Training and Education in Professional Psychology

Hidden Gems Among Clinical Psychology Training Programs

Jennifer L. Callahan, Camilo J. Ruggero, and Mike C. Parent

Online First Publication, October 14, 2013. doi: 10.1037/a0034233

CITATION

Callahan, J. L., Ruggero, C. J., & Parent, M. C. (2013, October 14). Hidden Gems Among Clinical Psychology Training Programs. *Training and Education in Professional Psychology*. Advance online publication. doi: 10.1037/a0034233

© 2013 American Psychological Association
1931-3918/13/$12.00 DOI: 10.1037/a0034233

Hidden Gems Among Clinical Psychology Training Programs

Jennifer L. Callahan and Camilo J. Ruggero
University of North Texas

Mike C. Parent
University of Florida

As a result of CoA-mandated program disclosure being initiated in 2006, there are now sufficient data available to allow for analyses that compare clinical psychology programs on a range of variables, including student outcomes. This standardized data, in concert with other sources of publically available data (i.e., APPIC and ASPPB), allow for programs to be compared empirically in new ways. Using SEM, in this Study 80.6% of the variance in clinical psychology training programs' outcomes (i.e., internship match and licensure exam performances) was accounted for by predoctoral characteristics (measured by GPA and GRE scores). Analyses then identified programs that produced exceptionally better outcomes than expected, given their predoctoral characteristics. The identified top programs were next compared on a range of department level training-relevant variables to similar programs, but whose outcomes were equal to or worse than expected. Findings are discussed and future directions for research and policy are suggested.

Keywords: training, graduate education, clinical psychology, internship match, EPPP, accreditation, program evaluation

Over the past decade, there has been an increasing expectation that clinical psychology training programs engage in "truth in advertising" (Belar, 2000) by disclosing objective data about programs costs and outcomes in publically accessible formats. Initially, it was the Council of University Directors of Clinical Psychology (CUDCP) that responded to this call by passing a motion stating "the member programs of CUDCP support the provision of fuller disclosure about operations and outcomes of our educational endeavors" (*http://www.cudcp.us*). The Commission on Accreditation (CoA) subsequently adopted Implementing Regulation (IR) C-20 in May of 2006, requiring all accredited doctoral programs to provide detailed, up-to-date information on "student admissions, outcomes, and other data" (*http://www.apa.org/ed/accreditation/about/policies/implementing-regulations.aspx?item=27*).

Before these changes, few data were publically accessible to engage in any kind of objective evaluation of training programs. Because of the lack of empirical data, publicized rankings of program success were primarily reputational in nature. Perhaps the most widely recognized among these rankings is found within the popular press, via the *U.S. News & World Report* (*USNWR*) rankings of colleges and graduate programs. According the *USNWR's* reported methodology (Morse, 2012), psychology programs were identified by the American Psychological Association. *USNWR* decided to include each department only once, regardless of how many distinct degree programs were offered. Thus, the rankings usually reflect the department, rather than a specific program. Who was surveyed was not uniform across schools. The methodology description suggests that it was often the department chair, though it was also described as including a graduate studies director or even simply a faculty member teaching graduate students. It is unknown whether the previously adopted CUDCP resolution (January 2001; available at *http://www.cudcp.us*) that their nearly 200 member programs not participate in *USNWR* reputational surveying, increased the likelihood that only someone with peripheral knowledge about those programs would respond to the survey request. What is known is that the response rate for the discipline of psychology was only 25% (compare this with the 90% completion rate in the area of criminology, e.g.). Those who did respond ranked up to 10 programs that they thought represented the best programs within each specialty, regardless of their familiarity with the specialty itself. In essence, the rankings reflect reputational strength of departments with ratings often supplied by less-informed individuals rather than stemming from genuine peer review. Despite such methodological flaws, reputational rankings continue to be relied upon heavily by prospective students as well as in determination of awards and grant distributions (Hanish et al.,

Jennifer L. Callahan earned her PhD in clinical psychology from the University of Wisconsin-Milwaukee and completed her internship and postdoctoral training at Yale University. She is currently an associate professor and a director of clinical training at the University of North Texas. She holds board certification in clinical psychology and studies training and competency development.

Camilo J. Ruggero, an assistant professor at the University of North Texas, studies bipolar and depressive disorders, and has expertise in advanced quantitative methods for social sciences. He received his PhD in clinical psychology from the University of Miami and completed an internship and NIMH-sponsored research fellowship at Brown University.

Mike C. Parent is presently a doctoral candidate in counseling psychology at the University of Florida, and will be beginning a full-time position as an assistant professor in counseling psychology at Texas Tech University in August 2013. His research focuses on intersections of gender, sexuality, and behavioral health.

Correspondence concerning this article should be addressed to Jennifer L. Callahan, PhD, ABPP, University of North Texas, Department of Psychology, 1155 Union Circle #311280, Denton, TX 76203. E-mail: jennifer.callahan@unt.edu

1995; Ilardi et al., 2000), perhaps owing to their extensive and long history (for a good overview, see Diamond & Graham, 2000).

In contrast to reputational rankings, it was suggested by Kazdin (2000) that the rigor of research evaluation be applied to the evaluation of training programs. Initially, only limited data were available for such evaluations. These early investigations drew from publically available data in program brochures/Web sites, hand searches of journals, or from databases of published literature and generally approached program evaluation by examining faculty data. As a result, faculty research productivity and/or impact has been a common area of inquiry accomplished by examining number of publications and citations within specific journals or fields of study (see Morey, 2010 for a relatively recent example of a well-conducted study in this genre). Because of the limited outcome data that have historically been publically available, it has been more difficult to assess outcomes and make comparisons between programs. Those comparisons that have been made typically draw again from faculty-level data. For example, Ilardi, Rodriguez-Hanley, Roberts, and Seigel (2000) identified programs contributing the most core faculty members to training programs. Although important data, most graduates do not go on to faculty careers. The ability to conduct a comparative study examining outcomes as pertaining to more commonly attained professional benchmarks has been historically elusive.

With the advent of CoA-mandated program disclosure in 2006, there are now sufficient data available to allow for analyses comparing programs on range of variables, including program costs and outcomes (e.g., internship match rates). Such normative data allow for individual programs to evaluate how they are succeeding, while also enabling potential students to make informed decisions throughout the application and decision-making processes. Such data may be used to advocate for program level changes (e.g., negotiating internally with upper administration to provide greater financial support to incoming students that is commensurate with peer programs), to determine the impact of changes that have already been made (e.g., examining whether student qualifications of incoming students have improved as a result of some change in recruitment or selection procedures), or to identify needed areas of program development (e.g., addressing poor match rates).

There has been some discussion of using such data to identify underperforming training programs, most notably via the identification of thresholds for internship placement (CoA IR D.4–7[b], currently under revision). However, prospective students likely want to have more information than simply which programs to screen out of consideration. In the authors' mentoring experiences, undergraduate students more typically ask for assistance in identifying graduate programs they should screen into their consideration and/or suggestions on making decisions between competing offers for admission. Although many of the variables that inform such decisions are idiosyncratic (e.g., familial proximity, area of research interest), the publically available data disclosed via programs and other bodies provide a uniform data repository upon which useful comparisons might be made. By examining such data it becomes possible to not only set thresholds to identify underperforming programs it also provides the means to identify programs that are performing as expected or, perhaps, better.

With respect to evaluating training programs, different approaches are possible using publicly available data. The most straightforward involves evaluating programs based on unadjusted outcomes (e.g., a program's success of placing students into internships [Parent & Williams, 2010) or pass rates on the EPPP [Schaffer et al., 2012]). Such evaluations are important, easy to interpret, and are often of most direct interest to students and other stakeholders. However, they are less useful for evaluating program efficacy, per se: programs that take highly qualified students tend to have better outcomes (see review by Stedman, 2007). Such confounds make it difficult to know whether successful outcomes are a reflection of the program's training, a reflection of the predoctoral characteristics of their students, or some combination of these variables. Finding adequate models to rank and evaluate different programs remains contentious, a problem seen in other areas of education and medicine (e.g., Burgess et al., 2000; Rubin et al., 2004). Approaches that recognize and adjust for baseline differences in predoctoral characteristics are better able to identify the contributions of program training to outcomes than those that do not (cf. Goldstein & Spiegelhalter, 1996).

The purpose of the present study was to use publically disclosed data sources to identify clinical psychology programs that excel empirically at training, as reflected by two emerging professional benchmarks (i.e., success of programs' students in obtaining an internship placement and their success at passing the EPPP). The latter two benchmarks are hardly exhaustive: they ignore other important signals of success, such as research productivity. Nevertheless, they are standardized, making comparisons possible, and are a better basis of evaluation than having no data at all, which was often the case in the past.

Rather than simply use outcomes in isolation to rank programs, programs were ranked based on the degree to which their outcomes *exceeded* prediction given the predoctoral characteristics of each program's student body, as reflected by the typical Graduate Record Examination (GRE) and undergraduate grade point average (GPA) of the programs' incoming students. Both GRE and undergraduate GPA have repeatedly been found to be significant predictors of graduate school success (cf., meta-analytic findings reported by Kuncel, Hezlett, & Ones, 2001). Thus, outcomes were ranked only after adjusting for differences in the programs' typical incoming GRE and undergraduate GPA scores. Finally, exploratory descriptive analysis focused on programs that performed exceptionally better than predicted by comparing them on a range of variables to similar programs but whose performance with respect to outcomes was equal to, or worse than, expected.

Method

Participants

For the sake of interpretive clarity, the sample was drawn from the 233 accredited Clinical Psychology doctoral programs and excluded other accredited doctoral programs in psychology (e.g., Counseling Psychology, $n = 69$; School Psychology, $n = 59$; Combined, $n = 8$). Other exclusion criteria were as follows: program was not located within the United States, program located in a US territory, no data on the primary outcome measures (see below) for the years 2006 to 2010. One hundred eighty-three programs remained after enforcing the exclusion criteria. Among these programs, 74.3% were Ph.D. programs ($n = 136$) and 24.0% were Psy.D. programs ($n = 44$).

Measures

Predoctoral characteristics. GRE verbal and quantitative scores, as well as undergraduate GPA for each year from 2006 through 2010 were accessed via public disclosure data associated with each of the identified accredited programs. Because of space limitations, citations for each public disclosure Web page are not provided herein. However, in general, a basic Internet search for the name of the school (e.g., "university of . . .") in concert with the program type (e.g., "clinical psychology") was typically all that was needed to locate public disclosure data, which were often (though not always) labeled "student admissions, outcomes, and other data."

Emerging professional benchmarks. Two indicators of emerging professional benchmarks were used to evaluate training outcomes. The first involved internship match rates for each doctoral program as reported by the Association of Psychology Postdoctoral and Internship Centers (APPIC) for each year from 2006 through 2010, accessed via the APPIC Web site (APPIC, 2010). The second indicator involved the percentage of the program's examinees passing the Examination for the Professional Practice of Psychology (EPPP), as reported by the Association of State and Provincial Psychology Boards (ASPPB), for aggregate years 2006–2011 on the ASPPB Web site (ASPPB, 2011). The EPPP data were aggregated by ASPPB to include year 2011, and therefore could not be constrained to 2010 for the current study, but all other variables in the present study were restricted to end with year 2010 to coincide with significant changes in the APPIC match process beginning in year 2011 (i.e., elimination of the Clearinghouse and initiation of a two phase match system).

Department characteristics. Data tables for the National Research Council's *A Data-Based Assessment of Research-Doctorate Programs in the United States* (revised 5/3/2011) were accessed (data and methodologies may be accessed from *http://www.nap.edu/rdp*) to characterize faculty, student, and training variables associated with departments' housing select training programs.

Data Analyses

Analyses focused on identifying programs achieving better emerging professional benchmarks than expected given the program's predoctoral characteristics. Because data are not available linking specific incoming students with their particular emerging professional benchmarks, all analyses were conducted at the program level using aggregated data to approximate programs' typical incoming predoctoral characteristics on the one hand and programs' typical outcomes on emerging professional benchmarks on the other.

A structural equation model (SEM) was estimated, in which the latent variable of emerging professional benchmark outcomes was predicted by the latent variable of predoctoral characteristics (see Figure 1). Indicators for the latent predoctoral characteristics predictor variable were average GRE-Verbal scores, average GRE-Quantitative score, and average undergraduate GPA of each program's incoming class (averages were for years 2006–2010). Indicators for the latent emerging professional outcomes variable were EPPP pass rates and average internship match rates (averages were again for years 2006–2010). SEM analyses were conducted using Mplus (Muthen & Muthen, 2007) and involved full information maximum likelihood, a recommended approach for handling missing data (Enders, 2001; Schafer & Graham, 2002). By using this approach, no programs were excluded from SEM despite missing data on some of the variables. Most commonly, missing data were found among the variables for GPA or GRE scores (however, two thirds of the sample was fully complete across all three of these variables).

Parameters from the estimated SEM model were used to predict emerging professional benchmark outcomes, and then comparisons were made between each program's predicted outcomes and their actual outcomes (i.e., the residuals were calculated). Large positive discrepancies (i.e., residuals) reflect the fact that a program is producing better emerging professional benchmark outcomes than predicted given predoctoral characteristics. These discrepancies were used to rank programs. In other words, rankings reflect programs' outcomes controlling for differences in predoctoral characteristics, providing a better measure of the relative importance of program training. The top 10 programs based on this adjusted ranking are reported for the combined latent outcome, as well as for each specific outcome (i.e., EPPP passing and internship match rate).

Subsequent considerations of the data focused on identifying factors related to better than expected outcomes. Because this

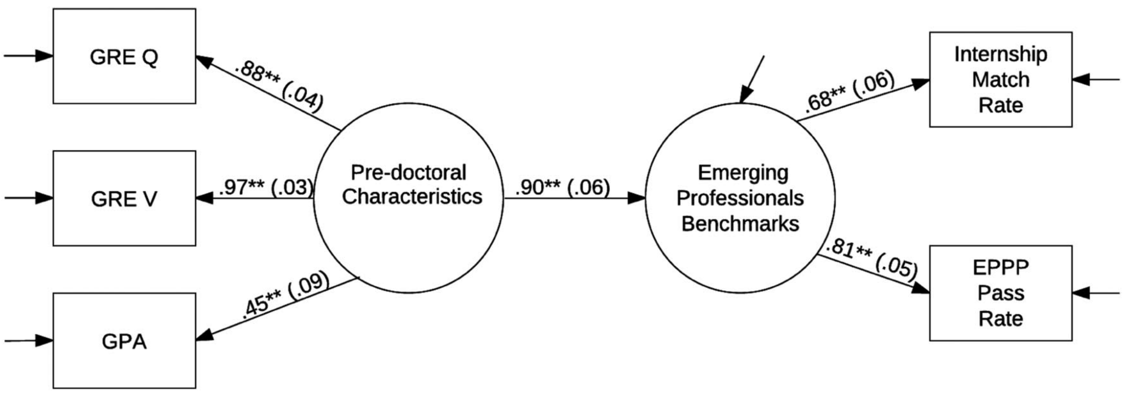

Figure 1. SEM of clinical psychology program students' success with emerging professional benchmarks predicted by predoctoral characteristics. Coefficients (standard errors) are standardized. R^2 for "Emerging Professional Benchmarks" = .81, *SE* = .11, $p < .001$. $^{*} p < .05$. $^{**} p < .01$.

represents the first study of its kind, a conservative approach to the data was taken and the focus was constrained to only those programs that empirically demonstrated clearly exceptionally better than expected outcomes, operationalized as evidencing outcomes that were more than two standard deviations better than predicted. Those programs were compared with programs with similar predoctoral characteristics scores but that performed either as expected or worse than expected on their early professional benchmarks on variables of interest provided by the National Research Council. Selected comparison programs were required to have complete data on predoctoral characteristics variables.

Results

Before analyses, data were screened for univariate outliers and departures from normal distributions. There were no significant departures from normality, but four programs were identified as outliers on multiple variables and had unusual characteristics (e.g., recently accredited). These programs were removed, leaving 179 programs for subsequent analyses. The SEM depicted in Figure 1 fit the data well, $\chi^2(4) = 5.80$, $p = .21$; CFI = .99; TLI = .98; RMSEA = .05; SRMR = .06. As shown in Figure 1, predoctoral characteristics (measured by GRE Verbal, GRE Quantitative, and undergraduate GPA) significantly predicted emerging professional benchmarks (measured by EPPP passing and internship match rates) and accounted for the great majority of variability ($R^2 = .81$, $SE = .11$, $p < .001$).

Using the estimated SEM model, predicted program outcomes were compared with their observed outcomes (i.e., residuals calculated), and discrepancies were used to rank programs. Table 1 identifies the top 10 programs that performed better than expected with respect to the combined latent emerging professional benchmarks outcome, EPPP pass rates, and internship match rates, respectively. To rule out the possibility that only programs admitting individuals with relatively lower predoctoral characteristic scores could have better than expected outcomes (or vice versa), the correlation between the latent predoctoral characteristics factor scores and the latent emerging professional benchmarks residuals was calculated. There was a significant positive relationship ($r = .24$, $p < .01$), indicating that, in general, programs admitting students with relatively better predoctoral characteristics also tended to perform better than expected on emerging professional benchmarks.

Four programs, as shown in column 1 of Table 1, performed exceptionally better than predicted (i.e., operationalized as having residual scores more than two standard deviations above the mean) on the combined emerging professional benchmark adjusted outcomes. Program rankings did not significantly correlate with *USNWR* department rankings. To elucidate departmental characteristics that might have contributed to the substantially better than expected outcomes, these four programs were each individually linked to four comparison programs. Comparison programs were chosen for each of the top four programs based on their latent benchmark factor scores. We took each program's benchmarks and found schools that had the same or better benchmarks, but that did NOT perform better than predicted. The *closest* such program that had complete data available was yoked to its top sister school. Each yoked comparison program was a Ph.D. granting program in a public university. Exceptionally performing programs ($M = 42.0$, $SD = 17.6$) did not significantly differ from the comparison programs ($M = 41.7$ years, $SD = 28.9$) in terms of how long they have been accredited by the Commission on Accreditation (dates of initial accreditation were drawn from the Education Directorate of the American Psychological Association (http://www.apa.org/ed/accreditation/programs/clinical.aspx). None of the exceptionally performing programs are accredited by the Psychological Clinical Science Accreditation System (PCSAS), though two of the comparison programs are (*http://www.pcsas.org/accredited-programs.php*).

Data obtained from the National Research Council were obtained for each of the identified programs. For each variable

Table 1
Ranking of Top 10 Programs According to Adjusted[a] Program Outcomes (i.e., Emerging Professional Benchmarks Combined, EPPP Passing, and Internship Match Rates)

Emerging professional benchmarks combined[b]		EPPP passing rate		Internship matching rate	
Program	Ranking (standardized residuals)	Program	Ranking (standardized residuals)	Program	Ranking (standardized residuals)
Texas Tech U.	1 (3.02)	Western Michigan U.	1 (2.08)	Eastern Michigan U.	1 (2.20)
Georgia State U.	2 (2.82)	Texas A&M	2 (1.93)	Georgia State U.	2 (2.16)
Western Michigan U.	3 (2.11)	Texas Tech U.	3 (1.83)	Wichita State U.	3 (1.97)
Case Western Reserve U.	4 (2.01)	American U.	4 (1.58)	Bryn Mawr College	4 (1.76)
Bryn Mawr College	5 (1.82)	U. of North Texas[c]	5 (1.54)	Michigan State U.	5 (1.72)
Binghamton U./SUNY	6 (1.73)	U. of Wisconsin-Madison	6 (1.53)	U. of Nevada-Las Vegas	6 (1.70)
Eastern Michigan U.	7 (1.64)	U. of Tulsa	7 (1.46)	Washington U.	7 (1.66)
Sam Houston State U.	8 (1.59)	U. of South Carolina	8 (1.45)	Texas Tech U.	8 (1.62)
U. of Maine	9 (1.57)	Indiana U. of Pennsylvania	9 (1.40)	U. of Texas-Austin	9 (1.57)
U. of Montana	10 (1.52)	Idaho State U.	10 (1.38)	U. of Delaware	10 (1.52)

Note. Standardized residuals are reported in parentheses next to rankings.
[a] Outcomes were adjusted for differences in program students' typical pre-doctoral GRE and undergraduate GPA scores. [b] Based on the latent outcome from the SEM model in Figure 1. [c] There are two accredited programs in Clinical Psychology at this university; one broad program and one specific to Clinical Health Psychology and Behavioral Medicine. This is the broad program in Clinical Psychology.

presented in Table 2, a simple difference score was computed (ranked program − comparison program) and is presented in Table 3. Independent samples t tests were run to compare the exceptionally performing programs with their comparison programs on each variable. The only significant difference in scores was for percentage of non-Asian minority students enrolled in exceptionally performing ($M = 0.24$, $SD = 0.06$) and comparison ($M = 0.10$, $SD = 0.04$) programs; $t(5) = 3.36$, $p = .02$; 95% confidence interval, 0.03, 0.25. All other variables were nonsignificantly different.

Discussion

The current investigation sought to identify programs that are especially good at graduate training in professional psychology. The model subjected to SEM demonstrated a good fit with the data. Consistent with reports elsewhere (cf. meta-analysis by Kuncel, Hezlett, & Ones, 2001) the results support the validity of GRE and undergraduate GPA in predicting outcomes, including EPPP scores (Stedman & Schoenfeld, 2011), and underscore the need for continued inclusion of these data as indicators of incoming student quality and potential for successful outcomes in psychology graduate training. Results further indicated that beyond predoctoral characteristics, program training plays an important role in outcomes; several programs produce outcomes that are exceptionally better than predicted. Inspection of Table 1 highlights those latter programs, and it is worth noting that these are not necessarily those with highest reputational rankings (they do not significantly correlate with *USNWR* rankings). They also do not appear to be reflective of significant departmental differences on key data points, as indicated by examination of data from the National Research Council. It is not that highly regarded programs do not produce good outcomes. Typically, such programs attract well-qualified applicants who subsequently demonstrate good outcomes (see review by Stedman, 2007). But the present study sought to go beyond the predoctoral characteristics and identify those programs that are excelling at training students and producing significantly better outcomes than predicted. In short, these programs represent the hidden gems in the training community. Applicants and their mentors may want to consider this information when making decisions about where to apply for admission or which doctoral program offer of admission to accept.

Table 2
Descriptive Statistics on Departmental Characteristics for Exceptional Programs and Yoked Comparison Programs

Variable	Exceptional programs mean/*SD*	Comparison programs mean/*SD*
Publications per allocated faculty	0.99/0.53	0.74/0.23
Cites per publication	1.78/0.73	1.85/0.71
Percent faculty with grants	47.3%/18.3%	43.9%/17.7%
Percent faculty interdisciplinary	22.0%/20.4%	15.5%/17.2%
Percent non-Asian minority faculty	10.0%/5.5%	3.2%/1.8%
Percent female faculty	42.8%/9.9%	52.9%/18.5%
Awards per allocated faculty	0.08/0.09	0.15/0.13
Percent 1st yr. students with full support	91.7%/16.7%	100%/0%
Percent 1st yr. students with external funding	4.8%/9.5%	0%/0%
Percent non-Asian minority students	24.1%/6.0%	10.1%/4.3%
Percent female students	76.3%/9.4%	68.0%/8.9%
Percent international students	6.0%/3.6%	8.7%/10.3%
Average PhDs 2002 to 2006	9.8/5.7	8.5/7.4
Percent completing within 6 Years	40.1%/10.5%	42.5%/15.1%
Time to degree (in years)	6.0/0.2	6.2/0.8
Percent students Pursuing academic positions	43.2%/5.9%	53.5%/5.9%

Note. Data drawn from the National Research Council database.

Nationally, the rise in numbers of female faculty (American Psychological Association, 2012) and emphasis on interdisciplinary research (van Rijnsoever & Hessels, 2011) have been linked with more early career, junior faculty (Rhoten & Pfirman, 2007). Visual inspection of Table 3 might suggest the hypothesis that programs good at training tend to have proportionately more faculty in advanced ranks with many years of training experience. Unfortunately, this could not be determined from the department level data available. Although each of the value-added programs lists faculty on their Web sites, these data reflect current faculty and not necessarily the faculty composition during the data capture window for this study. Future studies examining this possibility are encouraged.

Several limitations deserve mention. Most importantly, we were limited to looking at only two indictors of emerging professional outcomes. Although important, they are rough proxies, with neither being a direct measure of success after graduate school (e.g., passing the EPPP[1] does not necessarily mean that one will be successful in practice). Moreover, neither reflects outcomes more salient to academic careers, where research productivity is more relevant. However, these outcomes do capture significant benchmarks facing the typical doctoral graduate in clinical psychology. Existing studies on placement rates in academic or research positions and number of publications by graduates, or future studies investigating research impact or grants awarded may identify a different set of programs as excelling in training. Of course, analyses were constrained by the data available; in this study it is notable that the GRE shares method variance with the EPPP. The identification and standard collection of additional benchmarks is strongly needed.

Notably, GPA and GRE scores were not publically disclosed, and therefore unavailable for analyses, among one third of programs in this study. Given that CoA does not require public disclosure of these variables, missing data on these variables were expected. However, missing data were not imputed in the SEM. Rather, we used full information maximum likelihood (FIML), which is the recommended approach (Enders, 2001; Schafer & Graham, 2002). By using this approach, all available data (which included consideration of more than GPA and GRE scores) were used in the SEM, with no programs excluded from SEM as a result of missing data on some variables. To examine the potential impact of including versus excluding programs with missing data, we ran the model again but excluded programs with incomplete data on the GPA and GRE variables. The model was again supported. However, because that approach would unnecessarily limit the sample given the data that are available and result in more

[1] Although we have no means to calculate reporting accuracy, we have heard complaints from training directors suggesting that EPPP data may not be accurately tied to specific doctoral programs.

Table 3
Differences on Departmental Characteristics Between Exceptional Programs and Their Yoked Comparison

Variable	Pair 1 differences	Pair 2 differences	Pair 3 differences	Pair 4 differences
Publications per allocated faculty	0.205	0.717	−0.244	0.53
Cites per publication	−0.602	0.612	0.174	0.169
Percent faculty with grants	−3.41%	18.83%	−0.33%	−0.45%
Percent faculty interdisciplinary	6.45%	11.99%	37.5%	−26.86%
Percent non-Asian minority faculty	−2.92%	14.58%	1.79%	4.46%
Percent female faculty	−8.11%	−20.08%	−42.86%	11.77%
Awards per allocated faculty	−0.008	0.137	0	−0.263
Percent 1st yr. students with full support	0	0	0	−33.33%
Percent 1st yr. students with external funding	19.05%	0	0	0
Percent non-Asian minority students	5.37%	15.37%	−2.55%	11.88%
Percent female students	−2.03%	0.3%	−3.74%	30.22%
Percent international students	−3.08%	9.09%	9.09%	−17.22%
Average PhDs 2002 to 2006	−5.2	14.6	1.6	0.2
Percent completing within 6 years	−1.3%	−6.615	−25%	22.74%
Time to degree (in years)	0	0.5	0.71	−1.26
Percent students in academic positions	−8.5%	−9.0%	−21.3%	−7.9%

Note. Pair differences reflect the discrepancy between the value reported for the top four ranked schools for emerging professional benchmarks and each program's yoked comparison school on the indicated variable, drawing from the National Research Council database.

limited implications we present only the results of the SEM using FIML. An implication of the SEM findings here is that GPA and GRE scores do seem to be very salient in the current training paradigm for our field. As a result, CoA should consider including these variables among the reporting requirements for accredited programs.

A final limitation that merits fairly extensive discussion is the decision to use match rates as reported by APPIC in analyses. Alternatively, we could have drawn match rates from the public disclosure data that each accredited program are required by the CoA to publish. Previous findings, reported by Parent and Williamson (2010), document that these two sources of data frequently fail to correspond. Unfortunately, APPIC and CoA used different standards to compute match rates during the years under consideration in the current study (2006–2010), which may account for the poor correspondence. Before the 2011 match, APPIC match rates reflected only those students who obtained an internship on "match day." Thus, any student who obtained an internship via the now defunct "Clearinghouse" was not included by APPIC in the doctoral programs' computed match rate. However, CoA standards (specifically Implementing Regulation C-20; CoA, 2011) required that programs report the number of students who "obtained" an internship, which includes both those students who secured an internship on match day and those students who subsequently obtained an internship via the Clearinghouse or other mechanism. In considering which metric to include in analyses, we determined that the APPIC match rate was preferable. Although APPIC match rates are often lower than those in programs' public disclosure data, we noticed that this was not always the case. Occasionally, programs' public disclosure data were exactly the same as reported by APPIC. Although this might mean the program never placed any students via the Clearinghouse, it could also simply reflect an unawareness of a program that CoA standards allowed for inclusion of additional students in their public disclosure data.

An additional complication was how students seeking an internship were identified. APPIC included all students who registered for the match, regardless of whether they subsequently withdrew formally or informally (i.e., by not submitting any applications). In contrast CoA did not stipulate that such students must be included. Rather, CoA noted that programs must report in their public disclosure data those who "sought or applied" for internship. As an example of how these subtle differences can result in a discrepancy, consider the student who has registered for the match but does not complete the dissertation proposal by a program-specified deadline. The program may not allow the student to seek an internship. Similarly, a student who faces a medical crisis or life change (e.g., pregnancy) after registering for the match might decide not to submit any applications. APPIC would consider these types of students to have gone unmatched when computing the program match rate. However, the students' programs would not likely view such students as seeking or applying for internship.

Thus, although the APPIC match rates may not have fully captured programs' success with students obtaining internships, the strength of the APPIC match rates was that they were uniform across programs and therefore more suitable for inclusion in analyses. Since the elimination of the Clearinghouse, APPIC is now tabulating match rates so that students who match in either phase of the match (Match I or Match II) are represented. Although this does not remedy the discrepancy in how students seeking internships are identified, it does solve at least one problem with computing match rates. Once sufficient data are accumulated from the two-phase matching system, it may be useful to update this study.

In summary, the current study found predoctoral characteristics (as indicated by GRE Verbal, GRE Quantitative and undergraduate GPA) account for the great majority (80.6%) of variability in emerging professional benchmarks outcomes (measured by EPPP pass rates and internship match rates). However, select programs appear to be value-additive during training and evidence a student body that is achieving markedly better than expected. In this study we took a conservative approach to identifying exceptional programs by focusing only on programs that demonstrated outcomes two or more standard deviations better than predicted, based on their predoctoral characteristics. The data, as we note above, contain inherent imperfections, and this conservative approach was taken to reduce the likelihood of the findings being spurious. Although it was tempting to use a more inclusive approach, such as including all programs that performed at least one standard deviation better than predicted, our concern was the findings might not replicate well or create the impression of a vanity listing.

Applicants to doctoral training in clinical psychology, as well as their mentors, are encouraged to consider the performance of programs on salient outcomes (as well as other important information such as research fit, financial incentives, proximity to family, long-term career goals, etc.) in determining which programs to apply to and how much weight to give an offer of admission from a value-additive program. As a policy implication, CoA is encouraged to begin using the data repositories resulting from their implementing regulations (or perhaps those from other sources as was done in this study) during program reviews. Programs that deviate markedly from normative data drawn across accredited programs could be empirically identified. Such an approach would potentially provide balance to the current review process in which programs are asked to selectively identify their own outcomes, their own measures, and their own analyses to formulate a narrative supporting their continued accreditation. Although we are not suggesting that the current approach be unilaterally dismissed, we do suggest that developing normative expectations that could be examined across programs fosters the best interests of public safety.

References

American Psychological Association. (2012). Demographic shifts in psychology. Retrieved from http://www.apa.org/workforce/snapshots/2003/demographic-shifts.aspx

Association of Psychology Postdoctoral and Internship Centers. (2010). APPIC match: 2000–2010: Match rates by doctoral program. Retrieved from http://www.appic.org/downloads/APPIC_Match_Rates_2000-10_by_Univ.pdf

Association of State and Provincial Psychology Boards. (2012). Psychology licensing exam scores by doctoral program. Retrieved from http://www.asppb.net/i4a/pp./index.cfm?pageid=3571.

Belar, C. D. (2000). Revealing data on education and training. *Clinical Psychology: Science and Practice, 7,* 368–369. doi:10.1093/clipsy.7.4.368

Burgess, J. F. Jr., Christiansen, C. L., Michalak, S. E., & Morris, C. N. (2000). Medial profiling: Improving standards and risk adjustments using hierarchical models. *Journal of Health Economics, 19,* 291–309. doi:10.1016/S0167-6296(99)00034-X

Commission on Accreditation. (2011). Policy statements & implementing regulations. Retrieved from http://www.apa.org/Ed./accreditation/about/policies/implementing-regs.pdf

Diamond, N., & Graham, H. D. (2000). How should we rate research universities? *Change, 32,* 20–33. Retrieved from http://www.pha.jhu.edu/~zbt/graham/change.htm doi:10.1080/00091380009601745

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling, 8,* 128–141. doi:10.1207/S15328007SEM0801_7

Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society, Series A, 159*(3), 385–443. doi:0035-9238/96/159385

Hanish, C., Horan, J. J., Keen, B., St. Peter, C. C., Ceperich, S. D., & Beasley, J. F. (1995). The scientific stature of counseling psychology training programs: A still picture of a shifting scene. *The Counseling Psychologist, 23,* 82–101. doi:10.1177/0011000095231009

Ilardi, S. S., Rodriguez-Hanley, A., Roberts, M. C., & Seigel, J. (2000). On the origins of clinical psychology faculty: Who is training the trainers? *Clinical Psychology: Science and Practice, 7,* 346–354. doi:10.1093/clipsy.7.4.346

Kazdin, A. E. (2000). Evaluating the impact of clinical psychology training programs: Process and outcome issues. *Clinical Psychology: Science and Practice, 7,* 357–360. doi:10.1093/clipsy.7.4.357

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin, 127,* 162–181. doi:10.1037/0033-2909.127.1.162

Morey, L. C. (2010). Leading North American programs in clinical assessment research: An assessment of productivity and impact. *Journal of Personality Assessment, 92,* 207–211. doi:10.1080/00223891003670133

Morse, R. (2012). Methodology: Graduate social sciences and humanities rankings. Retrieved from http://www.usnews.com/education/best-graduate-schools/articles/2012/03/12/methodology-graduate-social-sciences-and-humanities-rankings

Muthen, L. K., & Muthen, B. O. (2007). *Mplus. Statistical analyses with latent variables. User's guide* (5th ed.). Los Angeles, CA: Muthen & Muthen.

Parent, M. C., & Williamson, J. B. (2010). Program disparities in unmatched internship applicants. *Training and Education in Professional Psychology, 4,* 116–120. doi:10.1037/a0018216

Rhoten, D., & Pfirman, S. (2007). Women, science, and interdisciplinary ways of working. Inside Higher Ed. Retrieved from http://www.insidehighered.com/views/2007/10/22/rhoten

Rubin, D. B., Stuart, E. A., & Zanutto, E. L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics, 29,* 103–116. doi:10.3102/10769986029001103

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147–177. doi:10.1037/1082-989X.7.2.147

Schaffer, J. B., Rodolfa, E., Owen, J., Lipkins, R., Webb, C., & Horn, J. (2012). The examination for professional practice in psychology: New data-practical implications. *Training and Education in Professional Psychology, 6,* 1–7. doi:10.1037/a0026823

Stedman, J. M. (2007). What we know about predoctoral internship training: A 10-year update. *Training and Education in Professional Psychology, 1,* 74–88. doi:10.1037/1931-3918.1.1.74

Stedman, J. M., & Schoenfeld, L. S. (2011). Knowledge competence in clinical and counseling training and readiness for internship. *Journal of Clinical Psychology, 67,* 1–5. doi:10.1002/jclp.20740

van Rijnsoever, F. J., & Hessels, L. K. (2011). Factors associated with disciplinary and interdisciplinary research collaboration. *Research Policy, 40,* 463–472. doi:10.1016/j.respol.2010.11.001

Received July 10, 2012
Revision received May 24, 2013
Accepted May 29, 2013 ■