Adv in Health Sci Educ (2017) 22:741–759
DOI 10.1007/s10459-016-9709-2

A primer on the statistical modelling of learning curves in health professions education

Martin V. Pusic[1] · Kathy Boutis[2] · Martin R. Pecaric[3] · Oleksander Savenkov[4] · Jason W. Beckstead[5] · Mohamad Y. Jaber[6]

Received: 13 October 2015 / Accepted: 11 August 2016 / Published online: 3 October 2016

Abstract Learning curves are a useful way of representing the rate of learning over time. Features include an index of baseline performance (y-intercept), the efficiency of learning over time (slope parameter) and the maximal theoretical performance achievable (upper asymptote). Each of these parameters can be statistically modelled on an individual and group basis with the resulting estimates being useful to both learners and educators for feedback and educational quality improvement. In this primer, we review various descriptive and modelling techniques appropriate to learning curves including smoothing, regression modelling and application of the Thurstone model. Using an example dataset we demonstrate each technique as it specifically applies to learning curves and point out limitations.

Keywords Competency · Growth models · Learning curves · Mastery · Medical education · Regression models · Statistical modelling

Electronic supplementary material The online version of this article (doi:10.1007/s10459-016-9709-2) contains supplementary material, which is available to authorized users.

✉ Martin V. Pusic
martin.pusic@nyumc.org;
http://www.med.nyu.edu/biosketch/pusicm01

[1] Institute for Innovations in Medical Education, New York University School of Medicine, 550 First Avenue, MSB G109, New York, NY 10016, USA

[2] The Hospital for Sick Children, and University of Toronto, Toronto, ON, Canada

[3] Contrail Consulting Services, Toronto, ON, Canada

[4] Department of Medicine, New York University School of Medicine, New York, NY, USA

[5] University of South Florida College of Nursing, Tampa, FL, USA

[6] Department of Mechanical and Industrial Engineering, Ryerson University, Toronto, ON, Canada

Introduction

The learning curve is an excellent representation of the goal-directed learning that occurs with practice to achieve a desired level of expertise (Pusic et al. 2015). The y-intercept represents the learner's prior knowledge. The slope is proportional to the efficiency of learning given that the mean increase in performance per unit of study varies according to the difficulty of the material, the quality of instruction or some combination thereof. With continued effort, the learner crosses a mastery threshold after which the focus can switch to performance maintenance as opposed to improvement (Pusic et al. 2012) though the highest level of expert often chooses an adaptive expertise approach in which a lifelong attitude of improvement is cultivated (Kalet and Pusic 2014). The learning curve asymptote or plateau is a property of the learning system, representing the maximal learning potential given endless repetitions (Jaber and Guiffrida 2004; Jaber 2006; Pusic et al. 2015).

While considerable information can be gleaned from a qualitative assessment of the learning curve, in this paper we explore the degree to which mathematical modelling of the learning curve relationship can be of use to health professions educators and researchers (Flavio et al. 2011; Jaber 2006).

We describe an approach to the statistical modelling of representative health professions education tasks that can be repetitively practiced. In explicitly modelling the longitudinal, growth nature of learning during effective practice, we demonstrate the feasibility of predicting individual learning trajectories to the benefit of individuals and groups of learners. Using both cognitive (radiology image interpretation) and psychomotor procedural (laparoscopy) examples, we begin with the equivalent of descriptive statistics: scatter plots and moving averages. We then describe the process of linear and nonlinear regression modelling for learning curves, considering both data driven fitting of models and theoretical frameworks. We separately describe learning curve models derived from group data compared with those compiled at the individual level. We finish by discussing potential applications of these techniques.

Conceptual basis

As early as 1909, Robertson developed a *learning equation* based on the principles of autocatalysis seen in physical chemistry (Robertson 1909; Singer and Willett 2003). His equation had a logistic trajectory, arriving at an asymptote as the amount "to be learned" decreased. Thurstone later (1917) found that the theory held for women learning to type, with the important parameters being the y-intercept, efficiency (a variant of slope) and the asymptote (Thurstone 1919). The ogive appearance to the curve means that there are diminishing returns, relative to effort expended, as one approaches the asymptote. This is consistent with modern conceptualizations of expertise (Ericsson 2004).

A large number of formulae have been applied with success to model learning including the Thurstone equation, logistic and power-law regression, each of which we consider (Flavio et al. 2011; Ramsay et al. 2001). With novice learners, we have found that linear regression can be a helpful first approximation in that it is easily determined and explained. It also has the advantage of a "goodness-of-fit" metric, r-squared, which is easily calculated and compared, a fact that we will use to advantage when we examine the learning of a group of individuals (Altman 1990).

We propose that the variability in the degree to which a learning curve relationship holds can be useful to health professions educators. There is a large amount of empirical data validating the learning curve conceptualization. Ritter and Schooler describe the learning curve's universality well when they say…

> "From short perceptual tasks to team-based longer term tasks of building ships, the breadth and length of human behavior, the rate that people improve with practice appears to follow a similar pattern. It has been seen in pressing buttons, reading inverted text, rolling cigars, generating geometry proofs and manufacturing machine tools, performing mental arithmetic on both large and small tasks, performing a scheduling task, and writing books (Ritter and Schooler 2001)."

Often when we apply statistical modeling techniques, typically in research settings, we are trying to draw inferences from empirical data to inform a theory. However, given the wide acceptance of learning curves as a fundamental principle of psychology, it may be more advantageous to go in the opposite direction—that is, we can assume, under conditions of practice with adequate ongoing feedback, that if a learning curve model does not hold then the *fault is likely not with the theory but rather lies with the conditions of learning*. In this way of thinking, we are not asking the question "is there a relationship between learning and time or effort", but rather "if there isn't a relationship, why not?" This orientation allows us to investigate situations where the learning relationship holds for some learners, but not others. Are some learners not engaged? Are there developmental differences such that this intervention is not appropriate for these learners? Thus, a learning curve with a statistically significant *negative* slope is a call to attention for the educator; similarly, a model that quantitatively demonstrates the relationship for some learners but not others has provided interesting information on the learning method.

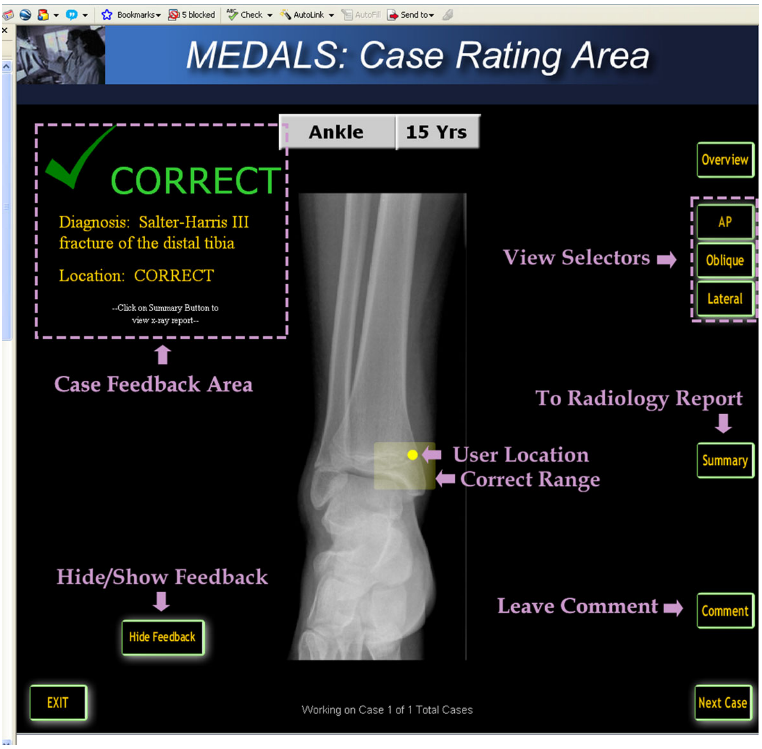
Example learning curve dataset

To describe the statistical modeling of learning curves, we will both draw on literature examples and demonstrate analyses using an example dataset which we describe in this section. Our emphasis is on demonstrating the analyses as opposed to presenting the research findings for which we refer the reader to the prior reports (Boutis et al. 2010; Pusic et al. 2015).

Learning curve models can be applied to assess learners' development over time when the following conditions are met: (a) there are at least three repeated observations per individual with (b) adequate sample size with the minimum being reported as low as 22 individuals but generally closer to 100; more repeated observations per individual result in increased power and (c) where maximum-likelihood methods are used, the repeated measure needs to be normally distributed; however, learning curves can be fitted by other means as well. The reader is referred to Curran et al. for a relevant summary (Curran et al. 2010).

Education intervention and post-intervention testing

We prospectively collected an initial pool of 234 ankle radiographs that were obtained to exclude the possibility of ankle fracture (Boutis et al. 2010). Each case included the three standard ankle radiograph views as well as the staff pediatric radiologist's report. Cases were categorized as either normal or abnormal based on the official radiology report. Case-



CORRECT

Diagnosis: Salter-Harris III fracture of the distal tibia

Location: CORRECT

--Click on Summary Button to view x-ray report--

Working on Case 1 of 1 Total Cases

Fig. 1 Screen capture from after the learner submits their answer (feedback screen). The purple annotations do not appear in the actual program. The *yellow dot* represents the learner's designation of the abnormality, just inside the "hotspot" representing the correct location (Boutis et al. 2010). (Color figure online)

to-case variability in terms of severity, quality etc. was accepted as being useful pedagogically.

Computer program to present radiographs

In our prior report, we described the computer program used to present cases to participants (Boutis et al. 2010). Subjects classified each case as either 'Normal' (no fracture), or 'Abnormal' (fracture present). Abnormal answers also required the user to indicate the location of the suspected abnormality with a mouse click. Once a participant commits to his/her diagnosis, the program immediately provided feedback by highlighting the pathology on the abnormal images and providing the radiologist's report (Fig. 1).

Participants In this review, we contrast the learning of 38 participants; 20 relatively novice medical students (MS), against that of more advanced physicians in training: 6 residents (RS), and 12 pediatric emergency medicine fellows (FL). The dataset is suitable for our demonstration in that it shows both inter-individual and inter-group variability in both prior knowledge and rate of learning. Each participant was presented with 234 randomly assigned ankle radiographs using a web-based application. The small number of participants has advantages for straightforward demonstration of the analyses, at the cost of leaving some of our comparisons underpowered. The dataset is provided as a supplemental appendix.

Item-coding We considered each completed case as one item. Normal radiograph items were scored dichotomously. Abnormal radiographs were scored correct if the participant both identified them as abnormal and correctly indicated the lesion on the radiograph. We defined "accuracy" as the dependent variable in our analyses, which takes on the value 0 or 1 for an individual case or, at the test level, the proportion correct.

Descriptive learning curves

In this section, we describe analytic methods that give insights as to the learning by individuals. We consider, in turn, scatter plots and moving averages.

Scatter plots

In order to generate a learning curve, we graph an index of performance or learning on the y-axis against an index of learning effort on the x-axis with the expectation that performance improves with increasing learning effort (Jaber and Sverker 2004; Pusic et al. 2015). In some cases, the effort variable is easily represented: for example counts of how many times an operation is repeated. Figure 2 shows a scatter diagram of the time taken (speed) by one surgeon learning to perform a laparoscopic cholecystectomy (Ahlering et al. 2004). As surgeons gain psychomotor expertise, their speed improves such that the relationship is nonlinear with much larger speed gains initially (Ramsay et al. 2001). Each point in Fig. 2 represents the time taken by one surgeon doing one operation somewhere in a sequence of operations. While there is variability in the repeated unit (operations) in terms of difficulty and other factors, the overall trend is clearly seen.

Binning

In some cases, like the radiograph interpretation example that we will further develop, the x- or y-axis variable must be aggregated to properly show the learning relationship in meaningful units. First, consider another learning curve of surgical operating times as they decrease with cumulative experience represented by the number of cases performed (Fig. 3).

Here, the learning curve has been smoothed by taking the average time over a *bin* of 10 cases and then demonstrating how the 10-case average of time changes with experience. This is an example of a connected-line scatterplot of the binned operative times. "Smoothing" results in less random noise in the scatterplot potentially making the underlying learning curve relationship more clear. This approach works well for situations where the y-axis is a continuous interval measure such as time, since continuous measures can be averaged (and therefore smoothed) with the same meaning throughout the scale.

For an individual learning the visual diagnostic skill of interpreting radiographs, a single case has less meaning as an index of an individual's overall performance. Instead their score across an aggregation of cases (again a *bin*) is a more reliable measure of performance. Determining a sensible number of cases per data point, or "bin size", depends on the nature of the task. In Fig. 4, we have plotted the Average Standard Error of Measurement from the individuals' scores on the ankle radiographs described earlier, against the number of cases within an individual's bin.

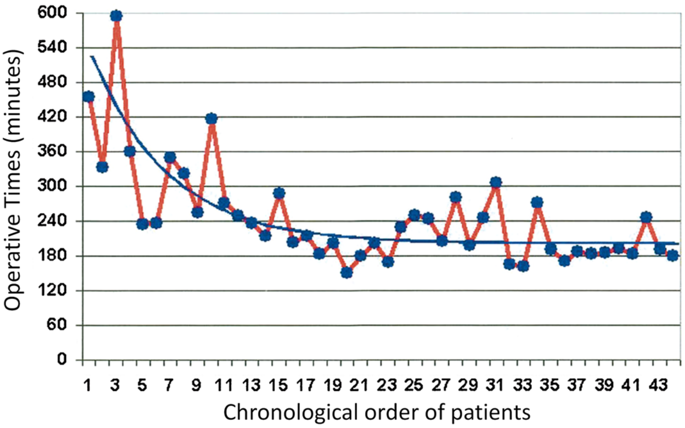


Fig. 2 Connect-line scatterplot of a single surgeon doing prostate surgery. Total operative time per single case graphed against the number of cases completed. http://www.laparoscopytoday.com/2006/01/robotassisted_r.html

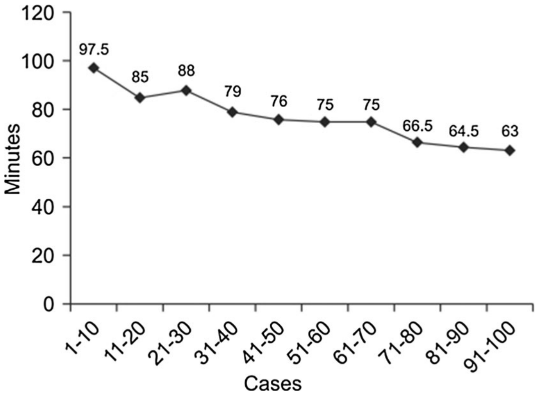


Fig. 3 "Smoothed" learning curve. Average times over 10-case bins for a single surgeon performing single-port laparoscopic cholecystectomy (Koo et al. 2012)

We can see that it takes about 10 cases to reduce the error in the estimate of accuracy by about 40 % (from ~20 % to 13 %), but doubling the bin size to 20 reduces the SEM by only 3 more percentage points (to 10 %). Each individual in the ankle radiograph study completed 234 cases. One way of binning this data is to aggregate the data from every 18 cases interpreted, resulting in 13 individual testlets ($13 \times 18 = 234$) though any number of other combinations of cases per testlet are possible, trading off individual data point precision against overall number of data points. A scatter diagram based on 13 testlets of a representative individual completing the radiographs is shown in Fig. 5.

Moving averages

Smoothing functions attempt to capture important patterns in a set of data, while leaving out noise or other rapidly fluctuating phenomena that do not bear on the overall trend

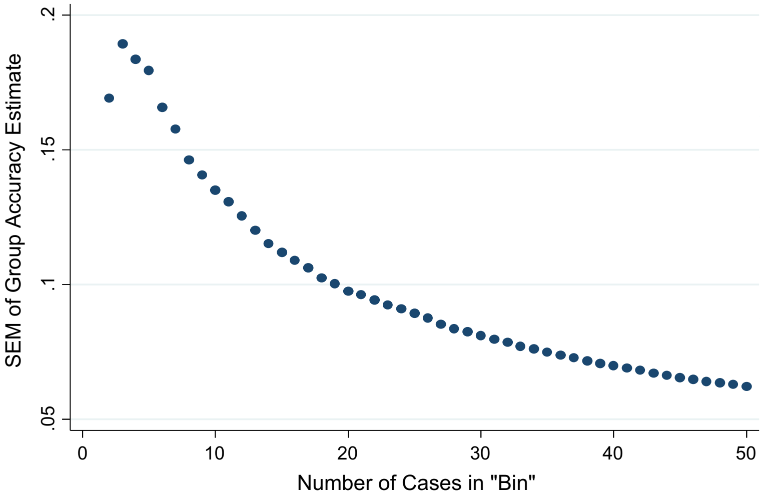


Fig. 4 Standard error of measurement for the accuracy across individuals reading ankle radiographs, as described in the text

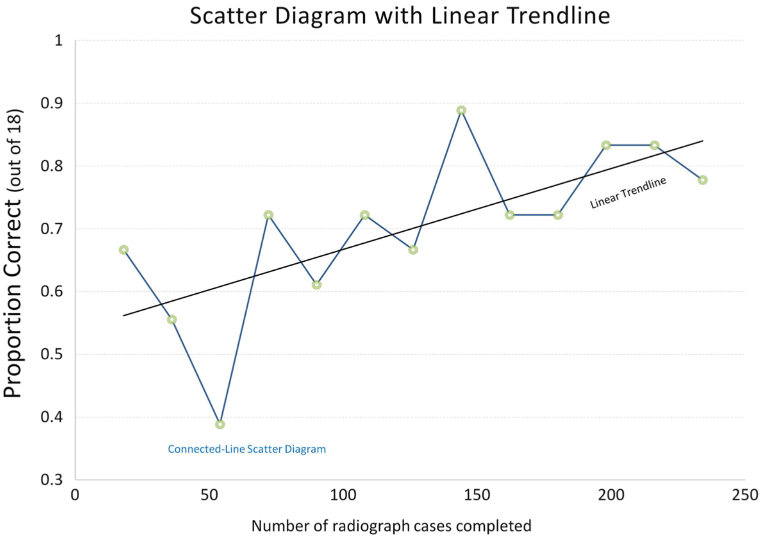


Fig. 5 A scatter diagram showing the results of a single individual's learning curve (MS002) upon completion of 234 radiograph cases with immediate feedback. The results have been aggregated into 13 bins of 18 cases. In addition, a regression trendline has been drawn through the 13 datapoints

(Simonoff 1996). Two of the most commonly used smoothing functions are *cumulative* and *simple* moving averages (Table 1). The *cumulative* moving average takes each new data point in a sequential series and incorporates it into the estimate of the average. While useful for demonstrating overall learning trends, cumulative moving averages' (e.g. Figure 6—green line) current estimate of the participant's accuracy is always weighed down by early cases, and therefore does not fully reflect the learning that has occured. *Simple*

Table 1 Summary of descriptive learning curves at the individual level

Type	Legend (see Fig. 6)	Formula for last graphed point	Formula for second last graphed point	Consecutive points independent?	Comment
Connected lines scatter plot	Blue jagged line	Average of accuracy for cases 216–234	Average of accuracy for cases 198–216	Yes, the data points are generated from non-overlapping data	Advantage is that this type of graph is straightforward to generate; can have considerable noise
Predicted regression line	Grey trendline	Interpolated from regression formula: Accuracy = $\beta_0 + \beta_1 n$, where β_0 (y-intercept) and β_1 (slope) are parameters, and n is the bin number		Yes, regression generated from the 13 independent observations	Maximum likelihood approximation which can be judged using model fit measures like r2. Assumes linear relationship
Simple moving average	Yellow moving average	Average of prior 18 points (cases 216–234)	Average of prior 18 points (cases 215–233)	No, the two adjoining points share 17 common data-points	Smoothed but prone to tradeoffs with regards to bin size
Cumulative moving average	Green moving average	Average over all cases to that point (cases 1–234)	Average over all cases to that point (cases 1–233)	No, the two adjoining points share n-1 common data-points	Maximum smoothing but later points underestimate actual ability/performance

Refer to Fig. 6 "Cumulative and Simple Moving Averages"

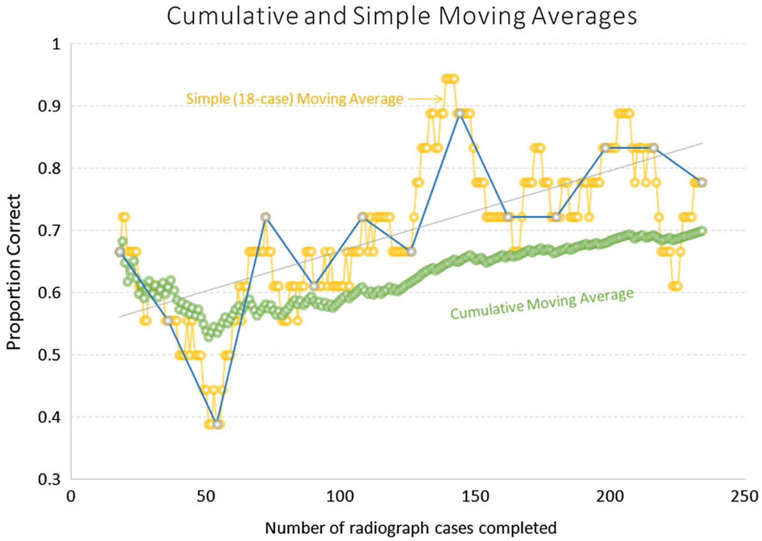


Fig. 6 The same scatter diagram as shown in Fig. 5 is now overlaid with cumulative (*green*) and 18-case simple (*yellow*) moving average curves. Note that the cumulative moving average underestimates the learners functioning at the end of the learning experience. (Color figure online)

Table 2 Linear regression parameters for 38 learners ordered by the amount of variance explained (r^2), where "SE" is standard error. In this case, the degree to which a linear model holds is highly variable between learners

Learner	Intercept	SE (inter)	slope	SE (slope)	r^2
MS019*	0.537	0.037	0.019	0.021	0.655
FL009	0.444	0.095	0.030	0.012	0.605
FL010	0.551	0.040	0.024	0.005	0.499
MS002*	0.529	0.047	0.024	0.006	0.473
FL005	0.655	0.032	0.014	0.004	0.429
FL012	0.530	0.075	0.025	0.009	0.416
MS016	0.618	0.038	0.019	0.005	0.413
FL008	0.537	0.057	0.020	0.007	0.402
MS018	0.485	0.059	0.019	0.007	0.397
MS013	0.344	0.175	0.035	0.020	0.379
RS005	0.596	0.064	0.019	0.008	0.372
FL011	0.574	0.081	0.027	0.010	0.362
MS012	0.571	0.031	0.018	0.004	0.338
FL001	0.589	0.039	0.018	0.005	0.332
RS002	0.432	0.163	0.026	0.020	0.324
RS001	0.639	0.031	0.011	0.004	0.245
MS014	0.431	0.135	0.023	0.017	0.223
RS004*	0.370	0.078	0.019	0.038	0.221
MS008	0.742	0.049	−0.015	0.006	0.214
FL007	0.541	0.110	0.016	0.014	0.179
MS017	0.403	0.075	0.017	0.009	0.168
MS015	0.489	0.078	0.015	0.010	0.152
FL003	0.781	0.034	0.007	0.004	0.127
MS020	0.684	0.137	−0.008	0.016	0.120
MS001	0.584	0.109	0.012	0.014	0.098
MS007	0.579	0.060	0.009	0.008	0.097
FL002	0.704	0.040	0.007	0.005	0.091
RS006	0.658	0.135	0.006	0.016	0.067
MS003	0.536	0.070	0.009	0.009	0.067
MS010	0.645	0.075	0.007	0.009	0.057
MS006	0.684	0.058	0.004	0.007	0.029
MS009	0.830	0.163	−0.012	0.020	0.027
FL004	0.784	0.069	−0.002	0.009	0.019
MS005	0.630	0.067	−0.003	0.008	0.014
MS011	0.639	0.062	0.002	0.008	0.014
FL006	0.747	0.082	−0.005	0.010	0.011
MS004	0.654	0.060	0.002	0.008	0.010
RS003	0.664	0.117	0.000	0.014	0.001

moving averages use a subset (or bin) of fixed size, and therefore fixed precision, across the whole length of the learning curve (Simonoff 1996). The issue of optimal bin size also applies to the size of the window used in a simple moving average. Simple moving averages provide a generally more accurate ending estimate of performance where the final estimate is the average score for the final bin (of size 18 in our example from Fig. 6—

yellow line). By contrast, in the cumulative moving average, the final estimate is across *all* repetitions and so incorporates data from the entire set (of size 234 in the same example, Fig. 6—green line). As mentioned, this has the effect of underestimating the final performance since the cumulative moving average is penalized by including early repetitions when the participant is not as facile with the task as they are at the end.

A particular type of the cumulative moving average used in health education for assessing procedural competence is the CUSUM (Cumulative Sum) analysis where success at a procedure, over sequential repetitions, is plotted taking into account a known or estimated acceptable failure rate (Bolsin and Colson 2000). Finally, curve-fitting using regression techniques can be considered a form of smoothing, but differs in that the emphasis is more explicitly placed on matching the data to a mathematical function that is ideally theory-based (whereas smoothing functions are non-parametric) (Simonoff 1996), (Priestley and Chao 1972).

Learning curve regression models for individuals

In this section, we explore the degree to which regression modeling can be used to draw inferences about the learning of an individual under the conditions of repetitive practice. We use the same dataset as for the prior exploration of descriptive learning curves including and, where applicable, the same binning of 13 testlets of 18 cases each. While any number of "linking functions" can describe the relationship between performance and learning over time, we will consider in turn, logistic, linear and power models (Pusic et al. 2015; Ramsay et al. 2000).

Logistic regression is used to model the relationship between a categorical dependent variable (in the case of the radiograph cases, dichotomous accuracy) and one or more independent variables such as the number of cases completed. In Fig. 7 we see a graph of the logistic function for the same learner (MS002) considered in the Moving Averages section. There are important distinctions from other types of learning curves. First, the y-axis is not an actual continuous outcome such as Accuracy), but rather a transformed probability

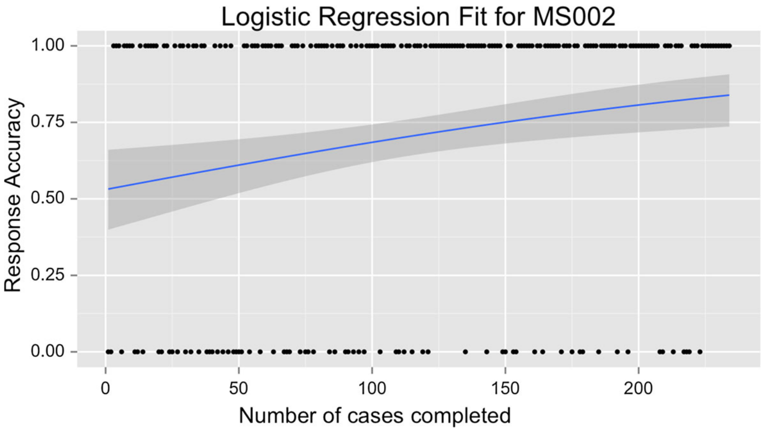


Fig. 7 Logistic regression model of learner MS002's progression through the 234 radiology cases. Each case is answered either correctly (scored 1.0 and dot on the *top line*) or incorrectly (scores 0.0 and dot on the *bottom line*). The y-axis is the predicted probability of a correct outcome given completion of x radiograph cases. *Shaded area* is the 95 % confidence interval about the prediction line

generated as the outcome of a Bernoulli distribution. As a result, the relationship between the trial data, a series of successes or failures, is more difficult to relate to an individual's path through the material. A key advantage of logistic regression modeling is that there is no need to "bin" the data since each datapoint can be meaningfully represented.

Linear and power-law regression models can be used at the individual level to relate performance to the number of cases completed. They differ from logistic regression in that the performance (y-axis) variable is a continuous measure: for the surgical operation example, it would be time per case; for the radiology example, the number of cases accurately diagnosed per bin of 13 cases. In Fig. 8 we show both relationships, the power law being theoretically more applicable to a learning situation where ongoing learning effort is likely to result in diminishing performance returns as the learner goes along (Ericsson 2008; Singer and Willett 2003).

For the situation of discrete outcomes, as for the radiology example, the question arises as to whether to use logistic or linear/power regression to generate learning curves. We suggest using each according to their advantages. Logistic regression modeling obviates the need for subjective determination of bin sizes and therefore the estimated parameter for learning rate is relatively standard. On the other hand, linear/power regression curves have the advantage of a directly-observed y-axis measure and a more fine-grained representation of the learner's path and how it differs from those of others.

In Table 2, we have listed the linear regression parameters of all 38 of the learners in our study in order to demonstrate the inter-individual variability in the degree to which a linear regression model fits. Similar inter-individual variability is seen for a Power model (data not shown). Using an index of model fit (r^2), shows that the models vary in the amount of variance explained from 0 (learner RS003), indicating no relationship whatsoever of Accuracy to number of cases completed, to an individual where fully 65 % of the variance in Accuracy is explained by the amount of practice (learner MS019). Each of these learners completed the same 234 cases albeit in a different random order.

In summary, learning curve models at the individual level are able to represent the inter-individual variability seen, from completely unsuccessful learners to those who follow the theoretical trajectory very well. We could speculate that the degree to which an individual follows a theoretical statistical model reflects the success of the learning environment for that individual and thus constitutes an indirect assessment that the necessary elements for learning are present, including engagement/motivation as well as match between feedback and the developmental stage of the learner.

Learning curve models for group-level data

Next, let us turn to the statistical analysis of the averaged learning curves of different groups for the purposes of comparison. For group comparisons we contrast the learning of two groups: the 20 medical students and the 18 more-expert residents and fellows under the same conditions of deliberate practice. The dependent variable is the accuracy score of the individual on each 13-case bin.

Repeated measures ANOVA

If the repeated measurements are independent (for normally distributed responses) we can use standard ANOVA techniques to estimate the difference between two groups of

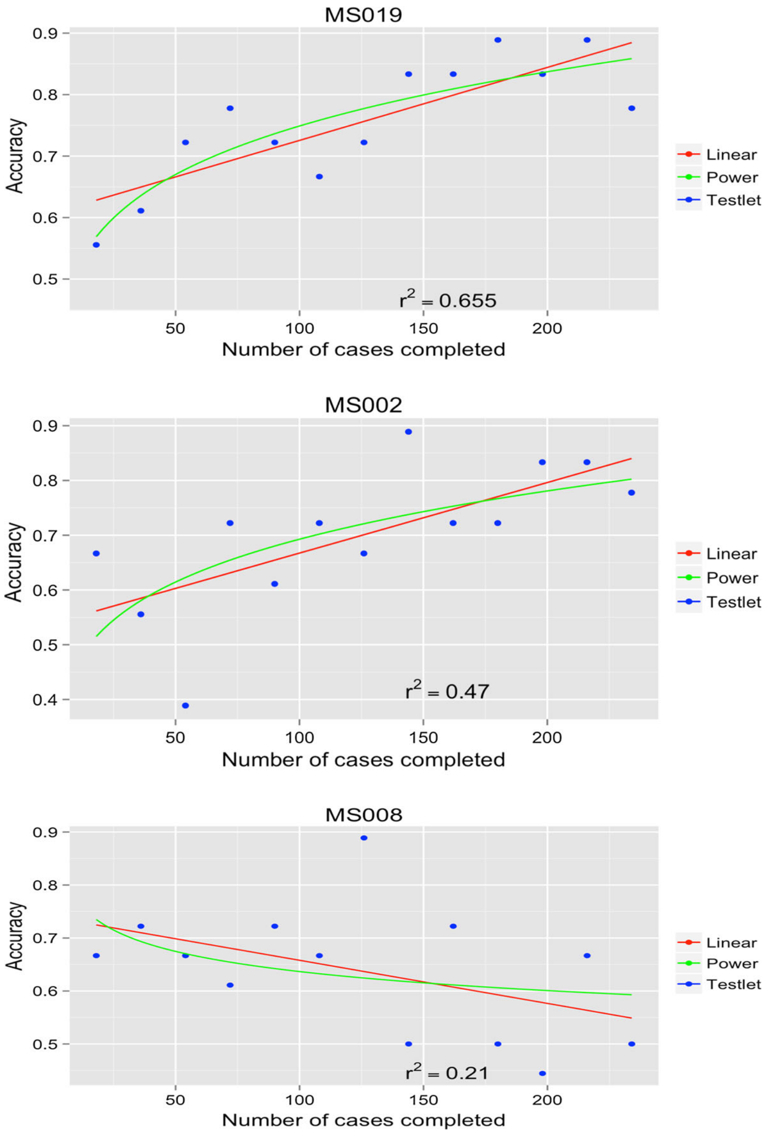
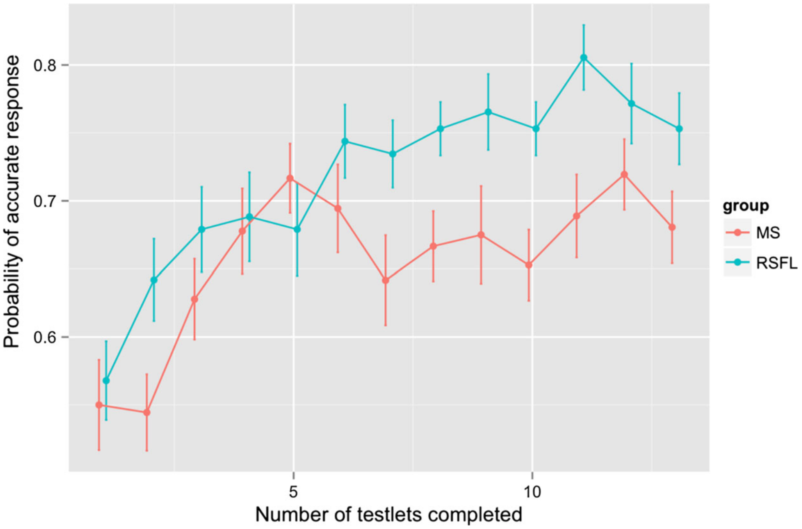


Fig. 8 Three individual learning curves chosen to show the degrees of correlation (r-squared) seen with 13 testlets of 18 cases each. The middle pane shows the same learning curve as in the moving average in Figs. 5 and 6. Model parameters for each are shown in Table 2, indicated by an *asterisk* (*) The r-squared displayed on each figure is for the linear model. Note that the slope of the learning curve for learner MS008 is negative to a statistically significant degree

participants. However, in fact the observations are *not* independent so we instead use Repeated Measures ANOVA which takes into account the correlation of observations from the same individual.

From the analysis (Fig. 9) we conclude that the "group" effect on accuracy is significant. The "Number of Cases" effect was also significant, meaning that the groups'



	Df	Sum Sq	Mean Sq	F value	Significance
Between Individuals					
Group	1	0.4668	0.4668	11.23	0.0019
Residuals	36	1.4966	0.0416		
Within Individuals					
# of Testlets	12	1.45	0.117	8.54	0.000
group:NumTestlets	12	0.219	0.01823	1.329	0.198
Residuals	432	5.924	0.01371		

Fig. 9 Repeated measures ANOVA for 13 testlets of 18 cases each. Bars on graph represent standard errors for the group means

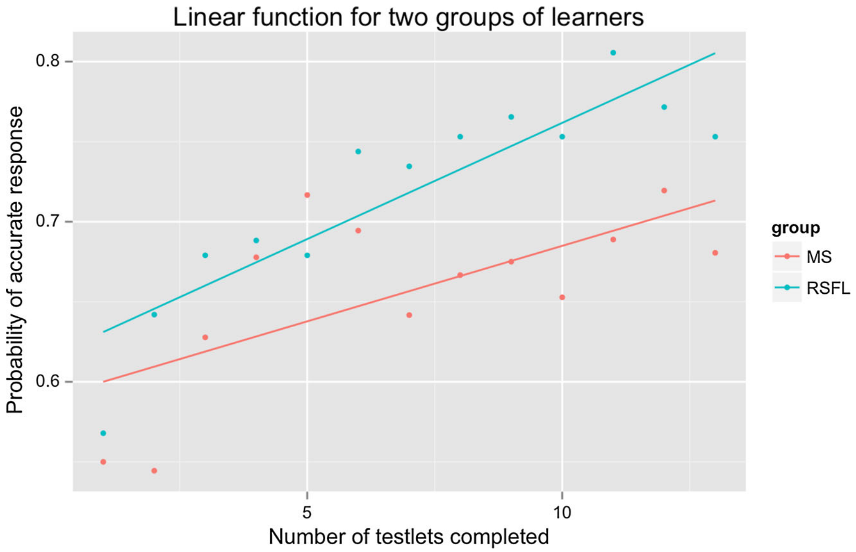
responses improve over time. The interaction term was not statistically significant suggesting that the dependence of learning on the number of cases is not statistically different between groups.

Linear and power models

Linear and power models can be useful for summarizing group-level learning curve data. In Figs. 10 (linear) and 11 (power), we see that for each group separately, there is a statistically significant slope indicating a relationship between the number of testlets completed and level of accuracy.

Comparing the slopes of the students to those of the residents and fellows using a *t* test revealed a more positive slope for residents of +0.006 with a 95 % CI of 0.0, +0.014. The difference in y-intercepts did not achieve statistical significance (mean difference +0.03; 95 % CI –0.038, +0.092).

Power models have the theoretical advantage of better representing the phenomenon of “diminishing returns” seen with the repeated deliberate practice that characterizes elite



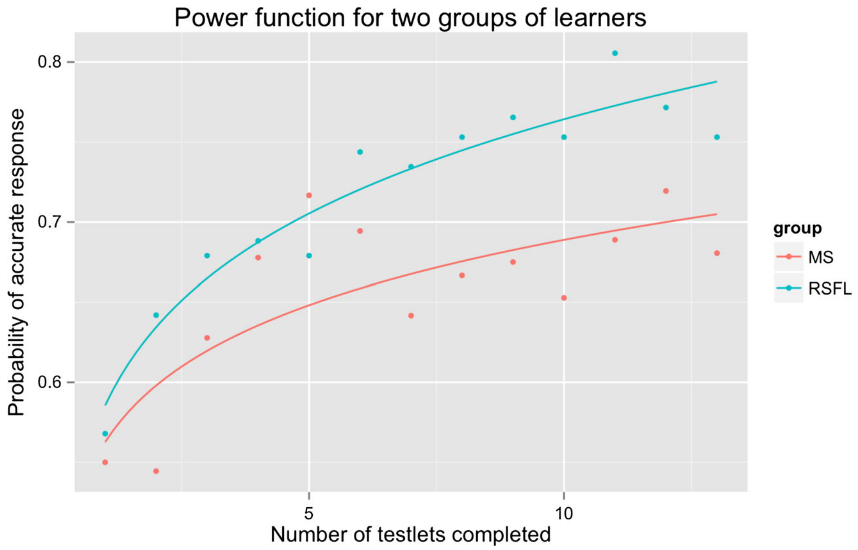
Medical Students (MS)				
	Estimate	Std. Error	t value	Significance
Intercept	0.59	0.025	23.4	0.000
Slope	0.0094	0.0031	2.96	0.013
Residents and Fellows (RSFL)				
Intercept	0.617	0.019	32.7	0.000
Slope	0.015	0.0023	6.1	0.000

Fig. 10 Linear regressions at the group level

performance. In a linear model, a large number of iterations would seem to result in the same rate of improvement with each additional testlet, something that is not observed in practice where later repetitions result in less improvement per repetition.

The Thurstone learning model

The Thurstone function (see Supplemental Appendix) adds one more parameter to the previously discussed indices of prior knowledge (y-intercept) and learning efficiency (slope); namely the asymptote (Singer and Willett 2003; Thurstone 1919). Based on the asymptote, we can predict the maximum possible learning for each group, under the conditions of the experiment. From the modeling results the coefficients for the asymptotes were 0.72 (SE = 0.032) for medical students and 0.84 (SE = 0.031) for the residents and fellows. Not only do fellows/residents learn at a greater rate, but they would be expected to attain a different maximum learning benefit from the intervention. Note that the addition of the third parameter (asymptote) makes this formula require greater participant numbers to generate a fitted model (Fig. 12).



	Estimate	Std. Error	t value	Significance
Medical Students (MS)				
Intercept	0.562	0.024	23.5	0.0000
Learning efficiency	0.088	0.022	4.03	0.0012
Residents and Fellows (RSFL)				
Intercept	0.586	0.013	44.03	0.0000
Learning efficiency	0.116	0.0115	10.03	0.0000

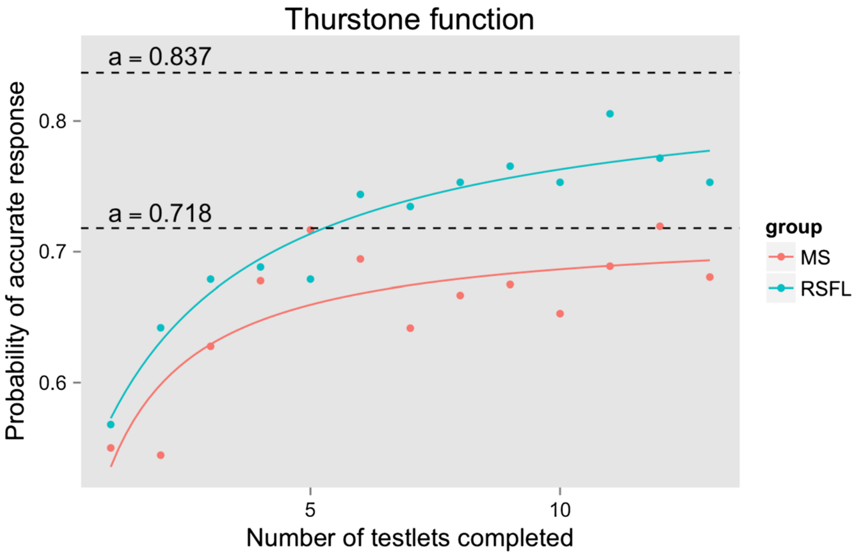
Fig. 11 Power law regressions at the group level

Discussion

In this review, using radiology and surgery learning data, we have shown that the learning curve describing the relationship between productive learning effort and performance can be successfully modeled using a number of descriptive (scatter plots, moving averages) and modeling techniques (regression) both at the individual and group levels. Our intent was to show how specific mathematical approaches to the description of the learning process can yield insights beyond those of the graphics alone.

Learning curves are a particular instance of statistical models termed *growth curve models* which allow for the estimation of “inter-individual variability in intra-individual patterns of change over time” (Curran et al. 2010). They represent an advance in that they provide a finer grained examination of the trajectories of individuals than do more traditional summary statistics, change scores, or even repeated measures ANOVA with its strict assumptions.

We have distinguished between individual-level and group-level learning curves and recommend determining both in order to fully assess the effectiveness of repeated practice.



Medical Students (MS)				
	Estimate	Std. Error	t value	Significance
Asymptote	0.718	0.033	21.8	0.000
Intercept	0.33	0.37	0.89	0.39
Efficiency	0.897	1.57	0.57	0.582
Residents and Fellows (RSFL)				
Asymptote	0.837	0.032	26.5	0.000
Intercept	0.467	0.069	6.735	0.000
Efficiency	2.498	1.426	1.751	0.11

a=asymptote (dashed horizontal lines)

Fig. 12 Thurstone learning curves at the group level

Fitting a group level model establishes that the learning curve relationship holds and allows determination of the average prior knowledge (y-intercept) and average efficiency (slope) as well as, in the Thurstone model, an estimate of the learning potential (asymptote) of the practice system. These norms provide context for individual-level curves in much the same way that a class average informs the interpretation of one individual's test score. If the learning curve relation does not hold at the group level (poor model fit), then it calls into question the entire system. On the other hand, if the relation holds at the group level, it does not mean that the system works for each learner. Individual-level curves, and the degree to which they fit (e.g. model r^2) allows determination of the inter-individual variability of the learning experience, valuable information for those responsible for managing the group's educational experience.

Assessment in medical education is often divided into that which serves formative or summative purposes (Epstein 2007). Learning curves add information on the *rate* and *path* of learning to an assessment, which is an advantage for each of these.

In formative assessment, learning curves can be used to individualize adaptive learning, thus providing motivation and direction for future learning. For a given level of prior knowledge, the number of repetitions required, on average, to achieve mastery could be estimated, allowing for individualized learning schedules. In current conceptualizations of summative assessment, the focus is on having sufficient psychometric rigor to allow a defensible binary decision as to whether a candidate has achieved a pre-determined level of competence (Epstein 2007). By adding rate information, a learning curve conceptualization changes the question from “will she/he pass?” to a question that is more consistent with a growth mindset: namely, “when will she/he pass?” which allows the full variety of learning paths, but still ensures the protection of the public.

The examples in this paper all show models of directly observed data so that the measurement errors in the predictors (principally the effort variable) and the performance measure are reflected in the modeling results. An alternative is the latent variable approach where the measurement error is taken into account allowing a potentially more accurate representation of the abstract construct of interest. In the radiology example, a latent variable representing “radiograph interpretation ability” would allow the modeling of unobservable learning trajectories with the errors deducted. This approach generally requires a larger sample size and meeting the other requirements of latent class models (Downing 2003). Another extension of the learning curve approach is to consider group and individual trajectories within the same “multi-level” model which, while somewhat more complex, has the added advantage of allowing exploration of group by individual interactions (Curran et al. 2010; Detry and Ma 2016; Singer and Willett 2003).

There are limitations that need to be kept in mind. A learning curve statistical model is only as good as the precision, reliability and validity of the measurements that go into it. For example, immediate measures of performance like those we describe in our examples are likely not as meaningful as those obtained after a retention interval (Dubrowski 2005). Other limitations more specific to our example data include the non-random convenience sampling and the fact that the participants generally completed the cases in a non-proctored environment with varying levels of engagement (Boutis et al. 2010). The fact that we do detect learning effects despite this real-world noise is hopeful. The radiology task lends itself to this approach as the unit of analysis (the case) is repeatable with a defined amount case-to-case variability. This may not be true of other health-professions skills where direct indices of performance may be defeated by the extent of the case variability.

In the end, learning curve modeling can be helpful in Health Professions Education where there are valid measures that allow discrimination of inter-individual variability in the individual’s learning trajectory whether psychomotor, knowledge or attitudinal in nature. Being able to describe that variability mathematically, whether in terms of the prior knowledge, rate of learning or maximum learning potential can provide the health professions educator or researcher with greater insight into their learning system.

In an online Supplemental Appendix we have listed the details of our data and the algorithms used to derive the models. We have limited the scope of this review to learning curves but we urge investigators to consider similar issues for forgetting or experience curves (Jaber 2006; Pusic et al. 2012).

Conclusion

We have shown a number of mathematical modeling techniques for demonstrating the degree to which a learning curve relationship holds for either groups or individual learners. The learning curve model can serve as a barometer of the efficiency and effectiveness of a learning system. Group level learning curve model parameters can guide the health professions educator by providing estimates as to how many repetitions are required, on average, for individuals to achieve mastery. Individual level mathematical models can make manifest the inter-individual variability in the rate of learning.

References

Ahlering, T. E., Woo, D., Eichel, L., Lee, D. I., Edwards, R., & Skarecky, D. W. (2004). Robot-assisted versus open radical prostatectomy: a comparison of one surgeon's outcomes. *Urology, 63*(5), 819–822. doi:10.1016/j.urology.2004.01.038.

Altman, D. (1990). Relation between several variables. In *Practical statistics for medical research* (pp. 325–364). London: Chapman & Hall/CRC.

Bolsin, S., & Colson, M. (2000). The use of the cusum technique in the assessment of trainee competence in new procedures. *International Journal for Quality in Health Care, 12*, 433–438. doi:10.1093/intqhc/12.5.433.

Boutis, K., Pecaric, M., Seeto, B., & Pusic, M. (2010). Using signal detection theory to model changes in serial learning of radiological image interpretation. *Advances in Health Sciences Education: Theory and Practice, 15*(5), 647–658. doi:10.1007/s10459-010-9225-8.

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve frequently asked questions about growth curve modeling. *Journal of Cognition and Development,*. doi:10.1080/15248371003699969.

Detry, M. A., & Ma, Y. (2016). Analyzing repeated measurements using mixed models. *JAMA, 315*(4), 407–408. doi:10.1001/jama.2015.19394.

Downing, S. (2003). Item response theory: Applications of modern test theory in medical education. *Medical Education*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1046/j.1365-2923.2003.01587.x/full.

Dubrowski, A. (2005). Performance versus learning curves: what is motor learning and how is it measured? *Surgical Endoscopy, 19*(9), 1290. doi:10.1007/s00464-004-8261-y.

Epstein, R. (2007). Assessment in medical education. *New England Journal of Medicine, 356*, 387–396.

Ericsson, K. A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine, 79*(10), S70–S81.

Ericsson, K. A. (2008). Deliberate practice and acquisition of expert performance: A general overview. *Academic Emergency Medicine: Official Journal of the Society for Academic Emergency Medicine, 15*(11), 988–994. doi:10.1111/j.1553-2712.2008.00227.x.

Flavio, S., Fogliatto, F. S., & Anazanello, M. J. (2011). Learning curves: The state of the art and research directions. In M. Y. Jaber (Ed.), *Learning curves: Theory, models, and applications* (pp. 3–22). Boca Raton, FL: CRC Press.

Jaber, M. Y. (2006). Learning and forgetting models and their application. In A. B. Badiru (Ed.), *Handbook of industrial and systems engineering* (pp. 1–27). Boca Raton, FL: CRC Press.

Jaber, M. Y., & Guiffrida, A. L. (2004). Learning curves for processes generating defects requiring reworks. *European Journal of Operational Research, 159*(3), 663–672. doi:10.1016/S0377-2217(03)00436-3.

Jaber, M. Y., & Sverker, S. (2004). A numerical comparison of three potential learningand forgetting models. *International Journal of Production Economics, 92*(3), 281–294. doi:10.1016/j.ijpe.2003.10.019.

Kalet, A., & Pusic, M. (2014). Defining and assessing competence. In A. Kalet & C. Chou (Eds.), *Remediation in medical education* (1st ed., pp. 3–15). Boston, MA: Springer.

Koo, E. J., Youn, S. H., Baek, Y. H., Roh, Y. H., Choi, H. J., Kim, Y. H., et al. (2012). Review of 100 cases of single port laparoscopic cholecystectomy. *Journal of the Korean Surgical Society, 82*(3), 179–184. doi:10.4174/jkss.2012.82.3.179.

Priestley, M. B., & Chao, M. T. (1972). Non-parametric function fitting. *Journal of the Royal Statistical Society. Series B (Methodological), 34*, 385–392. Retrieved from http://www.jstor.org/stable/2985075.

Pusic, M., Boutis, K., Hatala, R., & Cook, D. (2015). Learning curves in health professions education. *Academic Medicine: Journal of the Association of American Medical Colleges, 90*(8), 1034–1042. doi:10.1097/ACM.0000000000000681.

Pusic, M., Kessler, D., Szyld, D., Kalet, A., Pecaric, M., & Boutis, K. (2012). Experience curves as an organizing framework for deliberate practice in emergency medicine learning. *Academic Emergency Medicine, 19*(12), 1476–1480.

Ramsay, C. R., Grant, A. M., Wallace, S. A., Garthwaite, P. H., Monk, A. F., & Russell, I. T. (2001). Statistical assessment of the learning curves of health technologies. *Health Technology Assessment (Winchester, England)*, *5*(12), 1–79. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11319991.

Ramsay, C. R., Grant, A. M., Wallace, S. A., Garthwaite, P. H., Monk, A. F., & Russell, I. T. (2000). Assessment of the learning curve in health technologies. A systematic review. *International Journal of Technology Assessment in Health Care, 16*, 1095–1108. doi:10.1017/s0266462300103149.

Ritter, F. E., & Schooler, L. J. (2001). The learning curve. In N. J. Smelser, & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioural sciences* (1st edn., pp. 8602–8605). Oxford: Elsevier. Retrieved from http://ritter.ist.psu.edu/papers/ritters01.pdf.

Robertson, T. (1909). A biochemical conception of the phenomena of memory and sensation. *The Monist*, 367–386. Retrieved from http://www.jstor.org/stable/27900191.

Simonoff, J. S. (1996). Smoothing methods: A nonparametric/parametric compromise. *Smoothing methods in statistics* (1st ed., pp. 1–8). New York: Springer Science & Business Media.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.

Thurstone, L. L. (1919). The learning curve equation. *Psychological Review, 34*, 278–286.