Overall Proficiency Assessment in Point-of-Care Ultrasound Interventions: The Stopwatch is not enough

Matthew S. Holden, Zsuzsanna Keri, Tamas Ungi, Gabor Fichtinger

School of Computing, Queen's University, Kingston, Canada
72mh@queensu.ca

Abstract. With the shift in the medical education curriculum to a competency-based model, objective proficiency assessment is necessary. In this work, we use exploratory factor analysis to assess which primitive metrics convey unique information about proficiency in point-of-care ultrasound applications. We retrospectively validate the proposed methods on three datasets: FAST examination, femoral line, and lumbar puncture. We identify a minimal set of metrics for proficiency assessment in each application. Furthermore, we validate that overall proficiency assessment methods are unaffected by the removal of redundant metrics. This work demonstrates that proficiency in point-of-care ultrasound applications is multi-faceted, and that measuring completion time alone is not enough and application-specific metrics have added value in proficiency assessment.

Keywords: Surgical skills assessment, ultrasound-guided interventions.

1 Introduction

Medical education is undergoing a shift from a traditional time-based model to a competency based model, where trainees graduate only upon achieving a competency benchmark. With increasing demands on expert clinician time, this necessitates automatic methods for proficiency assessment.

Accordingly, there has been a proliferation of methods of objective, automatic technical proficiency assessment for many clinical applications. These methods perform computation on data from a different sources including: hand or tool motion tracking data, video data of the surgical field or operating room, surgeon status information from wearable sensors (e.g. eye gaze, cognitive load, muscle activity), or quantification of resulting tissue. Reviews of methods for proficiency assessment for medical interventions training can be found in [1] and [2].

Computing overall proficiency from a combination of primitive performance metrics is common practice. This is because primitive metrics are straightforward to compute, easy for trainees to understand, and readily interpreted into actionable feedback. Furthermore, they can be used to capture application-specific information that generic assessment methods cannot. Fraser et al. and Stylopoulos et al. first addressed this, proposing a sum of normalized features [3] and a sum of z-scores [4], respectively. Subsequently, Allen et al. showed that using support vector machines for overall proficiency classification outperformed either of these methods [5]. Oropesa et al. confirmed that support vector machines likewise outperform linear discriminant analysis and adaptive-neuro fuzzy inference for classification overall proficiency classification [6]. Modern machine learning techniques have also been applied to this problem [7].

It is interesting to consider which metrics are critical for overall proficiency assessment and which metrics are redundant. Primarily, metrics must be valid for distinguishing novices from experts. Several valid metrics used in the assessment, however, may measure the same aspect of proficiency and correlate strongly, while others may address different aspects of proficiency. Redundant metrics may be removed to reduce system complexity without reducing assessment accuracy or feedback quality. Metrics addressing different aspects of proficiency, on the other hand, must remain to achieve a complete assessment with feedback specific to each aspect of proficiency.

In this work, we seek to evaluate which primitive metrics are sufficient and necessary for a complete assessment of technical proficiency in point-of-care ultrasound applications. In particular, we address whether simply measuring completion time is sufficient for overall proficiency assessment and the role of application-specific metrics.

2 Methods

2.1 Primitive Metric Validity

While most primitive metrics are designed to measure a clinically important quantity, it must still be show that they correlate with proficiency. To this end, we examined primitive metrics from both novices and experts, and assessed whether there is a difference between metrics for the two groups. Metrics which did not show evidence of validity were removed from subsequent analysis.

First, we used the Mann-Whitney U test (α=0.05) to determine if there is a statistically significant difference between the two groups for each metric. We used Cliff's Δ to quantify the effect size. Additionally, we measured the information gain associated with splitting on each metric. The information gain indicates how well splitting the data improves the groups' purity, where large information gain indicates that a metric distinguishes novices from experts effectively. We further assessed if the split produced significantly different groups using Fisher's exact test (α=0.05).

2.2 Primitive Metric Redundancy

Metric redundancy is most commonly computed using correlation, where a strong correlation indicates a high likelihood of redundancy. As an initial check, we computed the correlation between each pair of metrics.

Subsequently, we performed Exploratory Factor Analysis (EFA) on the primitive metric values. EFA expresses each primitive metric as a linear combination of some set of latent factors. Two primitive metrics which are similar linear combinations of the latent factors would be considered redundant. Furthermore, when combined with expert knowledge, the latent factors can be interpreted as aspects of technical proficiency and their importance can be identified. For this study, we used the principal components methods and chose the smallest number of factors explaining at least 90% of the variance in the data. Two primitive metrics were considered redundant if they both had loading factors greater than 0.90 on the same latent factor.

2.3 Assessment Using Non-Redundant Primitive Metrics

Once we identified which metrics were redundant using EFA, for each set of redundant metrics we chose one "representative" metric. This metric was chosen to be the metric with the best loading on each of the latent factors. We then computed an overall proficiency classification for each participant using the "representative" metrics for both the traditional sum of z-scores method [4] and the support vector machine method [5]. For the sum of z-scores method we used equal weighting. For the support vector machine method, we normalized the data on the range [0, 1] and used the radial basis function.

We compared the proficiency classification accuracies achieved using the "representative" set of primitive metrics with the accuracies achieved using all primitive metrics. The area under the receiver-operator characteristics curves was computed for each metric set for each of the sum of z-scores method and the support vector machine method. We determined whether the areas under the curves was different for the metric sets using the Hanley-McNeil test (α=0.05).

2.4 Datasets

We retrospectively analyzed datasets from three point-of-care ultrasound training applications: FAST ultrasound examination, femoral line insertion, and freehand lumbar puncture. In each case, we used previously computed metric values based on tool tracking data. In each case, the metrics were specifically designed by experts to capture relevant information on proficiency while performing the intervention.

In the FAST ultrasound training dataset, a group of fourteen novices and fifteen intermediates performed a complete FAST examination on a healthy volunteer on each of the four regions of interest (hepatorenal, splenorenal, pericardial, and pelvic regions) [8]. The ultrasound probe was tracked relative to the volunteer, and the following primitive metrics were computed: completion time, percentage of expert-defined points of interest missed, and ultrasound probe path length.

The femoral line insertion dataset included ten novices and four experts performing an ultrasound-guided insertion on a commercially available simulation phantom [9]. The motion of the operators' hands was tracked relative to the phantom model, and the following primitive metrics were computed: completion time, probe hand path length, needle hand path length, probe hand rotational actions, needle hand rotational actions, probe hand translational actions, and needle hand translational actions.

The lumbar puncture dataset included twenty-three novices and five experts performing freehand lumbar puncture on a commercially available lumbar spine model [10]. The pose of the operators' hands and needle was tracked relative to the phantom model, and the following primitive metrics were computed: completion time, left hand path length, right hand path length, needle tip path length, tissue damage caused by needle, needle tip path length in tissue, and needle tip time in tissue.

3 Results

3.1 Primitive Metric Validity

For the FAST dataset, all metrics were significantly different between novices and intermediates, thus all metrics were kept for subsequent analysis. For the femoral line dataset, probe hand and needle hand rotational actions were not significantly different between novices and experts, thus these two metrics were removed. All other femoral line metrics were kept. For the lumbar puncture dataset, all metrics were significantly different between the novice and expert groups, thus all metrics were kept (**Table 1**).

Table 1. Validity of metrics for each dataset. MW indicates the p-value for the Mann-Whitney test; Δ indicates the non-parametric effect size; F indicates the p-value for Fisher's exact test; IG indicates the maximal information gain associated with splitting on that metric.

Dataset	Metric	MW	Δ	F	IG
	Completion time (s)	<0.001	0.40	<0.001	0.10
	Points missed (%)	<0.001	0.58	<0.001	0.21
	Probe path length (mm)	<0.001	0.44	<0.001	0.08
	Completion time (s)	0.002	1.00	<0.001	0.60
	Probe hand path length (mm)	0.024	0.80	0.015	0.33
	Needle hand path length (mm)	0.024	0.80	0.011	0.36
	Probe hand rotational actions	0.056	0.68	0.070	0.26
	Needle hand rotational actions	0.607	0.20	0.221	0.16
	Probe hand translational actions	0.006	0.93	0.005	0.42
	Needle hand translational actions	0.002	1.00	<0.001	0.60
	Completion time (s)	<0.001	1.00	<0.001	0.47
	Left hand path length (mm)	0.001	0.93	<0.001	0.32
	Right hand path length (mm)	0.007	0.79	0.003	0.22
	Needle tip path length (mm)	0.006	0.81	0.001	0.23
	Tissue damage (mm^2)	0.010	0.76	0.001	0.25
	Needle path in tissue (mm)	0.026	0.65	0.026	0.14
	Needle time in tissue (s)	0.022	0.67	0.008	0.15

3.2 Primitive Metric Redundancy

The correlation matrices for each dataset are shown in **Fig. 1**. Using EFA, two latent factors were found for the FAST dataset, accounting for 91% of the variance. Two latent factors were found for the femoral line dataset, accounting for 98% of the variance. Three latent factors were found for the lumbar puncture dataset, accounting for 93% of the variance. The loading plots for each dataset are present in **Fig. 2**.

For the FAST dataset, completion time and probe path length both primarily load on one latent factor, and points missed loads primarily on the other latent factor. We interpret the first latent factor to be "efficiency" and the second latent factor to be "thoroughness". For the femoral line dataset, needle hand path length loads primarily on one latent factor and probe hand translational actions loads primarily on the other latent factor. We conjecture the first latent factor to be "needle hand efficiency" and the second latent factor to be "probe hand efficiency". All other primitive metrics cross-load on the two latent factors. For the lumbar puncture dataset, tissue damage caused by needle, needle tip path length in tissue, and time needle in tissue load primarily on one factor, left hand path length and right hand path length load primarily on another, and needle tip path length loads primarily on a third factor. Completion time cross-loads. We hypothesize these three latent factors to be respectively "needle insertion efficiency", "landmarking efficiency", and "needle placement efficiency".

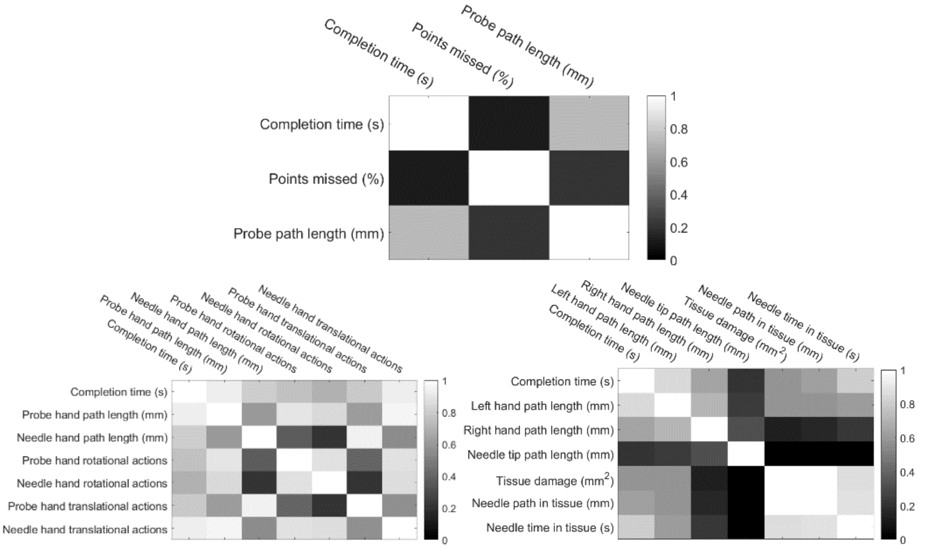

Fig. 1. Correlation matrices for metrics in the FAST (top), femoral line (left), and lumbar puncture (right) datasets. White indicates high correlation; black indicates low correlation.

Based on the primitive metric loadings, the following metrics were kept as "representative" metrics. For the FAST dataset, completion time and points missed were kept. For the femoral line dataset, needle hand path length and probe hand translational actions were kept. For the lumbar puncture dataset, right hand path length, needle tip path length, and tissue damage were kept.

3.3 Assessment Using Non-Redundant Primitive Metrics

Differences in the areas under the curves using all primitive metrics and using a "representative" set were insignificant for all datasets using both the sum of z-scores and support vector machine methods (**Table 2**). The greatest change in area under the curve was 0.052, for the lumbar puncture dataset using the z-score method (**Fig. 3**).

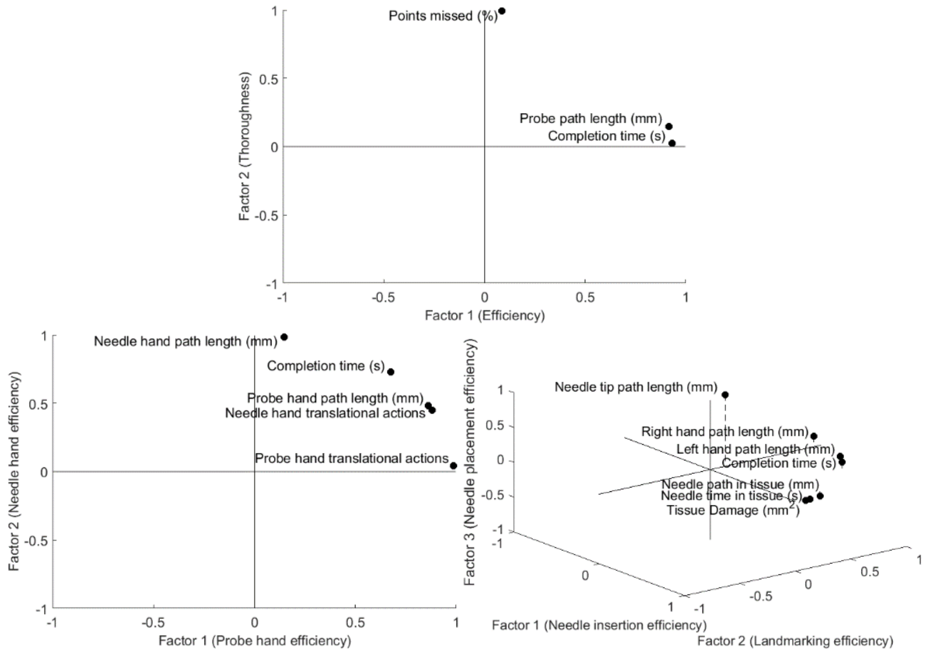

Fig. 2. Loading plots for metrics onto presumed factors in the FAST (top), femoral line (left), and lumbar puncture (right) datasets.

Table 2. Area under the curve (AUC) for each method of overall proficiency assessment. All AUC indicates the area under the curve using all metrics, and Rep. AUC indicates the area under the curve using only the "representative metrics". p-value indicates the p-value for the Hanley-McNeil test.

Dataset	Sum of Z-Scores			Support Vector Machine		
	All AUC	Rep. AUC	p-value	All AUC	Rep. AUC	p-value
FAST	0.84	0.83	0.45	0.84	0.84	0.44
Femoral Line	1.00	1.00	0.50	1.00	1.00	0.50
Lumbar Puncture	0.97	0.91	0.31	1.00	0.96	0.25

4 Discussion & Conclusion

In each dataset, the majority of the reported metrics were determined to be valid. There were strong correlations between many of the metrics, and exploratory factor analysis indicated that the metrics were associated with two to three latent factors. We interpreted the meaning of these latent factors using domain-specific knowledge. Taking only the most representative metrics for each factor, we achieved accuracies for overall proficiency assessment that were not significantly different from accuracies using all metrics. This indicates that many of the metrics could be removed; however, completion time cannot be used alone to measure proficiency. Furthermore, it shows application-specific primitive metrics have added value in proficiency assessment.

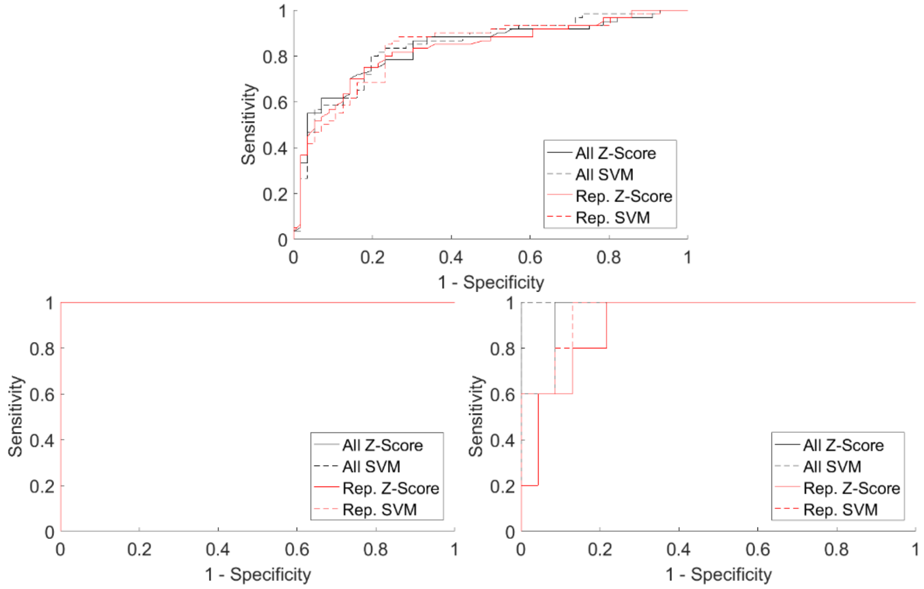

Fig. 3. Receiver operator characteristic curves for overall proficiency assessment for the FAST (top), femoral line (left), and lumbar puncture (right) datasets. Black lines indicate all metric were used; red lines indicate only "representative" metrics were used. Solid lines indicate the sum of z-scores method was used; dashed lines indicate the support vector machine method was used.

This study, however, is not without limitations. Primarily, for the femoral line and lumbar puncture datasets, the sample size is limited with the expert group including four and five participants respectively. This can be especially problematic for EFA. The other main limitation is that we have used experience as a proxy for ground-truth proficiency. This does not account for experts who have developed bad habits or have an “off day”. Ideally, ground-truth proficiency should be determined by a panel of experts using a valid objective assessment tool. Finally, our analysis assumes a monotonic relation between each metric and proficiency, which may not always be the case.

We suggest that these results will extend to other ultrasound-guided and freehand interventions. Here we have tested three different interventions, and our metric reduction techniques seem to apply well to each application, yielding less than six percent difference in proficiency classification for all datasets. We suggest this analysis could be used in other point-of-care ultrasound applications to identify which primitive metrics may be removed to reduce setup complexity and factors contributing to proficiency.

Finally, for each of these datasets, we have more than one latent factor contributing to proficiency. In particular, the application-specific metrics have added value and completion time alone is insufficient for assessing these factors. In fact, there may be additional factors which are not measured by the primitive metrics we chose. One should be aware of all such factors when computing an overall proficiency score. We suggest that providing a report card that addresses each of these factors may better allow trainees to understand which aspects of their intervention require the most improvement.

5 References

[1] C. E. Reiley, H. C. Lin, D. D. Yuh, and G. D. Hager, "Review of methods for objective surgical skill evaluation," *Surg. Endosc.*, vol. 25, no. 2, pp. 356–366, 2011.

[2] S. S. Vedula, M. Ishii, and G. D. Hager, "Objective Assessment of Surgical Technical Skill and Competency in the Operating Room," *Annu. Rev. Biomed. Eng.*, vol. 19, no. 1, 2017.

[3] S. A. Fraser, D. R. Klassen, L. S. Feldman, G. A. Ghitulescu, D. Stanbridge, and G. M. Fried, "Evaluating laparoscopic skills," *Surg. Endosc. Other Interv. Tech.*, vol. 17, no. 6, pp. 964–967, 2003.

[4] N. Stylopoulos *et al.*, "Computer-enhanced laparoscopic training system (CELTS): bridging the gap," *Surg Endosc*, vol. 18, no. 5, pp. 782–789, May 2004.

[5] B. Allen, V. Nistor, E. Dutson, G. Carman, C. Lewis, and P. Faloutsos, "Support vector machines improve the accuracy of evaluation for the performance of laparoscopic training tasks," *Surg Endosc*, vol. 24, no. 1, pp. 170–178, Jan. 2010.

[6] I. Oropesa *et al.*, "Supervised classification of psychomotor competence in minimally invasive surgery based on instruments motion analysis," *Surg. Endosc.*, vol. 28, no. 2, pp. 657–670, 2014.

[7] B. D. Kramer, D. P. Losey, and M. K. O'Malley, "SOM and LVQ Classification of Endovascular Surgeons Using Motion-Based Metrics," in *Advances in Self-Organizing Maps and Learning Vector Quantization: Proceedings of the 11th International Workshop WSOM 2016, Houston, Texas, USA, January 6-8, 2016*, E. Merényi, M. J. Mendenhall, and P. O'Driscoll, Eds. Cham: Springer International Publishing, 2016, pp. 227–237.

[8] M. S. Holden, T. Ungi, C. McKaigney, C. Bell, L. Rang, and G. Fichtinger, "Objective Evaluation Of Sonographic Skill In Focussed Assessment With Sonography For Trauma Examinations," in *CARS 2015—Computer Assisted Radiology and Surgery Proceedings of the 29th International Congress and Exhibition Barcelona*, 2015, pp. S79–S80.

[9] R. McGraw *et al.*, "Development and Evaluation of a Simulation-based Curriculum for Ultrasound-guided Central Venous Catheterization," *CJEM*, pp. 1–9, May 2016.

[10] C. T. Yeo, C. Davison, T. Ungi, M. Holden, G. Fichtinger, and R. McGraw, "Examination of Learning Trajectories for Simulated Lumbar Puncture Training Using Hand Motion Analysis," *Acad. Emerg. Med.*, vol. 22, no. 10, pp. 1187–1195, 2015.