Modest analytics: using the index method to identify students at risk of failure

Tim Rogers
University of South Australia
Learning and Teaching Unit
Adelaide, South Australia
tim.rogers@unisa.edu.au

Cassandra Colvin
University of South Australia
Learning and Teaching Unit
Adelaide, South Australia
cassandra.colvin@unisa.edu.au

Belinda Chiera
University of South Australia
School of Information Technology and Mathematical Sciences
Adelaide, South Australia
belinda.chiera@unisa.edu.au

ABSTRACT
Regression is the tool of choice for developing predictive models of student risk of failure. However, the forecasting literature has demonstrated the predictive equivalence of much simpler methods. We directly compare one simple tabulation technique, the index method, to a linear multiple regression approach for identifying students at risk. The broader purpose is to explore the plausibility of a flexible method that is conducive to adoption and diffusion. In this respect this paper fits within the ambit of the *modest computing* agenda, and suggests the possibility of a *modest analytics*. We built both regression and index method models on 2011 student data and applied these to 2012 student data. The index method was comparable in terms of predictive accuracy of student risk. We suggest that the context specificity of learning environments makes the index method a promising tool for educators who want a situated risk algorithm that is flexible and adaptable.

Categories and Subject Descriptors
G.3 [**Probability and Statistics**]: Correlation and regression analysis; robust regression

General Terms
Algorithms, Experimentation, Theory.

Keywords
Predictive models for student performance, modest computing, index method, regression.

1. INTRODUCTION
Given the increasing quality and efficiency pressures facing universities it is not surprising that learning analytics has placed considerable focus on the early identification of students at risk of failure [7, 9, 17]. The prediction, or forecasting, of students at future risk has consequently been a high institutional priority [22, 27]. Following the social sciences generally, the predictive method of choice has been regression analysis (of one kind or another) using a constellation of demographic (static) and engagement (dynamic) variables, the latter usually emanating from the LMS traces students leave as they navigate the online elements of their courses [e.g. 18].

This common reliance on regression is entirely uncontroversial, but its hegemonic status in many branches of the social sciences has served to sideline other promising approaches that are far simpler and much easier for people from non-statistical backgrounds to understand [2, 19, 20]. As Armstrong [3] put it "Regression analysis is clearly one of the most important tools available to researchers. However, it is not the only game in town". We present here an investigation of the merits of a plausible, simpler alternative to regression for identifying students at-risk that may be a better tool if our aim is not only predictive accuracy but also organisational diffusion and the embedding of change.

This paper seeks to extend the concepts and framework of *modest computing* [13, 14] to analytics. There are several reasons to favour the parsimonious and simple predictive approach over the expensive and complex, but perhaps the most important is that the simple is often a better predictor than the complex, or at least comparable or practically equivalent, as the 'fast and frugal heuristic' literature has documented [20]. 'Practically equivalent' is a key term here. The aim of the modest computing agenda is not simplicity for simplicity's sake, but to invest in high leverage initiatives and methodologies that will have significant impact for relatively minimal investment. Whatever the technique, it should be a means to the end of *orchestrating* [13] the educational interface, not an end in itself. Modest computing seeks rather "to emphasize adoption, scale and sustainability more– to shift from pure invention to a "diffusion of innovation" perspective." [13]. As attested to by a wide range of organisational change and adoption approaches [e.g. 24, 28], when the mechanisms on offer are easy to adopt, configurable at the user's end, and solve an actual practitioner problem, they are much more likely to be implemented, and implemented as they were intended. Conversely put, it is to avoid, where possible, hermetic, abstruse, overly complex approaches, especially those that serve the interests of the researchers at the expense of the practitioners, where "complex methods become an end in themselves, a ritual to impress others, and at the same time opportunities to learn how to do things better are missed" [20].

In exploring the potential of extending these modest computing principles to analytics, this paper marks a preliminary assessment of the possibility of a *modest analytics*. To this end we directly compare the utility of the *index method*—a simple tabulation technique—with the kind of regression analysis typically used to identify students at risk. What follows is a proof-of-concept, an

LAK'14, March 24–28, 2014, Indianapolis, Indiana, USA.

http://dx.doi.org/10.1145/2567574.2567617

empirical head-to-head between an established regression process and a lesser known, but far simpler, alternative approach.

2. INDEX METHOD

The forecasting literature shares two of learning analytics' core goals, to find the most effective means of identifying a future event or estimating a future trend. In educational terms these two goals may be expressed as, for example, the need to identify which students are likely to struggle in their studies and mapping the cumulative long term impacts of curriculum innovations on educational outcomes. Overall, the forecasting literature displays a catholic and utilitarian approach to the exploration of techniques and their combinations. For example, the realisation that expert judgement, if augmented, could still play a role in forecasting accuracy [8] led to a reassessment of previous research, along with a number of new studies that methodically combined judgements, or used them in tandem with statistical approaches. Such approaches as the Delphi technique and judgemental bootstrapping demonstrated advantages over numerical only, or expert opinion only, approaches in predicting outcomes in a range of practical areas [1, 2, 20].

Perhaps the most intriguing, and certainly the simplest, technique documented is the 'index method', often called the 'unit weight' approach. The method appears almost childishly uncomplicated, and hence unlikely to be of much consequence. Its origins, according to Armstrong [3], are with Benjamin Franklin, who essentially suggested the best way to make a difficult choice between two alternatives was to divide a piece of paper in half, on one side list the pros and the other the cons on the full range of known criteria, weight the criteria, and then tally up the resulting scores to make a selection.

Where the index method has been tested against various predictive alternatives, typically regression analyses, it has usually matched and in many cases bettered these more commonly used and more complex alternatives [4, 5, 11, 12, 16, 29]. For example, Armstrong and Graefe [5] compared the index method to a range of well-known regression approaches on the outcome of 29 US Presidential elections (1896 to 2008). They derived each presidential candidate's likelihood of victory simply by tabulating the accrued head-to-head score for one candidate over another on 60 biographical variables, from candidate height, to age, weight, tone of voice, home state and so on. They found that in terms of binary outcome (who won the election) the index method fared better than all 3 well-established regression approaches it was pitted against. In terms of predicting the vote share of the candidates, the index method outperformed all but one regression study, and here the difference was marginal.

Dana and Dawes [12] looked at a variety of real world distributions and compared regression coefficients to a range of alternative, and simpler, coefficients, including unit weights. They found that the "alternative prediction methods presented in this article quite often defeated ordinary least squares in our competition" (p. 328). With respect to the unit weight method, they found that "Although equal [unit] weights were not the most efficient of the alternative methods, they were practically as good. The decrement in explainable variation on validation caused by using unit weights as opposed to the best coefficients rarely exceeded 2% or 3%... In Wainer's (1976) words, estimating coefficients apparently "don't make no nevermind," or at least it doesn't for most social science data" (p. 324).

Despite the index method's comparative success, it is rarely reported in any literature and presumably rarely used. This is a pity, as its pragmatic value alone warrants further investigation. It has occasionally been used to estimate student performance vis-à-vis regression, but the results have been mixed; Armstrong [4] found 3 studies favouring the index method and 1 favouring regression. To our knowledge the method has not been used to predict student outcomes since 1974. Again, this is a pity because the index method, if it does indeed show itself to be a valid substitute for more complex risk algorithms, could provide a quick, flexible, practitioner-centred, and adoption friendly indicator of student risk that would be of great utility. For example, if educators alter their curriculum by emphasising or de-emphasising particular elements, such as online discussions or formative tests, they would be able to immediately update their student risk analysis to reflect these changes. This would be very difficult to do with an institutionally mandated, or commercial, risk algorithm. This would be possible with the index method because:

- Unlike regression analysis, the index method does not estimate weights from the data, so sample size is not a limiting factor. Small samples (say a previous class cohort) and limited observations per variable are no barrier to its use [5, 16, 29]
- There is no limit to the variables that can be incorporated, so the full range of domain knowledge can be brought to bear. This contrasts with regression models, where increasing variables can decrease the model's ex-ante predictive accuracy, given a constant sample size [5, 16, 20]
- It is straightforward to add or subtract variables to reflect new understandings/empirical findings and with no consequence to the method's integrity [5]
- It can incorporate judgement, and so context. Where data is scarce but expert opinion exists about the relevance and directionality of a variable (i.e. whether the variable is favourable or unfavourable for the relevant outcome), it can be included [3]
- It is easily understood and 'owned' by educators. They do not need to know any statistical theory to apply the method, although it helps to be able to critically evaluate prior studies that may be relied upon for help in identifying the variables to be included.

3. METHOD

3.1 Data

Four first year courses at the University of South Australia were chosen on the basis of their previous engagement with the early intervention program 'Enhancing Student Academic Potential' (ESAP). Each course had previously been identified as having a high percentage of at-risk students relative to other courses in their discipline area. The courses were from a range of disciplines: Accounting, Computer Science, Biology, and Business/Law. There were two years of data drawn, 2011 and 2012. 2011 was used to develop models that would then be applied to 2012. There were 1230 students in the 2011 data set and 1102 students in 2012.

3.2 Index Method

We selected the independent variables *a priori*, using the student attrition and learning analytics literature, asking the coordinators of the courses for their suggestions, and by eyeballing the data.

We only included variables that could be drawn from one of our online systems. The majority of the variables were demographic and performance based. Only one variable was dynamic – first logon to the course LMS. Other dynamic data was unreliable or proved too difficult to extract.

Each variable for each student was initially given a score of -1, 0, +1 based on its hypothesised effect on students' chances of passing the course. A score of zero meant the variable was neutral with respect to passing. Where knowledge of the effect of variables is good weighting can be applied [16] and we did this for GPA and for any previous notice of poor academic progress (an academic warning sent to students). Partly these weightings reflected the number of evaluations of the student implied by the variable. For example, a student on a second academic notification would have shown poor prior performance on 4 previous occasions in most instances, so the weighted score on this variable was 4. Each student's score was calculated by adding across the columns of variables where they achieved a positive or negative score. Scores were given for: gender, non-English speaking background, repeating the course, being enrolled full-time, for having no parent completing school, for having had a deferred exam in the study period just past, for failing, withdrawing late or only barely passing any course in the study period just past, for a fail or bare pass in a pre-requisite course, for late LMS log-in (post 7 days from course start), and, as stated above, low GPA or academic notice of poor progress. We used 2011 student data to set the index score and used a simple linear regression to develop a model to test on 2012 data for the same courses. This regression achieved an R^2 of 0.38 for the 2011 data. The final predictive model was:

Mark = -2.91(Index score) + 57.82

Applying the index model to the 2012 data to generate predictions for each student's 2012 score yielded a significant correlation in the expected direction of r = -.58, ($p < 0.05$). This result is illustrated in Figure 1 below, where predictions of the 2012 student marks were made using the index method and are shown in green, overlaying the actual marks (blue), with a correlation of 0.58 between the two series

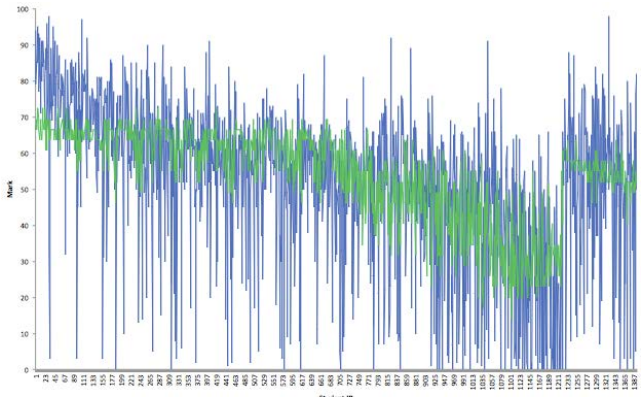

Figure 1: the actual (blue) and predicted (green) student marks are shown for the index method.

3.3 Regression

In contrast to the approach taken for the index method, the regression model was constructed through an experimental exploration of the independent variables, using backward elimination, to model student marks as the explanatory variable. A collection of both quantitative and qualitative independent variables was considered, where at each step, the variable with the smallest t value (significance at 0.05) was removed. The final predictive model was:

Mark = 16.71 - 2.64(Gender) + 0.29(Age) - 4.87(AcademicLoad) + 0.11(EarnedUnits) + 6.44(GPA)- 7.74(WithdrewPreviously) + 3.92(EnrolledNextYear) - 19.88(Counselling) - 17.38(PreviousNoticePoorProgress) (1)

which reported an R^2 value of 0.57, indicating the independent variables explain approximately 57% of the variability in Mark. The independent variables captured the student's Gender, their age, an Academic Load of full- or part-time, the number of units earned in their program to date, their GPA, whether they had withdrawn late from a course in the previous study period, whether they were enrolled in the following year, whether they had been recommended for academic counselling and whether they had been given a notice of poor academic progress.

Predictions of the 2012 student marks were made using (1) and are shown in green in Figure 2, overlaying the actual marks (blue), with a correlation of 0.70 between the two series. Testing of the model residuals indicated no severe issues with the assumptions of linearity, homoscedasticity, independence and Normality.

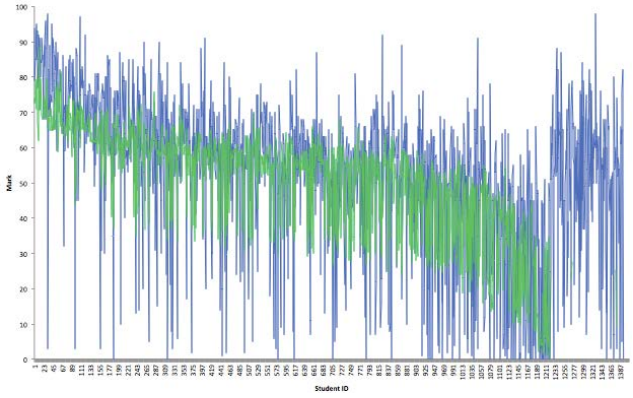

Figure 2: the actual (blue) and predicted (green) student marks are shown for the regression analysis[1].

4. DISCUSSION

In this comparative study we pitted a standard linear multiple regression based approach to identifying students at risk of failure against a simpler approach, the index method. Our findings are consistent with the preponderance of previous studies in showing the index method to be competitive with regression: we found the correlation coefficients for the model-derived scores and the actual 2012 scores for both regression (r = 0.70) and the index method (r = - 0.58) to be significant and largely in line with each other. While the regression approach yielded a higher correlation it should be noted that it also produced more variability in its predictions (which can be seen in the green regression model predictions spanning a larger range than the index model predictions). What the index model lost in correlation it picked up

[1] The gap in predicted marks on the right hand side due to variables used for prediction in the regression model having missing values at that end of the data set

in reduced variability – an important consideration when the purpose is prediction.

There are various explanations for this phenomena reported in the literature. The most prominent concerns the robustness of regression models i.e. the ability of the model to both fit the data used as a source for the model and predict in the novel population for which the model was built. Although in comparative studies regression is better at fitting the data than alternatives, it is a poorer predictive instrument in out-of-sample populations, which suggests overfitting of the data is involved [11, 20], although that explanation does not apply here. Two other explanations centre on the amount of information that the index method permits relative to regression. Einhorn and Hogarth [16] suggest that unit weight approaches are able to include more information as the included variables are not estimated from the data and so do not consume degrees of freedom, "therefore, one can have as many variables in the prediction equation as necessary" (p. 184). Several researchers have suggested that it might be the identification of the variables, rather than their weighting, that is important in *ex ante* predictions. In this case the power of the index method rests largely in the *a priori* selection of variables and their signs based on the hard empirical and theoretical work of earlier researchers [12, 16].

While in terms of linear prediction this paper found the regression approach to be marginally more accurate, this paper was not intended to be a reproof of linear multiple regression, which we acknowledge is a valuable technique. It was intended rather to pave the way for the exploration of simpler models that may facilitate the adoption and diffusion of risk management in educational settings. Simpler techniques may have these properties because they are better understood by a wide range of educators and are easy to adjust to changing circumstances and integrate with current practices.

There are several limitations to this study. Our test procedure used one year's data for model development, which is an artificial constraint not likely to be faced by model developers in situ. We attempted to address this shortcoming by cross-validating the regression model by using the 2012 data to construct a predictive model. The result was a model containing the same explanatory variables as in (1) with similar regression coefficients. Using more than one year's data for model development may (or may not) improve the robustness of the regression model, a possibility we will test in the future. We note, however, that the index method suffers from no such limitation, in that the model's robustness is free from the requirement of many years of data and with that, the assumption that the student cohort behaviour will remain stable over time, which is ultimately an intrinsic assumption of any regression model. We also used very few dynamic variables (e.g dwell time and other LMS data, physical class attendance, online library interactions etc.) as the extraction of this data at our University is still under development. The inclusion of this data may alter our findings, although the addition of a number of new variables could be expected to favour the index method [5]. The choice of an additive linear model for comparison does not exhaust the regression technique possibilities. Other approaches may yield different results. Our choice was based on the standard approach followed in the literature that has previously tested index and unit weight approaches, and findings that suggest complex and nonlinear models do not offer a greater predictive dividend [4].

It is also important to note that our findings do not specifically support the scenario we offered earlier of an educator using the index method to progressively update their risk identification model. Rather, we took the preliminary step of establishing the viability of the index method as a predictor of student risk by comparing it to an established approach on a common data set. Our follow-up research will investigate risk analysis in niche settings, where academics are able to integrate context specific variables that are closely tied to their curriculum design and teaching practice. We hope then to be able to investigate the flexibility of the method and its potential for adoption and diffusion.

One objection to our approach needs to be canvassed here as it goes to the heart of our rationale. Our comparison, so this criticism goes, is a straw man argument, or about to become one, as big data and the associated data mining methodologies will replace regression anyway, or supplement it in a way that is a stepwise change from the process used here. We certainly concede that there are data mining alternatives designed to neutralise regression's weaknesses with large data sets consisting of hundreds of variables [e.g. 15]. And, to be sure, the mining of big data sets has facilitated important discoveries, from the humanities [25] to biology [26]. However, two caveats should be sounded. Firstly, data mining has not shown itself to be superior to other methods in direct comparisons of forecast accuracy [2]. While education may produce student risk data that is uniquely amenable to exploitation via these methods, this has yet to be established [cf. 6].

This leads to our second caveat: the importance of context to higher education. There is considerable hope in the learning analytics literature that big data can make a significant contribution to our knowledge of student failure or dropout (e.g. the well-known Predictive Analytics Research (PAR) Framework, that pools data from six US universities [21]). One consequence of these attempts to distil new truths from vats of big data is that the more data is pooled, the more context is lost. Context, of course, is information, so omitting it is, by default, losing information. Context is also essential to interpretability: it is the Rosetta Stone of analytics. The danger of a large decontextualized data set is that meaning will be lost and the end result only relevant to whatever context is held in common. PAR's interim finding, that there is a "significant correlation between disenrollment and the number of concurrent courses in which students were enrolled" [7 p. 5] may be an underwhelming testament to this possibility.

Of course big data and cross-institutional projects are potentially valuable, and their 'macro' findings may serve to inform 'micro' practices [7]. However, modest analytic approaches like the index method that are flexible and adaptable by the user may play an important role also. A modest analytics can be sensitive to context, and facilitates the working *with*, rather than *on*, educators [23], which is likely to be advantageous to the diffusion and embedding of student risk management practices.

5. ACKNOWLEDGMENTS

Our thanks to Damien Edwards and Louise Spencer for gathering and sifting the data from our systems (the University's, that is).

6. REFERENCES

[1] Armstrong, J.S. 2001. Combining forecasts. *Principles of forecasting*. J.S. Armstrong ed. Springer. 417–439.

[2] Armstrong, J.S. 2006. Findings from evidence-based forecasting: Methods for reducing forecast error. *International Journal of Forecasting*. 22, 3 (2006), 583–598.

[3] Armstrong, J.S. 2012. Illusions in regression analysis. *International Journal of Forecasting*. 28, 3 (Jul. 2012), 689–694.

[4] Armstrong, J.S. 1985. *Long-range forecasting: From crystal ball to computer*. John Wiley.

[5] Armstrong, J.S. and Graefe, A. 2011. Predicting elections from biographical information about candidates: A test of the index method. *Journal of Business Research*. 64, 7 (2011), 699–706.

[6] boyd, d. and Crawford, K. 2012. Critical Questions for Big Data. *Information, Communication & Society*. 15, 5 (2012), 662–679.

[7] Buckingham Shum, S. 2012. UNESCO Policy Brief: Learning analytics. UNESCO. Moscow. http://iite. unesco. org/pics/publications/en/files/3214711. pdf.

[8] Bunn, D. and Wright, G. 1991. Interaction of Judgemental and Statistical Forecasting Methods: Issues & Analysis. *Management Science*. 37, 5 (May 1991), 501–518.

[9] Campbell, J.P., DeBlois, P.B., and Oblinger, D.G. 2007. Academic analytics: A new tool for a new era. *Educause Review*. 42, 4 (2007), 40.

[10] Crow, S.M., Hartman, S.J., Mahesh, S., McLendon, C.L., Henson, S.W., and Jacques, P. 2008. Strategic Analyses in Nursing Schools: Attracting, Educating, and Graduating More Nursing Students: Part I-Strengths, Weaknesses, Opportunities, and Threats Analysis. *The Health Care Manager*. 27, 3 (2008), 234–244.

[11] Czerlinski, J., Gigerenzer, G., and Goldstein, D.G. 1999. How good are simple heuristics? *Simple heuristics that make us smart*. G. Gigerenzer and P.M. Todd, eds. Oxford University Press. 97–118.

[12] Dana, J. and Dawes, R.M. 2004. The Superiority of Simple Alternatives to Regression for Social Science Predictions. *Journal of Educational and Behavioral Statistics*. 29, 3 (Sep. 2004), 317–331.

[13] Dillenbourg, P., Nussbaum, M., Dimitriadis, Y., and Roschelle, J. 2013. Design for Classroom Orchestration. *Computers & Education*. (2013), 1–34.

[14] Dillenbourg, P., Zufferey, G., Alavi, H., Jermann, P., Do-Lenh, S., Bonnard, Q., Cuendet, S., and Kaplan, F. 2011. Classroom orchestration: The third circle of usability. *CSCL2011 Proceedings* (2011), 510–517.

[15] Ebrahimi, M., Ebrahimie, E., Shamabadi, N. and Ebrahimi, M. 2010. Are there any differences between features of proteins expressed in malignant and benign breast cancers? *Journal of Research in Medical Sciences : The Official Journal of Isfahan University of Medical Sciences*. 15, 6 (2010), 299–309.

[16] Einhorn, H.J. and Hogarth, R.M. 1975. Unit weighting schemes for decision making. *Organizational Behavior and Human Performance*. 13, 2 (Apr. 1975), 171–192.

[17] Ferguson, R. 2012. The state of learning analytics in 2012: A review and future challenges. *Knowledge Media Institute, Technical Report KMI-2012-01*. (2012).

[18] Fritz, J. 2011. Classroom walls that talk: Using online course activity data of successful students to raise self-awareness of underperforming peers. *The Internet and Higher Education*. 14, 2 (Mar. 2011), 89–97.

[19] Gigerenzer, G. 1996. From tools to theories: Discovery in cognitive psychology. *Historical dimensions of psychological discourse*. (1996), 36–59.

[20] Goldstein, D.G. and Gigerenzer, G. 2009. Fast and frugal forecasting. *International Journal of Forecasting*. 25, 4 (Oct. 2009), 760–772.

[21] Ice, P., Díaz, S., Swan, K., Burgess, M., Sharkey, M., Sherrill, J., Huston, D., and Okimoto, H. 2012. The PAR Framework Proof of Concept: Initial Findings from a Multi-Institutional Analysis of Federated Postsecondary Data. *Journal of Asynchronous Learning Networks*. 16, 3 (2012), 63–86.

[22] Johnson, L., Adams, S., and Cummins, M. 2012. *Technology Outlook for Australian Tertiary Education 2012-2017: An NMC Horizon Report Regional Analysis. Austin, Texas: The New Media Consortium*.

[23] Jones, D. Three likely paths for learning analytics and academics. *The Weblog of (a) David Jones*. http://davidtjones.wordpress.com/2012/10/11/three-likely-paths-for-learning-analytics-and-academic-in-oz-higher-education/.

[24] Kotter, J.P. 1995. Leading change: Why transformation efforts fail. *Harvard Business Review*. 73, 2 (1995), 59–67.

[25] Michel, J.B., Shen, Y.K., A. P. Aiden, Veres, A., Gray, M.K., The Google Books Team, Pickett, J.P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M.A., and Aiden, E.L. 2011. Quantitative analysis of culture using millions of digitized books. *Science*. 331, 6014 (2011), 176–182.

[26] Nealson, K.H. and Venter, J.C. 2007. Metagenomics and the global ocean survey: what's in it for us, and why should we care? *The ISME Journal*. 1, 3 (2007), 185–187.

[27] Norris, D. and Baer, L. 2013. *Building organizational capacity for analytics*. Educause. http://net.educause.edu/ir/library/pdf/PUB9012.pdf.

[28] Rogers, E.M. 2010. *Diffusion of innovations*. Simon and Schuster.

[29] Schmidt, F.L. 1971. The Relative Efficiency of Regression and Simple Unit Predictor Weights in Applied Differential Psychology. *Educational and Psychological Measurement*. 31, 3 (Oct. 1971), 699–714.