then why would you propose measuring 'path length' just because another study had measured it as an explicit constraint on their task? Likewise, if the performance on a procedure does not depend on the force with which you grasp a tool, then why include that as a measure of performance on the task? If your performance on the task can be measured on the basis of the speed and accuracy of performance of the task, then why measure the characteristics of the heartbeat or the conductance of the sweaty hands of the participant? Or the `hot spots' of their eyemovements?

To be sure, each of these measures may (if you are lucky) correlate with the clinician's overall performance; but so might their performance correlate with the proportion of grey hairs on their head, or the price of their car in the parking garage. Each of these measures will probably correlate with performance. But at the end of the

day, the only real measure of performance will be to decompose the procedure into phases, and the phases into tasks, and the tasks into sub-tasks, until the leaf nodes are either `targeting' tasks, or `decision-theoretic choices'. At that point, each phase can either be evaluated objectively in terms of the speed and accuracy, or (inversely) the task time and the error rate. Yes, now *those* are controversial stances -- so let's discuss.

Bibliographic References

Cronbach, LJ, and Meehl, PE (1955) "Construct validity in psychological tests" Psychol. Bull., 1955, 52, 281-302.

Cronbach, Lee J, and Shavelson, Richard (2004) My Current Thoughts on Coefficient Alpha and Successor Procedures. Educational and Psychological Measurement. 64(3):391-418.

Dahlstrom,N, Dekker,S, van Winsen,R, and Nyce,N (2008) ``Fidelity and validity of simulator training" Theoretical Issues in Ergonomics Science. V.10(4), pp. 305-314

Faizal A. Haji, David B. Clarke, and Cian O'Kelly (2013) `How to build a national boot camp: Experience from the Canadian Neurosurgical Rookie Camp'. 2013 International Conference on Residency Education (ICRE)

Faizal A. Haji, David B. Clarke, Marie C. Matte, David M. Brandman, Susan Brien, Sandrine de Ribaupierre, Cian O'Kelly, Sean Christie, Patrick J. McDonald, Abhaya V. Kulkarni, Simon Walling and Anna MacLeod (2015)

``Teaching for the Transition: the Canadian PGY-1 Neurosurgery Rookie Camp"
The Canadian Journal of Neurological Sciences, v.42: 25-33.

Neumuth,T, Loebe,F, Jannin,P (2012) "Similarity Metrics for Surgical Process Models"
Artificial Intelligence in Medicine, v.54(1).

Nguyen,N, Eagleson,R, Boulton,M, de Ribaupierre,S (2014) Realism, criterion validity, and training capability of simulated diagnostic cerebral angiography. Studies in Health Technology Informatics, v.196, pp.297-303.

Roger Shepard and Jacqueline Metzler (1971) ``Mental Rotation of Three-Dimensional Objects" Science, v.171, issue 3972, pp.701-703.