A Data Analytics Framework to Support Training Effectiveness Evaluation

Chalinda Weerasinghe[1], Clarence Dillon[1], Paul Cummings[1], Michael Smith[1], Rodney Long[2]
ICF[1], Army Research Laboratory[2]

INTRODUCTION

As the Generalized Intelligent Framework for Tutoring (GIFT) framework becomes more pervasive as a research tool, the ability to evaluate student performance will be a critical consideration in its adoption. Having a common method for evaluating learner performance and course data in a simple user-intuitive way will aid course evaluators, instructional designers, and content managers to better assess the effectiveness of a course. In order to produce accurate measurement of learner performance within an intelligent tutoring system, a robust data analytics framework must include the ability to analyze course performance data, learner attributes, and learning strategies. Given the minimal amount of existing intelligent tutoring data, a tool for generating synthetic data would enhance the ability to build models and conduct robust experiments. This data can be used to establish and validate analytic findings, and develop baselines against which to compare observed performance. This paper introduces a data authoring tool and demonstrates its use.

The paper describes research that demonstrated an automated data generation capability that supports course (and performance) evaluation within the GIFT framework. We present a method to generate a simulated class using distributional properties that can be adjusted and visualized by the author. The data generation tool automatically creates dependencies between variables in a logical and consistent manner using a scale-invariant correlation measure (Kendall's tau, τ) from which data can be sampled and replicated. In addition to validating and calibrating the analytics engine, iterative simulations can provide expectations against which observed performance for new or modified GIFT instruction can be compared. The tool has been designed to allow intuitive inputs from users who do not possess strong statistical backgrounds by providing recommended settings and visual accompaniments to aide in data generation. These capabilities will be published to a web-based tool that will run alongside GIFT.

PREVIOUS RESEARCH

"The Changing Face of Military Learning," (Shatz, 2015) points to the complexity and rate of change in Army learning, and emphasizes the need to find new ways to empower learners. They specifically focus on increased investments in expanded set of training competencies, and developing more efficient and agile pathways to expertise. Methods for assessment and evaluation have been explored extensively in the psychology literature, but their application in the context of intelligent tutoring systems represents a less researched topic. The model for the analytical framework draws from foundational research published in Long et al. (2016a) that describes broad directions for analytics research in intelligent tutoring. Long, et al. (2016b) expanded on this research to develop a proof of concept system and user interface. This research explored the use of demographic factor analysis as a tool for performance effectiveness evaluation (Long et al., 2016b). The research successfully evaluated use cases to determine learning outcomes and gains, determinants of success, and recommendations for improvement, but also highlighted the need for a more robust data generation capability (Long et al., 2016b).

The current research builds on existing analytics framework development as described in Long et al. (2016b). Utilizing Basic Rifle Marksmanship as a use case, the research identified key demographic factors that formed the basis for simulated data. This research also introduced the application of the standard Kirkpatrick model for performance evaluation. Widely used in the realm of psychology and instructional systems design, this model posits four levels of evaluation that are used to evaluate performance (Kirkpatrick, 1994). The authoring tool focused on developing data to support what are known as Level 1, which focuses on the learner's reaction and satisfaction post-event, and Level 2, which focuses on the knowledge and skill gains that the learner exhibits.

METHOD

The research team simulated learner data following a process intended to imitate an objective (real-world) Learning Management System, Learning Records Store (LRS), and GIFT environment. An author, which could be an instructor, course designer, or researcher using GIFT, describes one or more learner "personas," which represent the types of learners who notionally populate a class. Persona attributes are user-defined characteristics of learners that broadly describe a group of learners: *e.g.*, "males, ages 18-24 from suburban, middle-class backgrounds and little previous job experience." We expect such data to be available in the objective learning environment: demographic and biographical data from learner profiles and biographical surveys; previous course and test performance from an LRS; and data from the current course, such as assessment scores and operationalized attitude surveys. In addition to variable names and the range of possible values (minimum/maximum or enumerated), authors provide information about distributional parameters (location, scale, and shape), using an interactive graphical tool. The user can specify bivariate dependencies between attributes, and the resultant joint distribution represents a narrative for each persona (see Appendix for screenshots of the authoring tool). Finally, the user specifies the mix of personas and the size of the class population: e.g., "class *XX* has 250 learners, 40% with *persona A* and 60% *persona B.*" The data authoring tool generates a heterogeneous set of learner profiles and learning records with statistical characteristics and dependencies that fit each persona generation scenario (collection of persona narratives).

These data sets form the basis for effectiveness evaluation utilizing the Kirkpatrick Model. The current focus is on Kirkpatrick Level 2, or student learning and knowledge gain. Also referred to as student performance, the most common Level 2 measurements are pre-test and post-test assessments given to the learners. With this simulated data, the factors of age and experience can then be analyzed with respect to these performance assessment outcomes. These measurements form the basis for automating the data analytics framework by providing a tool to assess relative impact on performance that is extensible to many different factors and assessment outcomes. Level 1 data, or learner reaction and satisfaction, is also readily available in GIFT and can be also be easily adapted to this data analytics framework.

Simulating Dependency

The research team utilized copula functions as an efficient method for describing bivariate joint probability distributions with uniform marginals (Joe, 1997; Nelsen, 2006). Sklar's theorem (Sklar, 1959) showed that any multivariate joint distribution can be broken into two components: the univariate marginal distribution functions that come together to make the joint distribution and the copula function that describes the dependency structure between these univariate distributions. The converse is also true: any set of univariate marginal distributions combined with an appropriate copula function will generate a multivariate joint distribution. Moreover, the multivariate joint distribution created is unique if the underlying marginal distributions are continuous. Even if they are not continuous, there is uniqueness over certain ranges of the marginals (Sklar, 1959). This research used the converse of Sklar's theorem to combine variables with arbitrary continuous margins with a copula family that described the desired dependency structure (described below) to form a unique, joint distribution. The tool currently implements the Archimedean class of copulas (Clayton, Frank and Gumbel families) due to their widespread use, flexibility and versatility

(Genest & Rivest,1993; Genest, Nešlehová & Ziegel, 2011).

Use Case

The research team tested data generated by the authoring tool by devising our own persona scenarios. We used methods and attributes developed in prior research (Long et al., 2016a; Long et al., 2016) and attempted to detect these personas and their narratives. The scenario is intentionally simple in an attempt to preclude spurious confirmation. We created *persona A* and *persona B*, each with four attributes: age (between 18-30 years); experience (0-10 years); pre-course assessment (pre-test) score (a proxy for general aptitude, between 0-100); and a predicted post-course assessment score (post-test, also on a 0-100 scale). In this use case we assumed all variables to be continuous.

In our narrative *persona A* learners were on average younger and less experienced, but scored higher on the post-test compared to *persona B* learners, who were older and more experienced, on average. Members in both A and B scored almost identically on the pre-test. Narratives for both *A* and *B* learners included dependencies across some attributes, which we describe next in detail, along with the distributional assumptions used for the simulation, summarized in Table 1.

Table 1. Persona Attributes, Distributions, and Parameters

	Persona A		Persona B	
attribute	*distribution*	*parameters*	*distribution*	*parameters*
age	normal	$\mu = 23$, $\sigma = 1.5$	normal	$\mu = 26$, $\sigma = 1.0$
experience	lognormal	$\mu = 3$, $\sigma = 0.1$	lognormal	$\mu = 4$, $\sigma = 0.1$
pre-test	normal	$\mu = 80$, $\sigma = 5.0$	normal	$\mu = 80$, $\sigma = 5.0$
post-test	normal	$\mu = 86$, $\sigma = 3.0$	normal	$\mu = 80$, $\sigma = 2.0$

Persona A learners' age and experience were positively correlated (moderate to strong, Kendall's $\tau \approx 0.6$), but at lower ages there was lower experience with more variation (more probability of falling at this end), while at higher ages there was higher experience with little variation (less probability of falling at the high end). The pre-test score was independent of age and experience. Instead, their pre-test score was strongly, positively correlated with post-test assessment ($\tau \approx 0.85$) so that those who scored lower on the pre-test scored lower with more variation on the post-test (more probability of falling at the low end), while those who scored higher at the pre-test scored higher in the post-test with little variation (less probability of falling at the high end).

Persona B learners' age and experience were weakly and positively correlated ($\tau \approx 0.3$). But at lower ages, there was lower experience with less variation (less probability of falling at this end) and at higher ages there was higher experience with more variation (higher probability of falling at the high end). The pre-test score was independent of age and experience for *persona B* learners as well. Their pre-test score was again strongly positively correlated with post-test assessment ($\tau \approx 0.8$), with those who scored lower on the pre-test scoring lower with more variation (more probability of falling at this end). Those who scored higher on the pre-test scored higher in the post-test with little variation (less probability of falling at the high end). Figure 1 shows these two personas and their narratives.

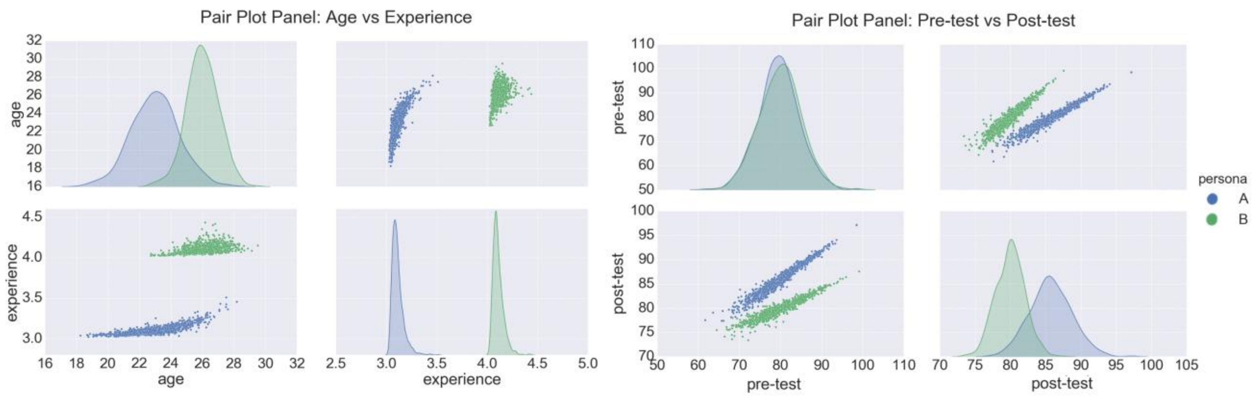

Figure 1. Sample of Attribute Clusters by Persona. From top-left to bottom-right, left to right across: the distributions for age; the scatterplot of experience and age; the distributions for pretest scores; the scatterplot of posttest and pretest scores; the scatterplot of age and experience; the distributions for experience; the scatterplot of pretest and posttest scores; and the distributions for posttest scores.

RESULTS

We present our results based on three levels of analytical validation of the simulated data. First, we review the extent to which the two personas can be recovered through analysis. Next, we validate between-variable relationships. Finally, we validate the parameters within variables, compared to the specifications for simulation. The statistical engine has an overall architecture that uses a three-level analytical design where incoming data is first examined for overall patterns using data mining tools, while uncovering underlying data distributions. At the next level the random variables are analyzed for underlying structures and relationships across variables, and finally within variable data patterns are analyzed.

Level 1 Validation

At the first level, the engine tried to recover the personas that the aggregate data described. The two personas were unique, and using two unsupervised learning cluster methods (k-means and agglomerative hierarchical) the system tried to identify the personas. Originally, the training set had 1000 students each for A and B (the class had 2000 students). Initially, k was specified as 2 (Table 2), and the system classified the observations accurately with an error rate of 0.0075 (15 out of the 2000 students were misidentified), where A had 15 students under-classified (data was standardized). The hierarchical clustering was able to identify the two clusters as well, while noting that one of the clusters could also be broken into two at a lower level. Table 2 also presents the deviations from total averages for the two clusters across the four variables, showing that our clusters represent two very different groups. It also validates persona scenario. The r-squared was moderate (53.3%), but improvements can only be achieved with higher classification error. Currently, code is being developed to locally optimize on the classification and find the best k out of a set of values (k<10) that fits the data if the training set's cluster assignment is known.

Once this was done the engine created histograms/kernel densities for the variables (see Figure 1 above), while generating the empirical distributions. The empirical distributions will also be used for data replication purposes. The densities clearly showed that the generated personas had the variable distributions prescribed to them. Even though groups A and B had distinct unimodal distributions when created, the aggregate data was unimodal only for the pre-test variable, as per the persona generation scenario. The Shapiro-Wilk, Kolmogorov-Smirnov, and Anderson-Darling tests showed that the persona variables followed their specified distributions (not shown).

Table 2: Level 1 Analysis Results

	cluster		error rate
persona group	*A*	*B*	
size	895	1015	0.0075
relative size	0.493	0.508	
R^2			0.533

deviations from total average		
persona group	A	B
age	-0.782	0.759
experience	-0.999	0.970
pre-test	-0.015	0.014
post-test	0.767	-0.744

Level 2 Validation

At the second level of analysis, the system tried to recover the between-variable relationships that were built in (Table 3). Kendall's τ tests correctly identified all four bivariate copula correlation values (with repeated simulation to validate these findings) to a very high degree of accuracy (p-values < 0.00001) within the two personas, while also noting that all of the other variables were not correlated with each other ($\tau < 0.03$ for all of them). For the most part, in the aggregate data the class showed dependencies that would have been anticipated given the persona narratives and overall scenario. The correlation between age and experience was $\tau \approx 0.68$, while between pre-test and post-test it was $\tau \approx 0.45$ (p-values < 0.00001). The only other significant result was a moderate negative relationship between experience and post-test ($\tau \approx -0.45$, p-value < 0.00001), which was part of the generation scenario since those in A (lower experienced) did better than those in B (more experienced).

To examine an explanatory model, multiple linear regression was used after validating some of the Gauss-Markov assumptions with post-test as the dependent variable. Table 3 shows model results for the entire class as well separately for the two personas. As expected, the linear model performed very well to the data with high goodness-of-fit measures. The model had low multicollinearity (variance inflation factors were all less than 2.75), which validated the linear structure used in the copula creation code with a random component introduced. All of the coefficients were positive and significant with the exception of experience, which had the largest negative effect on post-test scores (all p-values < 0.00001). The magnitude of the effect was not anticipated but also not surprising given the negative relationship between experience and post-test score, and the fact that experience was following lognormal distributions that were sufficiently far apart (see Table 1). Overall, the regression results validated the persona generation scenario, with all three variables explaining post-test scores well, and the model being a very good fit to the data.

Table 3: Level 2 Analysis Results

attribute pair	Correlation Results Persona A	Persona B	Aggregate Data
age - experience	0.5859*** (0.0211)	0.3253***(0.0211)	0.6757*** (0.0149)
pre-test - post-test	0.8618*** (0.0211)	0.8045***(0.0211)	0.4469*** (0.0149)
experience - pre-test	−0.0173 (0.0211)	0.0231 (0.0211)	0261* (0.0149)
experience - post-test	−0.0209 (0.0211)	0.0273 (0.0211)	−0.4444*** (0.0149)

The Kendall's tau value is given with the standard errors in parentheses, with $^{*}p < 0.1$ and $^{***}p < 0.0001$. Correlations used to build copulas appear in bold type. Not all correlations are presented; only those important to the results.

measure	Regression Results Persona A	Persona B	Aggregate Model	VIF
intercept	41.0953*** (1.5298)	48.6993*** (1.6876)	63.8740*** (0.4006)	
age	0.0319 (0.0226)	0.0221 (0.0225)	0.1573*** (0.0163)	2.5739
pre-test	0.5791*** (0.0043)	0.3815*** (0.0041)	0.4786*** (0.0039)	1.0017
experience	−0.7045 (0.5942)	0.0471 (0.4424)	−6.3875*** (0.0647)	2.7566
num. obs.	1000	1000	2000	
adj. R^2	0.9472	0.8945	0.9484	
std. error	0.6823	0.6666	0.8741	
F statistic	5978.4951 (< 0.0001)	2825.3421 (< 0.0001)	12250.23 (< 0.0001)	

The dependent variable is the post-test score. Coefficients are given with standard errors in parentheses for the variables, with $^{***}p < 0.0001$. All regressions have robust estimates. Parenthesis with the F-statistic indicates the p-value. VIF is variance inflation factor. Standard error is the standard error of the regression.

Level 3 Validation

The final analysis that was done was at the third level, where within variable statistics were examined to see if they followed variable parameters that were specified (Table 4 is for the entire class). The descriptive statistics that were generated showed that at the persona level the variables approximately had the location and scale parameters that were specified (not shown), while at the class level the data fell within the overall scenario bounds and had statistics that were expected (Table 4). Data graphing capabilities (box-plots) visually validated these findings. Overall, the data validation exercise was a success, with the statistical and analytical engine being able to recover the built in data and variable patterns dictated by the scenario modeled.

Table 4: Level 3 Analysis Results

measure	age	experience	pre-test	post-test
mean	24.5159	3.6106	79.8313	82.8793
std. error	0.0445	0.0112	0.1128	0.0861
median	24.8664	3.7675	79.8941	82.2737
std. deviation	1.9902	0.5018	5.0424	3.8483
skewness	-0.5890	-1.9553	0.1899	-0.4358
range	11	1	37	24
maximum	30	4	99	97
minimum	18	3	62	73
IQR	3.0699	0.9989	6.4993	5.7788
count	2000	2000	2000	2000

DISCUSSION

The results indicated that the automated analysis tool for GIFT successfully detected the persona groups from the learner population at each level of validation. Based on a number of measures these results were consistent and demonstrated statistical significance. Given the success of this relatively simple model, the data authoring tool can be used to develop more complex interdependencies for larger numbers of variables. Copulas have been demonstrated to be a highly scalable tool for modeling relationships with a number of desirable statistical properties. These results can also be scaled easily to apply to tens or even hundreds of thousands of students if operated on the right system architecture.

Clustering is a flexible tool that can be automated in a deployed environment to automatically categorize real-world data. This will allow insights to be delivered in a more seamless fashion by grouping outputs and students into groups that make sense intuitively. For example, an analytic output of the system could be that experienced students are performing poorly compared to inexperienced students and indicating lower levels of satisfaction. While this result is interesting in and of itself, an automated analytic tool-kit will need to define and identify these groupings and apply them to real world students, a capability demonstrated in the data authoring tool.

The data authoring tool and the automated performance analysis tool are both web-based applications, designed to support learning in a GIFT environment. The results can be used immediately in a GIFT environment to develop notional course performance data for authoring. The results can be used to implement a persona-based, baseline data authoring system. The models can be deployed within any Python environment and push data to a web-based interface. The data authoring tool and analytics engine could both be more useful if GIFT had a native capability to export course-level learner data that includes learner performance evaluations (scored and graded assessments). This would make real-time comparison of learner performance against subgroup baselines possible.

Results form the basis for a reusable testbed for learning analytics research. Results can be used in any environment with assessment data prior to employment. Simulation can be utilized post-employment as a kind of virtual laboratory to generate best case scenarios and isolate specific factors without the noise and messiness of empirical data. These outputs can then be used to calibrate analytics deployed in a number of different settings.

CONCLUSIONS AND FUTURE RESEARCH

Success here forms the basis for implementing and validating a robust, automated analytics capability for GIFT, research that is underway at the time of publication. By hypothesizing, identifying, and validating statistical relationships in a laboratory like setting, the data authoring tool has blazed a clear path to creating reproducible analytical results using empirical GIFT course data. By developing synthetic benchmark data, researchers and experimenters will be able to test their own theories and analytic applications in the GIFT framework with a tool fit for the intended environment. Once completed, the data authoring tool can be used in conjunction with the analytics engine to develop a continuous feedback loop to improve performance.

While initial development focuses on descriptive and inferential analysis, future extensions will easily transition to predictive applications. By developing and validating statistical relationships across increasingly complex sets of variables and interactions, the analytics engine can develop sophisticated learner models that can be evaluated with real-world GIFT interaction data. Once validated, the calibrated models can be used to rank success factors and identify problem areas earlier in the instructional process. For example, a notional data set depicting a relationship between education and performance could be developed in the authoring tool. Real world data could be used to validate this relationship, and predictive measures could be employed to change instructional strategies for these students. The benefit of an authoring environment is that while these relationships will play out across many variables over time, the individual effects can be isolated, studied, and ultimately employed to improve instruction.

Even though only bivariate copulas were used in the simulation, code is being developed to use multivariate copulas (using vines) in more general dependency settings (Bedford & Cooke, 2002; Kurowicka & Joe, 2011). Even though the code does not currently permit recovery of the underlying copula structures from the data, while only being able to obtain the bivariate dependencies that were built in, the vine code that is being developed will be able to identify multivariate copula structures in addition to identifying dependency strength.

Finally, while the initial demographic analysis variables are limited, it is straightforward to scale up and include more variables and inputs in the data authoring tool. Extending the simulation to include competency modeling and concept mapping is one such natural extension of this capability. In addition, granular event level data such as user activity in the Event Reporting Tool could yield additional predictive insights, as has been highlighted in other learning analytics research such as Chatti et al. (2012).

REFERENCES

Bedford, T.J. & Cooke, R.M. (2002). Vines- A new graphical model for dependent random variables. *Annals of Statistics 30*(4): 1031–1068.

Chatti, M.A, Dyckhoff, A.L., Schroeder, U., & Thüs, H. A (2012). Reference Model for Learning Analytics. *International Journal of Technology Enhanced Learning 4*(5/6), pp. 318-331.

Genest, C., J. Nešlehová, & J. Ziegel (2011). Inference in multivariate Archimedean copula models. *Test 20*(2), 223–256.

Genest, C. & L.-P. Rivest (1993). Statistical inference procedures for bivariate Archimedean copulas.

Journal of the American Statistical Association 88(423), 1034–1043.

Joe, H. (1997). *Multivariate Models and Dependence Concepts.* London: Chapman & Hall.

Kirkpatrick, D. L. (1994). *Evaluating training programs: The four levels.* San Francisco: Berrett-Koehler.

Kurowicka, D. & Joe, H. (2011). *Dependence modeling: Vine copula handbook.* World Scientific, Singapore.

Long, R., Smith, M., Dass, S., Dillon, C., & Hill, K. (2016a). *Data Analytics: Techniques and Applications to Transform Army Learning*. ModSim World Conference, NDIA, 2016.

Long, R., Smith, M., Dass, S., Dillon, C., & Silverman, J. (2016b). *Authoring Tools to Assess Training Effectiveness in Intelligent Tutors.* I/ITSEC Conference, NDIA, 2016.

Nelsen, R. B. (2006). *An Introduction to Copulas.* New York: Springer. Chapter 4.

Shatz, S. (2015). *The Changing Face of Military Learning*. Interservice/Industry Training, Simulation, and Education Conference

Sklar, A. (1959). Fonctions de r´epartition ´a n dimensions et leurs marges. Publl. Inst. Statist. Univ. Paris, 8, 229-231.

ABOUT THE AUTHORS

Chalinda Weerasinghe *is a research consultant who specializes in International Relations, Econometrics, Political Economy, and Development Economics. He undertook doctoral studies in Economics and Government and Politics at the University of Maryland, College Park with specializations in econometrics, international relations and development economics. He has an MS in Economics and an MS in International Relations from the Georgia Institute of Technology and a B.S. in Mathematics, Economics and History and Political Science (triple major) from Shorter University.*

Clarence Dillon *is a computational social scientist and Consultant at ICF. He authored the ontology for the Department of Defense strategic planning scenarios and established the first collaborative, semantic web platform used in DoD. Mr. Dillon holds a Bachelor of Arts in International Affairs, a Masters of Social Science in International Relations and he conducts graduate research in social complexity at George Mason University.*

Paul Cummings *is Senior Fellow in the Center for Advanced Learning Systems at ICF. He has over 19 years of technical and management leadership experience in the education, simulation, and training community. Mr. Cummings currently develops immersive technology systems for a wide range of clients. Mr. Cummings is a PhD student in the Krasnow Institute Computational Social Science program.*

Mike Smith *is a Senior Technical Specialist with ICF and has over 13 years of experience in data analytics, strategic planning, and management. Mr. Smith advises government clients on how to adapt analytics practices to improve their organizational performance. Mr. Smith has a B.A. in International Economics from Longwood University and a Master of Public Policy from Georgetown University.*

Rodney Long *is a Science and Technology Manager at the Army Research Laboratory in Florida and is* *currently conducting research in adaptive training technologies. Mr. Long has a wide range of simulation and training experience spanning 29 years in the Department of Defense and has a Bachelor's Degree in Computer Engineering from the University of South Carolina and Master's degree in Industrial Engineering from the University of Central Florida.*

APPENDIX

This appendix includes screen-capture images showing the user interface for creating Personas and identifying correlations between Persona attribute variables.

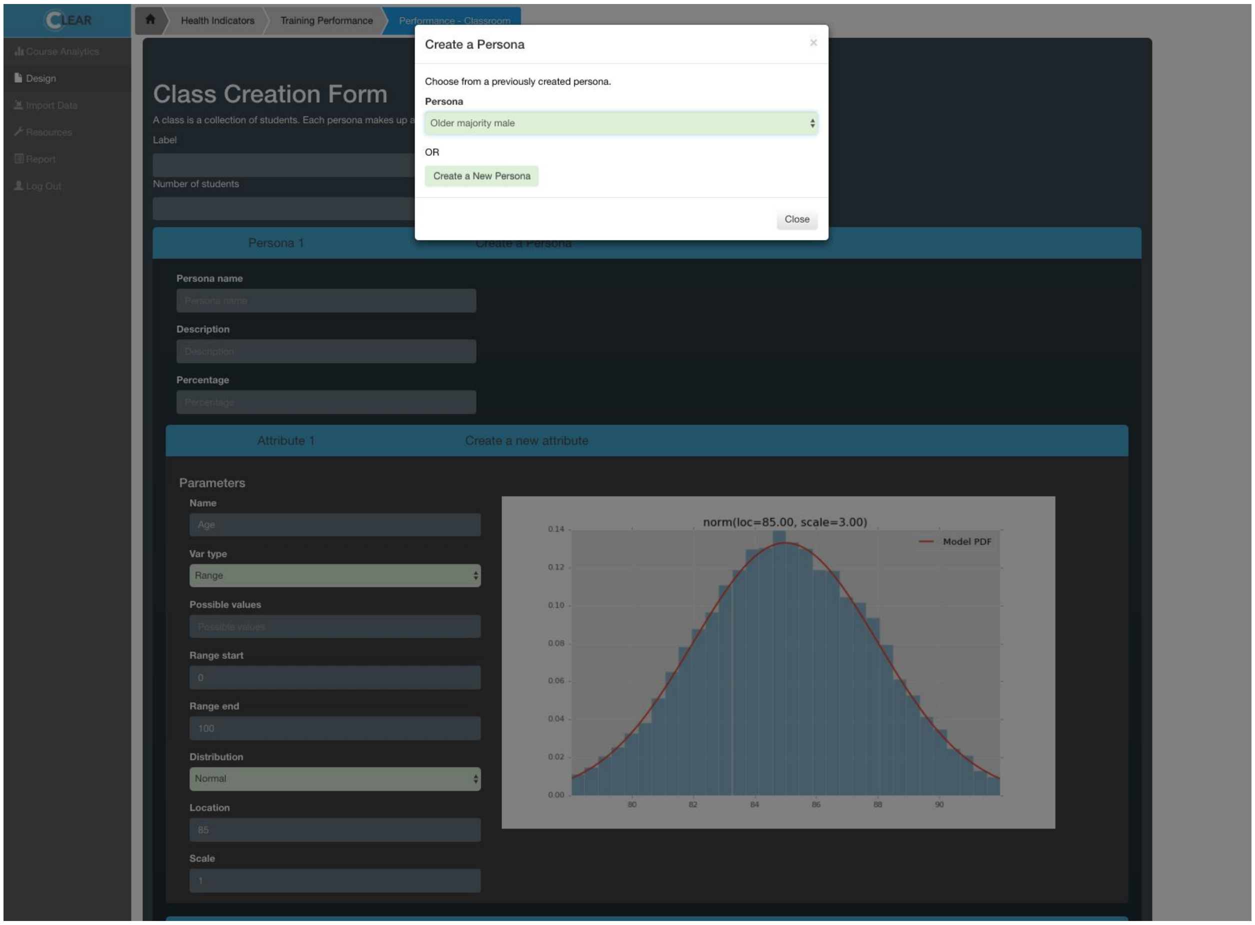

Figure 1. From the Class Creation Form, a user can design a persona and visualize the patterns of each attribute by itself. Form options are intended to be non-technical and to guide a user through the process in preparation to designing correlations in the next form. The user types a name for the attribute and uses a selection box to choose the type of variable—whether it is from the learner profile, the LRS, a biographical survey, or the current course. The user can choose if the attribute is a discrete variable with enumerated options or a continuous variable with minimum and maximum values over a range. The user can select the shape of the distribution from several options (normal, half-normal, lognormal, gamma and Chi-squared). After selecting the shape of the distribution, they select the location (mean) and scale (standard deviation).

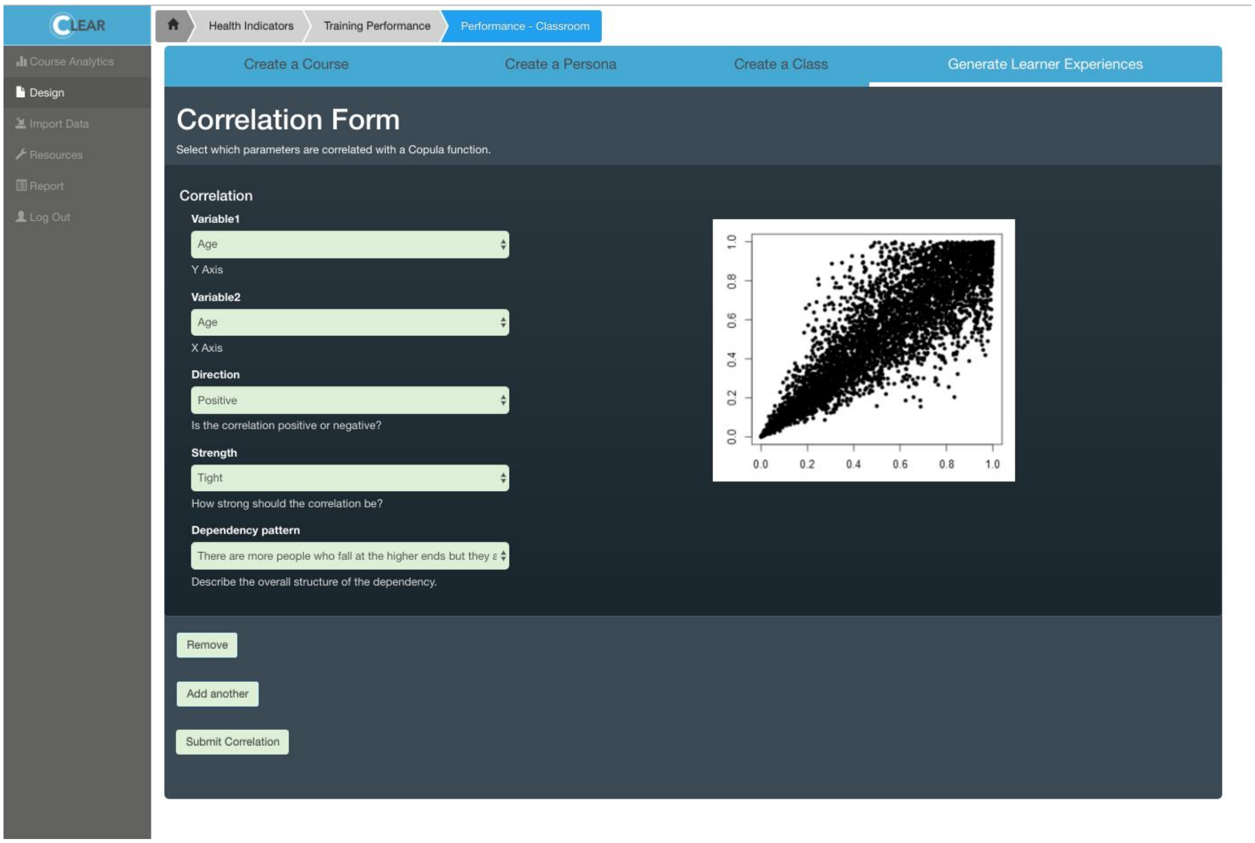

Figure 2. The Correlation Form guides a user through designing joint distributions between any two persona attributes. Options include the direction of correlation (positive or negative), the strength of the correlation (tight, moderate or loose) and the general pattern (tighter at the bottom and the top, tighter at the bottom and looser at the top, or looser at the bottom and tighter at the top). In the background, the application associates these descriptions with Frank, Clayton and Gumbel copula families, respectively. The plot to the right is visual example of the correlation generated by the form field parameters.