Accountability for Small Educator Preparation Programs

Prepared for the Education Professional Standards Board

by

Terry Hibpshman
University of Kentucky
Martin School of Public Policy and Administration

July 2016

“Little experience is sufficient to show that the traditional machinery of statistical processes is wholly unsuited to the needs of practical research. Not only does it take a cannon to shoot a sparrow, but it misses the sparrow! The elaborate mechanism built on the theory of infinitely large samples is not accurate enough for simple laboratory data. Only by systematically tackling small sample problems on their merits does it seem possible to apply accurate tests to practical data.” - - Fisher, R (1934).

In previous work (Hibpshman, 2013) we developed a model of accountability measurement for educator preparation programs (EPPs) approved by the Kentucky Education Professional Standards Board (EPSB). This model specified a general methodology for identifying deficits in program performance given the mission and goals of the EPSB, together with recommendations for specific measures of performance. The model was intended to be comprehensive: i.e., it provided for the evaluation of every performance issue that might be of interest to EPSB, either by recommending a specific method, or by providing for the development of methods for new issues that might arise in the future.

Although the 2013 model is comprehensive, it did not deal with one problem that has plagued the field of EPP accountability for some time: evaluation of small programs. Some programs, such as elementary education, enroll sufficient numbers by nearly all providers to allow for reasonably accurate estimates of program performance year to year; but many programs, even in larger institutions, enroll so few candidates that we cannot create estimates of program performance using traditional methods, even when data are combined across multiple years. Yet it is in our interest to evaluate these small programs as we would large programs, since public school children are taught by the teachers they produce, and we want to assure that, as with larger programs, smaller programs are of sufficient quality to prepare completers that promote adequate student achievement.

These small programs actually make up the majority of programs offered by Kentucky-approved educator preparation providers. They include such content areas as physics, earth science, and foreign languages – which are essential to the education enterprise but are generally not popular with K-12 students – and programs for teachers of low-incidence special education populations. Some common content areas, such as mathematics, English, or even Elementary Education have ample enrollments when offered by large providers, but may have insufficient enrollments to be evaluated with traditional methods when offered by small providers.

That small programs must exist is a consequence of two factors, curriculum requirements and teacher labor markets. Although the Commonwealth of Kentucky does not explicitly require some of the subjects for which educators are prepared by state-approved providers (such as physics, foreign languages, and engineering), these subjects are nonetheless highly desirable in a competitive national and state educational environment, and would be expected to be available to public school students, even if they are not popular. A substantial part of the curriculum of Kentucky K-12 schools consists of this type of subject matter.

Teacher labor markets matter because in Kentucky, teacher labor markets in most public school districts are dominated by a single provider.[1] In some districts the dominance of the largest supplier of teachers is so great that no other meaningful supplier can be said to exist.[2] Because this is true, unless there were some substantial change in markets, we can expect that, as a practical matter, if a district needs a teacher in a particular subject area, the teacher must come from the dominant provider. Providers that dominate districts must therefore train teachers in a very wide range of subject areas in order to fill job openings, and because some of these will be in subjects for which only a small number of placements exist in any given year, must necessarily operate small programs. To understand this issue, consider the case of physics. In earlier work we demonstrated that there are job openings for about 20 physics teachers statewide in any given year (Clements & Hibpshman 2008). But because these openings are likely to occur in widespread locations across the state, having a few large physics programs would guarantee that most of the jobs went unfilled, while guaranteeing that the majority of newly-trained physics

[1] By this we mean that a large plurality – often more than half – of the teachers employed by the district were trained by a single provider.

[2] E.g., districts in which 75% or more of the teachers, over a long period of time, are completers of a single provider.

teachers would be unemployed.[3] Although we would have no shortage of physics completers statewide, we would nonetheless effectively have a statewide shortage.

The present paper evaluates the nature and scope of the small-sample problem, and recommends some approaches to evaluating small programs. In section 1 we define what is meant by a "small program." In section 2 we evaluate the extent of the problem in Kentucky. In section 3 we suggest some approaches to evaluating the performance of these programs.

[3] Because research on teacher labor markets has shown consistently that completers do not often stray far from the geographic location of the program in which they are trained (Boyd et al., 2005).

Section 1
What is a small educator preparation program?

To answer this question, we begin with a consideration of the nature of our task in evaluating EPPs. What we want to know is whether an EPP is performing as well as, better than, or worse than our expectations. Our expectations might be defined by some target value (as when we desire completer Praxis pass rates to be greater than some value)[4], or they might be defined in terms of typical values for similar programs, or they might be based on a value negotiated with a particular provider. There will necessarily be some amount of error in our calculated estimates of program performance, either because of random fluctuations in the underlying causes of program performance, or because of data errors. Because some amount of error is unavoidable, we need some way to determine how confident we are that an apparent deviation from our expectations is in fact an indication that the program truly deviates from our expectations and is not due to chance[5].

Methods for making this type of judgement are well-established in the field of statistics. In general, statistical analyses establish the level of confidence that a result is truly different from some other value by use of the *standard error*, which is a measure of the likelihood that a statistic differs by some magnitude from the true value.[6] The standard error is inversely related to the size of the sample on which it is based, and the smaller the sample size, the greater the standard error. When a sample size is very small, the size of the standard error may be so great relative to the obtained value as to make it possible to detect only very large differences with any reasonable level of confidence.[7]

It is the difficulty of detecting small or moderate deviations from expected values that defines what is a small sample. A good rule of thumb used by many in statistics and the sciences is 30.[8] That is because when the sample size is at least thirty, the probabilities of the t distribution are functionally indistinguishable from those of the normal distribution. We note that this is really more a matter of convenience than of methodological purity: the t distribution by its nature corrects for distortions in the sampling distribution caused by small sample sizes, and meaningful comparisons can be made using t with very small samples – as few as 2 or 3, if conditions are right (Bland & Altman, 2009; De Winter, 2013).[9]

What defines a sample as small is a question of what methodology is appropriate for analysis of data about program performance. The frequently-mentioned minimum of 25 for studies of EPP performance is based, we believe, on the view that our usual practice – comparison of raw proportions – will generate a greater than desirable level of uncertainty if the sample size is very much less than that. There is perhaps some justice to this, but it is a problem only as long as simple comparison of unadjusted proportions is the predominant method.

Concerns over the difficulty of evaluating small programs has some unfortunate consequences. CAEP (CAEP, 2013), in its data requirements for state licensure programs, states that it will waive evaluation of programs with less than 10 enrollments in a three year period.[10] Application of these waivers would mean, in Kentucky, that more than half of our programs would receive waivers for every year for which data are available. This would mean that a substantial

[4] A “bright line.”
[5] i.e., is due to some factor within the control of the provider.
[6] More accurately, the standard error is the standard deviation of the sampling distribution of a statistic.
[7] The size of the standard error depends in part on the type of statistic used. In our case, because we typically use proportions to measure performance, the standard error tends to be especially large when the sample size is small.
[8] In some fields, a sample of a thousand or more may be viewed as small (Mitschele, 1991).
[9] Decisions about the meaningfulness of statistics based on samples of various sizes are the subject of very complex analyses regarding power and sampling methodology which are beyond the scope of this paper.
[10] For ease of exposition, we call this the “CAEP small sample rule.”

proportion of the teachers trained by these programs would not be subject to evaluation, and that programs that prepare teachers in some low-incidence subjects would never be evaluated at all. This would mean that we were managing the quality of only large, popular programs.

The small size of many of our programs is an example of a problem that exists in other fields, drawing reasonable inferences in circumstances where sample sizes are small for reasons organic to the object of study. Good examples of this type of problem outside of educator preparation are bone density studies of astronauts (Vico et al., 2000), and destructive materials studies (Garland & Greene, 2009). The total number of persons who have ever been in space is just a few hundred, and at any given time only a few can be assembled for a study. The cost of large-scale studies of materials toxicity can be substantial when the materials must be destroyed in the process of evaluation. The question of how the human body reacts to conditions in space is vital for the conduct of future space missions, and it would be irresponsible to forgo such studies because of small sample sizes. It would likewise be irresponsible for federal agencies to forego toxicity testing because of small sample sizes. We argue here that it would be equally irresponsible to forego studies of small EPPs.

Instead of applying naïve methods learned in our undergraduate statistics or research methodology courses, we have the option of applying less well-known but well-established methods designed for the types of problems that confront us. We will demonstrate two such methods in section 3, but it is important to be clear on the nature of what we are attempting to do. When we evaluate the performance of EPPs, we are not trying to conduct scientific studies to construct theories of teacher preparation.[11] Instead, we are trying to assemble evidence that will permit us to identify deviations from expected performance, so that we can intervene in order to assure improvements in teacher education. Formal criteria such as alpha levels and confidence bands, while they are necessary to our enterprise, are useful only insofar as they help us to make reasonable decisions about the performance of EPPs. There is nothing sacrosanct about a “confidence level” of 95% (Guyatt, Mills & Elbourne 2008; Lakins, 2014).

[11] Although these methods, applied over many studies of EPPs will, over time, lend themselves to theory construction Coppedge (2002).

Section 2
The extent of the small sample problem

To answer this question, we evaluated data from a research dataset constructed from the EPSB database in 2013. This dataset has reliable data for academic years close to 2012, but because there are known to be errors in earlier years of the database, we confined ourselves to the 6 most recent years, academic years 2007-2012. We applied a number of criteria to determine what proportion of teachers and programs would be deemed small, according to the following definitions:

- The proportion of programs that had less than 25 admissions for any year in the 6-year range
- The proportion of candidates admitted to all programs who would not be subject to evaluation, applying the minimum 25 admissions rule
- The proportion of programs that would have received waivers under the CAEP small sample rule for one or more years within the range

In addition, because it is often suggested that a solution to the problem is to combine data in three-year increments, we evaluated the proportion of programs and candidates who would not be evaluated if we did so. We also evaluated the question of whether, by combining programs into supernumerary classes ("program categories"), we could provide for evaluation of the majority of programs.

Results

If the CAEP small sample rule is applied to year-to-year data, we find that 64% of approved programs in Kentucky would receive waivers for one or more years in the 6-year range, and 56% would receive waivers for all years.

Only 2% of Kentucky programs would have 25 or more admissions for every year in range. These programs would constitute about a fourth of the teachers admitted in each of those years, i.e., about three fourths of the admitted candidates would not be subject to evaluation.

When data are combined in 3-year increments, 41% of programs would receive CAEP waivers for one or more years, and just 5% of programs would have 25 or more admissions for each of the three-year time periods.

Providers operate a large number of similar programs even within the same institution, and it is of interest to know whether combining data from similar programs might allow more programs and candidates to be evaluated. For this purpose we created "program category" classes by combining data within subject areas (e.g. combining data for all LBD programs, whether traditional or alternative route), and by combining programs that were within some subject superclass (e.g., combining all foreign languages into a single program category). Evaluating the admissions for these program categories, we found that 50% would be granted waivers under the CAEP small sample rule, and 35% would have waivers for the entire 6 year range. 64% of the program categories would have less than 25 admissions for at least one year of the range. About 35-40% of candidates would still not be subject to evaluation. Even when these program categories were compiled into three-year intervals, 21% would receive waivers under the CAEP small sample rule, and 11% would receive waivers for the entire range of available years.

Discussion

It is clear that the great majority of EPSB-approved programs are classifiable as small programs. Even when data are combined into program categories, and when data are combined into multiple year “buckets,” substantial numbers of programs and candidates would never be subject to evaluation, if the above-mentioned rules were applied. If we wish to comprehensively evaluate the performance of EPPs, we must apply strategies designed for these many small programs.

Section 3
Methods

We investigated a number of different strategies:

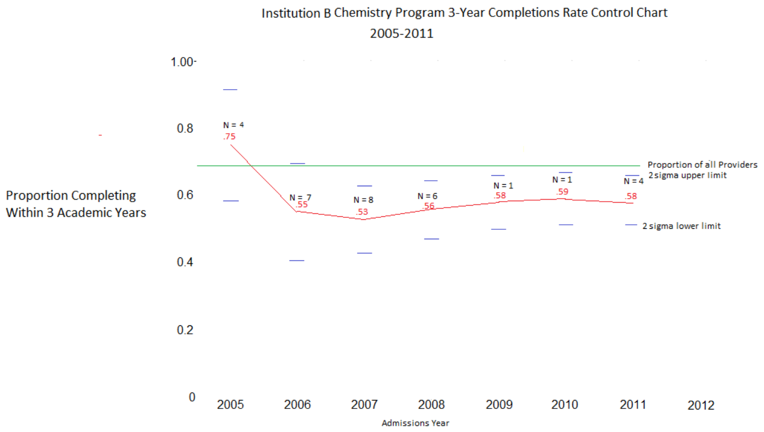
- Sequence analysis
- Binomial probability charting
- Bayesian approaches to statistical analysis, including:
    - Bayesian sequential analysis
    - Naïve Bayes analysis
- Survival analysis
- Exact methods designed for small samples, especially Fisher's exact test with Wilcox's correction
- Game theory

All of these methods are useful, but some of them either require difficult computations (Bayesian sequential analysis, exact methods, game theory), or were not easily interpretable by a person with minimal mathematical training (survival analysis, Naïve Bayes, Bayesian sequential analysis, game theory), or required the use of software packages not easily adaptable to our methodology (Bayesian analysis, survival analysis, game theory). Two, however – sequence analysis and binomial probability charting – proved optimal because of their simplicity, ease of computation, and easy interpretability.

Sequence analysis has a long history, beginning in 1929 with the seminal paper by Dodge and Romig (1929) at the Bell Systems Laboratory. The method works by accumulating data over time, and computing confidence intervals for the data accumulated at each time interval. The usual standard is a "3 sigma" (i.e., three standard error) boundary around the obtained value, and any value that falls outside the boundary is deemed to be remarkable.[12]

To give a concrete example, consider the following two charts drawn from our research data set:

Chart 1
Chemistry program completions for "Provider B"
2005– 2011



[12] Here we use a two sigma boundary because of the nature of our estimation problem, and to conform to the standard we use in our later, binomial, charts.

This is a very small program – only 31 candidates admitted over a seven year period, averaging 4.4 admission per year. The chart identifies this program as underperforming relative to other similar programs in just three years, with performance always below expectations after the first year. We would be justified in questioning whether this program requires some intervention to improve its completion rate. Further study would of course be necessary: if for example this is a program that typically admits a greater-than-average proportion of part-time candidates, the intervention would be different than if all candidates were full-time. If we were to posit a “bright line” standard for program completion, then this program might be subject to sanctions.

Consider chart 2, which displays the results for a different provider for the same subject, chemistry. This program admits an average of 4.6 candidates per year. The chart identifies it as overperforming in 4 years with just 11 admissions, and stays above the statewide level of performance thereafter. We would be justified in judging this program as exceptionally high-performing, and might want to refer the staff of Program B to this program for ideas about how to improve their completion rate.

Chart 2

Chemistry program completions for “Provider F”
2005– 2012

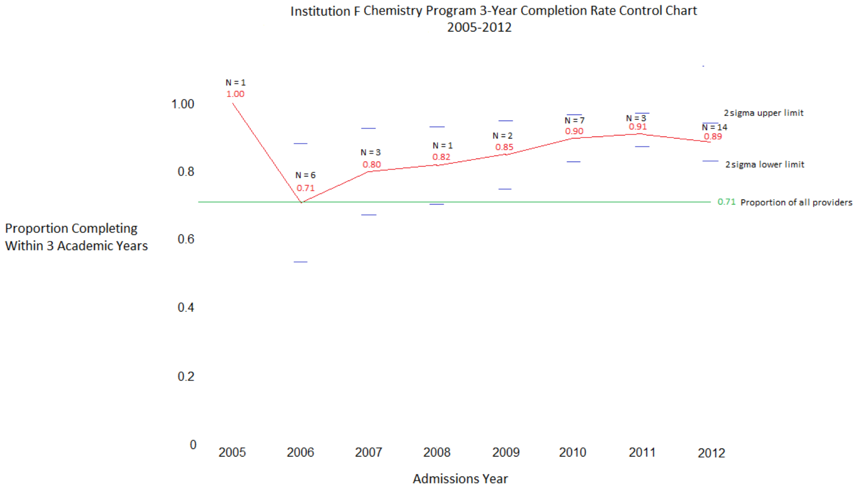
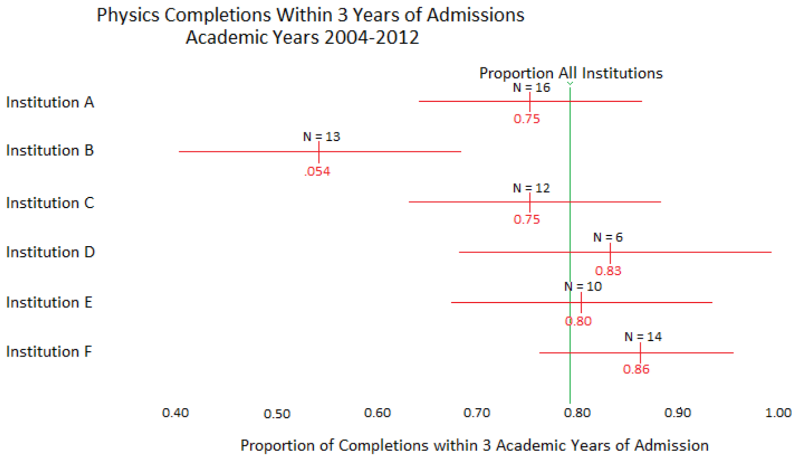


Chart 3 illustrates the use of what we here call a “binominal” chart.

Chart 3
Physics program completions within three years of admission
2004– 2012



All of these programs are very small – the largest admitted less than 2 candidates on average each year between 2004 and 2012. We assume that there is no bright-line standard for program completions, and use the statewide completion proportion as a rough measure of how well each program is performing relative to other programs in the same subject area. This chart functions by plotting the proportion of completions with a 2 standard-error confidence boundary.

Based on this chart we can say that just one of the six programs is remarkable. We are reasonably confident that "Institution B" is not performing as well as other similar Kentucky-approved physics programs. We can make this decision based on a total of 13 cases over 9 years – an average of 1.4 admissions per year.

Note that this chart loses some information that was available on the sequence chart, because it combines data for six years into a single measure. As a result, provider F, on this chart, no longer appears to be doing exceptionally well, as it did on the sequence chart. On the other hand, it does a fairly good job of identifying programs that markedly deviate from expectations, and this is what we want.

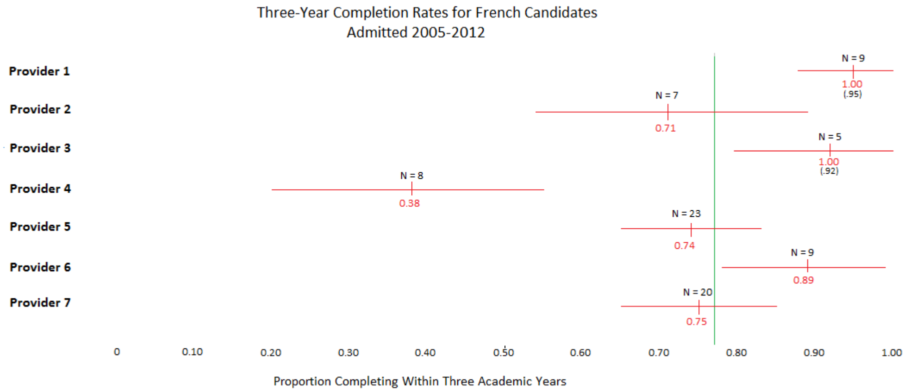
Chart 4 demonstrates the solution to one problem likely to occur when sample sizes are very small – creating a standard error when all of the cases are either successes or failures.[13] Although the maximum likelihood estimator of p (the proportion of successes) is just the number of successes divided by the number of observations (Chew, 1971), a finding that a program can be expected – based on historical records alone – to always be successful has little credibility (Lewis & Sauro, 2006). What we want is a measure of how confident we are that the program under study either exceeds or fails to achieve our expectations, and this requires the computation of a standard error. If all of the cases are successes or failures, the standard error will be 0.

Chart 4 corrects for this circumstance by use of the *adjusted Wald estimator*, which applies a correction developed specifically to deal with this problem (Lewis & Sauro, 2006).

[13] i.e., here we add an additional correction not incorporated into Chart 3.

Chart 4

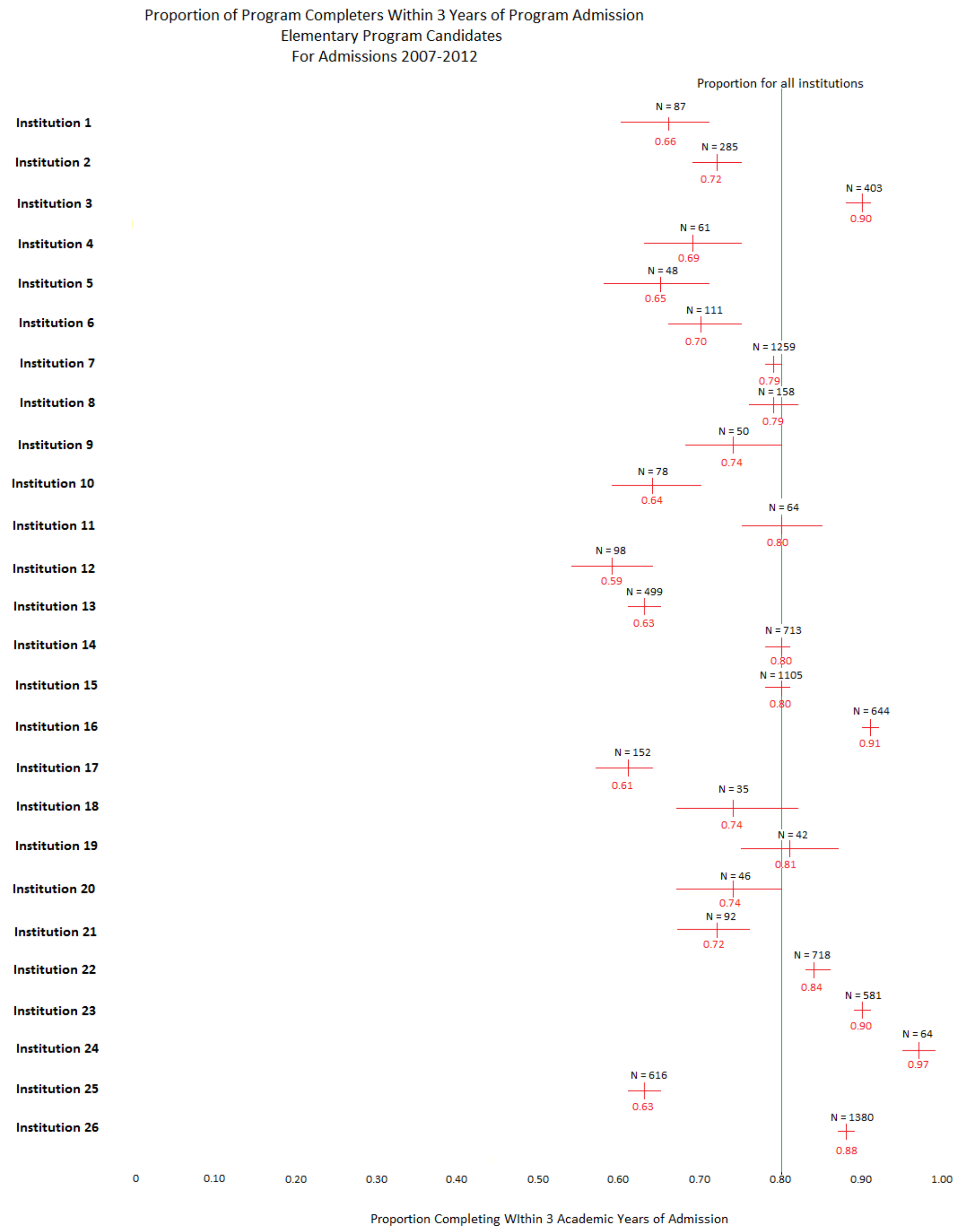
French program completions within three years of admission
2005– 2012



These are French program completions for seven programs that admitted candidates from 2005-2012. For two programs – Provider 1 and Provider 3 – all candidates completed within 3 years. The adjusted Wald estimator applied to the estimates for these two programs allows us to establish a confidence band, even though the unadjusted standard error would be 0. It is of interest to note that we can be reasonably confident that these two programs are performing better than the statewide expectation, even though the sample sizes are quite small – in one case, only 5 admissions. We are reasonably confident that the program offered by Provider 4 is underperforming.

These methods are not limited to use with small samples, and that solves a conceptual problem. One can argue that using these methods with small samples, and more traditional methods with larger samples, amounts to subjecting different programs to different standards. Although this isn't really true, we want to avoid the appearance of bias whenever possible. To illustrate the use of these methods with both large and small samples, we present Chart 5.

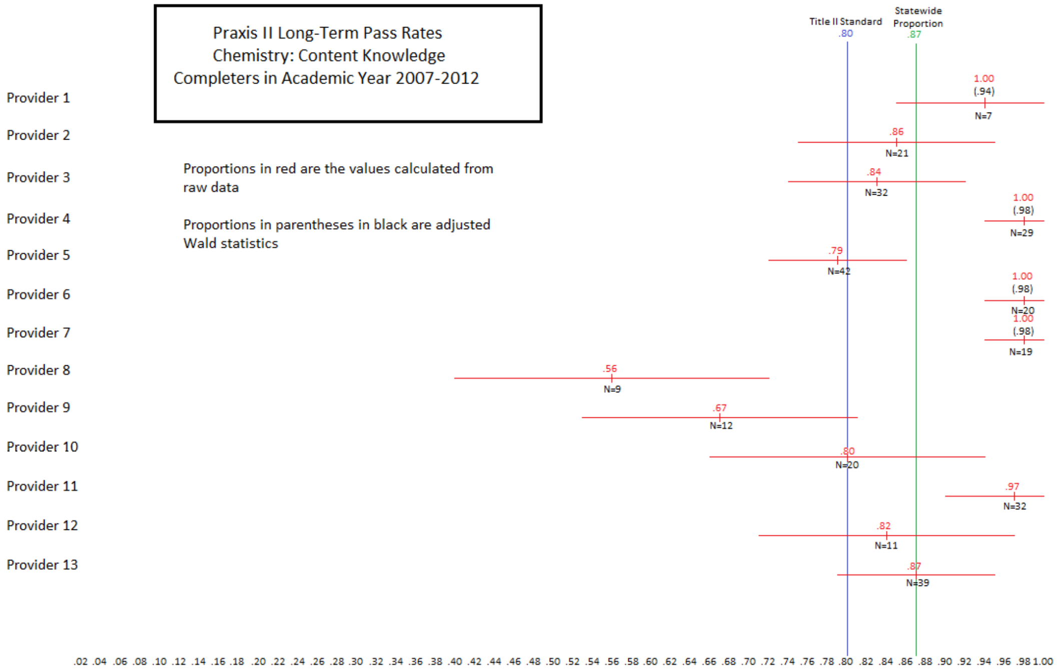
Chart 5
Elementary program completions within three years of admission
2007– 2012



Some of these programs are small – Institution 19, for example, admits an average of just 7 candidates per year. Others are very large, admitting 200 or more candidates per year.

This final chart illustrates the use of this methodology with a problem where a “bright line” standard applies.

Chart 6
Praxis 3-year Pass Rate
Chemistry: Content Knowledge
Candidates Completing in Academic Years 2007– 2012



The blue vertical line assumes we would have a fixed standard of 80% for acceptable performance. The green vertical line is the statewide pass rate computed from the data. Three programs have rates below the fixed rate – Providers 5, 8, and 9. Although all three have failed to meet our expectations, we would probably make somewhat different decisions based on this chart. We are not very confident that the pass rate for programs 5 and 9 is actually below 80%, and in these cases we would want to do additional analysis – perhaps prepare sequence charts. We are reasonably confident, however, that Provider 8 is failing on this standard, and would be justified in applying some remedy, based on these data alone.

Note one additional feature illustrated by this chart: the value of the adjusted Wald estimator. Although all of the candidates from Provider 1 passed the Praxis II test, we cannot confidently say, based on the adjusted Wald estimator, that this program is really doing better than other programs across the state.

The Way Forward

Based on the above analysis, we believe we have a workable method for performing accountability measurement for small programs. This method applies the following principles:

- We apply the binomial chart for all programs of the same type, and perform analyses with other methods when it is appropriate to do so

- We use a 6-year horizon, guaranteeing data from two completely disjoint data cycles[14]
- We chart all programs with at least 5 observations for each 6-year horizon

Using this approach, 2/3 of our programs can now be evaluated using a common, easily interpretable methodology. This leaves a third of our programs that are still too small to be subject to analysis. In principle, we can design methods for the evaluation of any program with 2 or more admissions, but it is unlikely that we would find deviations from expectations for such tiny programs except in cases where the deviations are very large. A better strategy, we believe, is to question whether these extremely small programs are viable. In some cases – such as preparation programs for teachers of low-incidence special education populations – it may be necessary to continue these programs. But in many cases, these programs are too small to be viable, and should be discontinued. When we identify these programs in the course of our analysis, we should query the providers as to their viability, and recommended that they be discontinued if a strong case cannot be made for their continued existence. When the decision has been made to continue a program despite its admission of less than 5 candidates in 6 years, we should then use surveys and other single-subject methods to acquire data about program performance.

[14] A data cycle is defined as 3 academic years, because our standard for program completion and Praxis II pass rates is within 3 years

References

Agresti A (1992). A Survey of Exact Inference for Contingency Tables. Statistical Science, Vol. 7, No. 1 (Feb., 1992), pp. 131-153.

Bachetti P (2010). Current sample size conventions: Flaws, harms, and alternatives. BMC Med. 2010; 8: 17.

Bacchetti P, Deeks S & McCune J (2011). Breaking Free of Sample Size Dogma to Perform Innovative Translational Research. Sci Transl Med. 2011 Jun 15; 3(87): 87ps24.

Berger, J. (1980). Statistical Decision Theory and Bayesian Analysis. New York: Springer-Verlag.

Bland M & Altman D (2009). Analysis of continuous data from small samples. BMJ 2009, 338: 3166.

Box J (1987). Guinness, Gosset, Fisher, and Small Samples. Statistical Science, 1987, 2(1), 45-52.

Boyd D, Lankford H, Loeb S & Wyckoff J. (2005). The Draw of Home: How Teachers' Preferences for Proximity Disadvantage Urban Schools. Journal of Policy Analysis and Management 24(1): 113-32, 2005.

Brown L, Cai T & DasGupta A (2001). Interval Estimation for a Binomial Proportion. Statistical Science 2001, 16(2), 101–133

Cai J, Zhou H & Davis C (1997). Estimating the Mean Hazard Ratio Parameters for Clustered Survival Data with Random Clusters. Statistics in Medicine, 16, 2009-2020 (1997)

Carlin B, Kadane J & Gelfand A (1998). Approaches for optimal sequential decision analysis in clinical trials. Biometrics; Sep 1998; 54, 3; 964-975

Casella G (1985). An Introduction to Empirical Bayes Data Analysis. The American Statistician, 39(2): 83-87.

Chew V (1971). Point Estimation of the Parameter of the Binomial Distribution. The American Statistician, Vol. 25, No. 5 (Dec., 1971), pp. 47-50.

Clements, S & Hibpshman T (2008). “Curriculum Complexity” in the American High School: Choosing Courses and Choosing Futures? A Report on a Limited Investigation and Proposal for a Larger Inquiry. Lexington, Kentucky: Partnership for Successful Schools.

Coppedge M (2002). Theory Building and Hypothesis Testing: Large- vs. Small-N Research on Democratization. Paper prepared for presentation at the Annual Meeting of the Midwest Political Science Association, Chicago, Illinois, April 25-27, 2002.

Council for the Accreditation of Educator Preparation (CAEP) (2013). Data Requirements: State Licensure Test Data. Retrieved from http://caepnet.org/accreditation/caep-accreditation/program-review-options/data-requirements on March 1, 2016.

De Vos A (2000). A primer in Bayesian Inference.

De Winter J (2013). Using the Student's t-test with extremely small sample sizes. *Practical Assessment, Research & Evaluation, Vol 18, No 10.*

Dodge H & Romig H (1929). A Method of Sampling Inspection. The Bell System Technical Journal, 1929 October: 613-631

Evans C, & Ildstad S (2001). Small Clinical Trials: Issues and Challenges. National Academies Press. Downloaded from http://www.nap.edu/catalog/10078.html

Fayers P, Ashby D & Parmar M (1997). Tutorial in Biostatistics: Bayesian Data Monitoring in Clinical Trials. Statistics in Medicine, 16, 1413-1430 (1997)

Fisher, R (1925). Statistical Methods for Research Workers, fifth edition. Edinburg: Oliver and Boyd, 1934. Downloaded from http://www.haghish.com/resources/materials/Statistical_Methods_for_Research_Workers.pdf

Garland S & Greene M (2009). Statistical Analysis of the Chemical Screening of a Small Sample of Unused Chinese and non-Chinese Drywall. U.S. Consumer Product Safety Commission, Division of Hazard Analysis.

Gehan E (1965). A Generalized Two-Sample Wilcoxon Test for Doubly Censored Data. Biometrika, 52(3/4), (Dec., 1965), pp. 650-653

Gregory (2010). Introduction to Bayesian Data Analysis. University of British Columbia. PowerPoint presentation.

Guyatt G, Mills E & Elbourne D (2008). In the Era of Systematic Reviews, Does the Size of an Individual Trial Still Matter? PLoS Med. 2008 Jan; 5(1): e4. Published online 2008 Jan 3.

Hibpshman, T (2013). Design of an Education Professional Standards Board (EPSB) Preparation and Accountability System for Teacher Training Programs. Frankfort, Kentucky: Education Professional Standards Board.

Hyde J (1977). Testing Survival Under Right Censoring and Left Truncation. Biometrika, 1977, 64(2), 225-230

Institute of Medicine (2001). Small Clinical Trials: Issues and Challenges. National Academy of Sciences, 2001.

IT Environmental Programs, Inc. & ICF Kaiser Incorporated (1994). Guidelines for Statistical Analysis of Occupational Exposure Data. Office of Pollution Prevention and Toxics, U.S. Environmental Protection Agency

Kammerman L (2014). Key Elements of Statistical Analyses for Studies with Small Populations (Pediatrics). Astrozenica PowerPoint presentation.

Kentucky Department of Education (KDE) (2016, May). Minimum High School Graduation Requirements. Frankfort, Kentucky, web page. Accessed at http://education.ky.gov/curriculum/hsgradreq/Pages/default.aspx.

Lai T (2001). Sequential Analysis: Some Classical Problems and New Challenges. Statistica Sinica 11(2001), 303-408.

Lakens D (2014). Performing High-Powered Studies Efficiently With Sequential Analyses. *European Journal of Social Psychology.*

Lan K, Rosenberger W & Lachin J (1995). Monitoring of Survival Data with the Wilcoxon Statistic. Biometrics, 51(3), (Sep., 1995), pp. 1175-1183

Lewis J & Sauro J (2006). When 100% Really Isn't 100%: Improving the Accuracy of Small-Sample Estimates of Completion Rates. Journal of Usability studies, 3(1), 136-150.

Lin D, Wei L & DeMets D (1991). Exact Statistical Inference for Group Sequential Trials.

Microsoft Research (2016). Fisher's Exact Test Tool Details. Accessed April 4, 2016 at http://research.microsoft.com/en-us/um/redmond/projects/MSCompBio/FisherExactTest/details.aspx

Mitschele J (1991). Small Sample Statistics. Journal of Chemical Education, Volume 68 Number 6 June 1991, 470-473

Mulholland H (1977). On the null distribution for samples of size at most 25, with tables. Biometrika (1977), 64(2), 401-9

Nakata T & Tonetti C (2014). Small Sample Properties of Bayesian Estimators of Labor Income Processes.

Nguyen X (2006). Sequential analysis: balancing the tradeoff between detection accuracy and detection delay. PowerPoint presentation

Novick M & Grizzle J (1965). A Bayesian Analysis to the Analysis of Data from Clinical Trials. Journal of the American Statistical Association, 1965, 60(309), 81-96.

O'Hagen A & Luce B (2003). a primer on Bayesian Statistics in Health Economics and Outcomes Research. 2003, Centre for Bayesian Statistics in Health Economics.

Olshausen B (2004). Bayesian probability theory.

Slavin R & Smith D (2008). Effects of Sample Size on Effect Size in Systematic Reviews in Education. Paper presented at the annual meetings of the Society for Research on Effective Education, Crystal City, Virginia, March 3-4, 2008.

Smith G, Seaman S, Wood A, Royston P & White I (2014). Correcting for Optimistic Prediction in Small Data Sets. Am J Epidemiol. 2014;180(3):318-324

StataCorp LP (2015). STATA Bayesian Analysis Reference Manual Release 14. College Station, Texas: STATA Corporation.

Sun D & Berger J (2008). Objective Bayesian analysis under sequential experimentation. IMS Collections Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh 3, (2008) 19-32

Tominaga D (2014). Statistical stage transition detection method for small sample gene expression time series data. Mathematical Biosciences, 2014, 254, 58-63

Tamura Y, Takehara H & Yamada S (2011). Component-Oriented Reliability Analysis Based on Hierarchical Bayesian Model for an Open Source Software. American Journal of Operations Research, 2011, 1, 25-32

Vico L, Collet P, Guignandon A, Lafage-Proust M, Thierry T Rehalia M & Alexandre C (2000). Effects of long-term microgravity exposure on cancellous and cortical weight-bearing bones of cosmonauts. The Lancet; May 6, 2000; 355, 1607.

Volinsky C & Raftery A (2000). Bayesian Information Criterion for Censored Survival Models. Biometrics, 56(1), (Mar., 2000), pp. 256-262.

Wetherill G (1961). Bayesian Sequential Analysis. Biometrika, 48(3/4), (Dec., 1961), pp. 281-292.

Woodall, W. (2000). Controversies and Contradictions in Statistical Process Control. 44th Annual Fall Technical Conference of the Chemical and Process Industries Division and Statistics Division of the American Society for Quality and the Section on Physical & Engineering Sciences of the American Statistical Association in Minneapolis, Minnesota, October 12–13, 2000.

Zar J (1987). A fast and efficient algorithm for the Fisher exact test. Behavior Research Methods, Instruments, & Computers 1987, 19 (4), 4/3-4/4.

Zaslavsky A (2001). Statistical Issues in Reporting Quality Data: Small Samples and Casemix Variation. International Journal for Quality in Healthcare, 13(6), 481-488.