Evaluating the Readiness of Special Education Doctoral Students to Apply the Standards of Evidence-Based Practice to Single-Case Research

Michael R. Mayton

West Virginia University

Jie Zhang

SUNY Brockport

Stacy L. Carter

Texas Tech University

Jennifer L. Suppo

Seton Hill University

Author Note

Correspondence concerning this article should be addressed to Michael R. Mayton, Department of Special Education, 508-F Allen Hall, West Virginia University, Morgantown, WV 26506-6122.

E-mail: michael.mayton@mail.wvu.edu

Abstract

How well doctoral students in special education are prepared to evaluate research as evidence-based practice (EBP) is likely to impact their careers, as well as the teachers they will train. In developing a method for evaluating the readiness of small cohort groups of doctoral students to apply a research-based model of EBP, an instrument and procedure were refined in a pilot evaluation and implemented within a multiple baseline design across participants. Participants' independent and instrument-guided performance in rating published research was compared to the ratings of two experts in single-case research design, yielding proportions of agreement across evaluation conditions. Results indicated group readiness to independently conduct the EBP evaluation and individual differences in readiness indicating the need for remediation.

Keywords: evidence-based practice, doctoral preparation, special education

Evaluating the Readiness of Special Education Doctoral Students to Apply the Standards of Evidence-Based Practice to Single-Case Research

At almost every level of law and policy in the United States, higher education faculty members in special education are increasingly facing the need to embed within their teacher preparation activities knowledge and applications regarding evidence-based practices (EBPs). Teacher accreditation standards and ethical principles of professional practice have embraced EBP terminology, and these terms and concepts are now often discussed within the context of other, more traditional teacher activities, as illustrated by the Council for Exceptional Children's Advanced Preparation Standard 5: "Special education specialists provide leadership to formulate goals, set and meet high professional expectations, advocate for effective policies and evidence-based practices and create positive and productive work environments" (Council for Exceptional Children, 2012, p. 5). Federal education law has been at the forefront of the EBP movement, as exemplified by the No Child Left Behind Act (NCLB, 2001), which reportedly makes reference to the term "scientifically based research" (SBR) more than one hundred times (Browder & Cooper-Duffy, 2003) and almost synonymously uses the terms "evidence-based" and "peer-reviewed research" (or PRR; Zirkel, 2008). The latest amendments to the Individuals with Disabilities Education Act (IDEA, 2007) also use the term SBR and refer to the use of interventions that are "research based." However, as Detrich and Lewis (2013) concluded in their analysis of the 10-year impact of NCLB, it appears that more progress has been made with the development of EBP processes and standards than has been made with finding out how to actually apply them. Others, such as Reed & Reed (2008), have analyzed the state of EBP knowledge and practice and concluded that there still is no practical, universal application of these concepts for practitioners in our field.

Research-Based Implications for Doctoral Training Programs

Extant research indicates that special education doctoral students are more likely to seek higher education faculty positions than their predecessors (Washburn-Moses & Therrien, 2008). Within these positions, they will train the next generation of special education teachers who will be expected to know and apply EBP principles. This fact has direct implications regarding teacher and student performance. Paulsen (2005) concluded, in a preliminary study involving a sample of 130 first grade students, that when preservice special education teachers are taught to use EBPs and are required to use them, both the measured success of the teachers and that of their students with disabilities tend to increase.

The implications for doctoral faculty may be just as high stakes. For example, Washburn-Moses (2008) reported that 70% of special education doctoral students (n = 619) indicated the desire to pursue faculty and research careers. More importantly, their overall satisfaction with their doctoral programs and perceptions of career preparedness were found to be highly correlated with (a) satisfaction with doctoral research experiences, and (b) their preparation for conducting independent research. Within doctoral preparation programs, a lack of emphasis on translating research to practice (EBP) could therefore have a negative impact on the satisfaction of preservice special education faculty. Washburn-Moses identified the need to investigate possible connections between satisfaction and faculty attrition in the field.

The overall picture is unclear as to how well our doctoral programs are preparing future faculty in terms of EBP principles and practices. However, at least one research-based indicator may provide some limited insight into this question. Benedict, Johnson, and Antia (2011) studied 48 deaf education teacher preparation programs to determine important status-indicating characteristics, including the type of faculty skills most needed. The authors concluded that the

emphasis for required faculty effort in deaf education teacher preparation programs has been more on teaching courses and less on preparing future teachers to use EBPs. This is a relevant finding, especially considering that "the nation continues to face a shortage of faculty who can generate new knowledge about effective practices, translate such findings into teacher preparation programs' curriculum, and prepare a sufficient supply of new and highly skilled teachers" (Smith, Robb, West, & Tyler, 2010, p. 25).

Assessment Efforts in Other Fields of Doctoral Study

Emanuel, Robinson, and Korczak (2013) underscore the importance of knowledge and skill assessment in doctoral study and provide a detailed description of the system of comprehensive assessment used at Towson University in Maryland to evaluate student progress at critical points in their Doctor of Audiology (AuD) program. Among the range of their formative and summative assessments of student learning and practice is the Audiology Basic Skills Assessment, or ABSA, which directly addresses the key learning goal of "competency in scientific and research foundations of practice" (Emanuel, et al. 2013, p. 18-19). The authors state that the ABSA has been found to have some predictive value (better than entry grade point average) in identifying students who are at risk for program failure, reportedly due to a range of factors that are related to student engagement in poor research-based practices. Therefore, if Towson AuD students fail the ABSA, they must meet with the program director to outline how they will remediate needed areas of science-based knowledge and retake the exam during the following year. In addition, Sabus (2008) studied 31 physical therapy doctoral students by developing and using an EBP scale that was administered both before and after an in-service and student project designed to increase targeted EBP competencies. Findings were encouraging and

included significant improvement in student competency scores after engaging in the programmed educational activities.

For special educators, the application of single-case research designs (SCRDs) in assessment and intervention is considered to be an integral EBP (Tankersley, Harjusola-Webb, & Landrum, 2008), as long as such applications conform to a set of essential EBP standards and quality indicators (e.g., Horner, et al. 2005). Considering that many doctoral students in special education will train special education teachers who will in turn directly affect the lives of students from potentially vulnerable populations, there is a need for doctoral programs in the field assess knowledge and application of EBP standards, especially in the area of SCRDs. In beginning to address this need, the current evaluation project was designed with three main purposes: (1) to develop and refine a method for evaluating the preparedness of small cohort groups of doctoral students in special education to apply essential EBP principles to the evaluation of SCRDs, (2) to explore the utility of a range of methods for validly analyzing evaluation results, and (3) to demonstrate a method for making individualized inferences regarding student readiness across eight EBP standards.

PILOT EVALUATION

Method

Purpose

The pilot evaluation was conducted in order to do a preliminary check regarding participant comprehension of the standardized instructions for the task (both written and spoken), the efficiency and accessibility of the protocol items, and the manageability of the overall task (e.g., the time required and the appropriateness of the workload for a typical doctoral student). It

was also conducted as an initial indicator of possible evaluation outcomes for the pilot participant's program cohorts, who would participate in the actual evaluation.

Participant

None of the doctoral student participants in the pilot or main evaluation were students of any of the authors of the current project, nor were they scheduled to take (or in need of) any class to be taught by any of the authors for the remainder of their graduate programs. It is also important to note that none of the authors were chair or a member of any of the participants' dissertation committees. For the pilot, one doctoral student (referred to as Judy) was chosen at random from those who volunteered to participate in the current project (see Table 1 for Judy's relevant characteristics).

Setting

Judy scheduled two blocks of free time with the first author, one for the independent performance session and one for the instrument-guided performance session, and was asked to bring no class materials such as textbooks or notes. Both sessions were completed in a vacant conference room with no windows, comfortable office chairs, ample lighting, and a large table that allowed materials to be spread out for easy access. The room was reserved for a sufficient block of time to ensure the participant would not be disturbed. Judy was given access to snacks and bottled water and was told that she could close the door if she wished (e.g., for quiet or privacy), go for bathroom breaks, come to the researcher's nearby office with procedural questions, or simply leave at any time if she wished to cease participation. Several sharpened pencils and a highlighter were also provided.

Article Selection

The articles used for the project were selected by searching the EbscoHost meta-database, which searched other relevant databases such as Academic Search Complete, Education Research Complete, Education Resources Information Center, PsycArticles, and PsycInfo. Using the terms *disability* and *single-subject*, the first 10 articles were selected from the resulting list of publications that met the following conditions: (a) from an academic, peer-reviewed journal; (b) a single-subject intervention study; and (c) conducted with persons with disabilities.

Instrumentation

Article packets. In order to potentially reduce the amount of information that Judy had to examine and process during rating sessions, the introduction and review of literature were removed from each article. What remained were the sections describing the method, results, and discussion for each study. In order to reduce possible bias that could affect the ratings, headers and footers containing journal and author information were obscured. Each article packet initially consisted of the following: (a) a one-page sheet with instructions, an example item with descriptions of the item parts, and a list of important things for participants to remember (e.g., "Please do not discuss any details of the task or requirements with other participants or persons likely to talk with other participants, either before or after subsequent sessions."); (b) a rating form; and (c) one of the altered articles.

EBP protocol for single-case research. The EBP protocol was developed from the Horner, et al. (2005) discussion of necessary components required to determine if single-case research can be classified as EBP (as described and utilized in Mayton, Menendez, Wheeler, Carter, & Chitiyo, 2013; and Mayton, Wheeler, Menendez, & Zhang, 2010). Across 23 items/indicators representing eight main EBP standards such, as *Participants and Setting* and *External Validity*, the Protocol provides a set of conditions/operational definitions that must be

satisfied before a rating of *acceptable* can be recorded for a particular indicator. (The full protocol can be obtained by contacting the first author.)

Research Design and Procedure

The design for the pilot evaluation was a simple case study design (also called an A-B design) that consisted of an independent performance condition (condition one) followed by an instrument-guided performance condition (condition two). During condition one, Judy rated three articles using only an outline of the EBP Protocol, upon which she marked *acceptable* or *not acceptable* for each brief prompt in the list (e.g., the prompt, "Operational definition: ___ acceptable ___ not acceptable," which is one of the indicators listed under the standard, *Dependent Variable*). During condition two, Judy rated the remaining seven articles using the full EBP Protocol (with written decision making algorithms/ operational definitions included for each item) to aid her in the decision making process. During both conditions, no assistance was provided to the participant in making determinations for the ratings, including definitions of terms or questions regarding the details of the studies. Only questions asking for clarification of the instructions or procedure of the task were answered. Across both conditions, the order in which the articles were presented to the participant was determined by random assignment.

An expert in single-case research design (the first author, who has been trained on the content at the doctoral level, implemented various types of designs as part of the education and treatment of children and adults, served on graduate student thesis and dissertation committees in regard to knowledge of these designs, and taught the content at the graduate level) independently rated each of the 10 articles using the full EBP Protocol, and the expert's ratings were compared with the participant's ratings to calculate a percentage of agreement for each article. As a reliability check on the ratings of the first expert, a second expert on single-case design (the third

author, who met the same qualifications as the first) independently rated all 10 studies using the EBP Protocol. Inter-rater reliability (point-by-point percentage agreement; Kazdin, 2010) of the two sets of expert ratings was calculated by dividing the number of agreements by the number of agreements plus disagreements, and multiplying the result by 100. Across 230 individual ratings (10 studies x 23 ratings each), agreement between the experts was calculated at 93%. This was deemed a sufficient level of expert agreement in order to proceed with comparing the first expert's ratings to Judy's ratings. The premise applied within this evaluation procedure was that if students were adequately prepared by their doctoral program to accurately apply EBP concepts to published research, the agreement of their independent ratings with those of experts should not differ significantly from the agreement of their ratings produced while using an instrument that provides specific EBP definitions and evaluation procedures.

Results

For the pilot, descriptive statistics were used to examine preliminary data in order to determine general implications regarding overall participant performance and the implementation of evaluation procedures. To address the main purpose of exploring methods for evaluating results, a range of more extensive analyses were conducted with data gathered from the participants in the main evaluation that follows.

Agreement with the first expert during condition one (for articles two, five, eight, one, and four, ordered as randomly presented to the participant) ranged from 48% to 83% per article (M = 70.6; SD = 13.52; Mdn = 74.0). Agreement with the second expert during condition one ranged from 48% to 78% per article (M = 70.4; SD = 12.68; Mdn = 74.0).

Agreement with the first expert during condition two (for articles seven, ten, three, six, and nine, ordered as randomly presented to the participant) ranged from 52% to 91% per article

(*M* = 71.2; *SD* = 16.98; *Mdn* = 65.0). Agreement with the second expert during condition two ranged from 57% to 91% per article (*M* = 71.4; *SD* = 16.46; *Mdn* = 65.0).

Discussion

In seeking to accomplish the main purposes of the pilot evaluation, it was noted during both conditions when Judy had a question regarding the clarity of information (wording, order of steps, etc.) within the standardized instructions and items/sections of the protocol. In addition, it was discovered that complete removal of the review of literature from each article also removed the research question(s)/statements of purpose, which were later deemed necessary for assessing the internal validity of the studies. This information was requested by Judy late in the time-line of the pilot, and a page with the heading, "Research Question/Statement of Purpose," was then inserted into the remaining article packet, just prior to the altered article. Included on the page was a direct quote of the question(s)/purpose from the corresponding article. Interpreted along with the identification of these potentially confounding variables, the high variability in the proportions of agreement with expert ratings (e.g., from 48% to 83% in condition one, and from 52% to 91% in condition, across both experts) indicated the need to do the following prior to implementation of the main evaluation: (1) revise instructions and protocol items for increased clarity, as noted; (2) add research questions/statements of purpose at the beginning of each altered article/packet; and (3) systematically vary the order of presentation of articles across participants (through the use of a counterbalanced presentation), in order to control for the possibility of order and practice effects.

MAIN EVALUATION

Method

Participants

Initially, five doctoral students in special education responded to the email invitation and volunteered to be participants in the current project. One student was randomly chosen to participate in the pilot, and the remaining four were scheduled for participation in the main evaluation. However, just prior to the start of the independent performance session, one student in the group chose not to continue participation. The remaining three participants are herein referred to as Lisa, Robert, and Bill (see Table 1 for relevant participant characteristics).

Procedure and Research Design

The setting and articles were the same as those described in the pilot, with time scheduled for each participant to complete their ratings with no other participant present. However, prior to the actual evaluation, standardized instructions, article packets, and the EBP Protocol were altered as previously indicated.

Article packets were provided in an order that varied across participants (see Table 2), to control for the possibility of order effects and ensure that participants would not rate the same articles on the same days (an attempt to control for the possibility of outside discussion that could affect results). The procedure that was followed during the two conditions was otherwise identical to that used in the pilot investigation. It is also important to note that the main purpose for using a multiple baseline procedure in the article presentation was not to conduct an intervention study that would demonstrate a functional relationship between a dependent variable and the introduction of an independent variable. The use of the procedure was meant to systematically control for practice effects during the repeated measurements that took place across the two evaluation conditions. The use of repeated measurements was thought to be a stronger approach to evaluation than, for example, a simple pre/post measures design, but the use of this procedure required increased controls for increased credibility of results. Within an

evaluation, differential effects across conditions may or may not be present, and therefore no predictive expectations are in place prior to implementation, as with an intervention study. In fact, the lack of significant differences across conditions would be a desirable finding in terms of how well doctoral students had been prepared for the task.

Results

Comparison of Agreement Across Conditions

The Mann-Whitney U test was used to compare participant levels of agreement with expert ratings across the two evaluation conditions. This test was chosen because (1) sampling procedures could not meet the assumptions inherent within parametric tests; and (2) within the multiple baseline procedure, measures of condition one and condition two agreement inherently varied in number and could not be matched across conditions. Although this nonparametric test does not require data to be normally distributed, percentages were first transformed into arcsin values in order to meet assumptions regarding specific levels of measurement (i.e., at least the ordinal level of measurement). This type of data transformation makes proportional variance estimates more stable, especially when using statistical tests of significance that require levels of measurement more robust than nominal data (as described in Hopkins & Chappell, 1994). Using the Mann-Whitney U test with the transformed data grouped by evaluation condition, it was concluded that the level of agreement across the two conditions was not significantly different ($U = 73.50$; $Z = 1.62$; $p = 0.105$).

To measure agreement directly, Cohen's kappa was calculated across evaluation conditions, participants, and experts. For two participants, kappa coefficients increased across both experts by relatively insignificant amounts from condition one to condition two (.06 - .08). It should be noted that, in these two cases, kappa already indicated "fair" levels of agreement

(0.34 and 0.44) during condition one (for one possible set of descriptors of kappa levels of agreement, see Altman, 1991). However, for the agreement between the third participant and one of the experts (Expert 1), there was a larger increase in kappa from condition one to condition two (+ 0.28), with much lower levels of agreement (0.09 and 0.04; or "poor") in the first condition. Agreement between this participant and the other expert (Expert 2) increased by only a relatively small amount (see Table 3).

Comparison of Agreement within Sections of the Protocol

A chi-square analysis was conducted within ratings from six of the eight sections of the protocol. Two sections, F and H, produced values that were lower than recommended for use with this procedure (i.e., less than 5, as discussed within the seminal paper, Lewis & Burke, 1949). Even though this type of judgment has since been labeled as being overly conservative with the chi-square test (e.g., Delucchi, 1983), the authors thought it best to adhere to this long-time standard. Ratings from sections F and H were therefore examined with Fisher's exact test, a test that can accommodate these lower values within a two-by-two contingency table, and no significance was indicated (section F: $p = 0.68$; section H: $p = 1.00$). However, chi-square analysis indicated that within Section E, *Experimental Control / Internal Validity*, there was a significant relationship between evaluation condition and the level of agreement with expert raters [χ^2 (1, $N = 45$) = 8.76, $p = 0.003$], meaning that level of agreement with experts across participants within this section tended to be highly dependent upon whether they were in condition one or condition two (see Figure 1). The significance of relationships between these variables within the five remaining sections of the protocol (sections A, B, C, D, and G) did not fall within the predetermined confidence interval (*p* values were, respectively, .82, .83, .30, .78, and .83).

The Mann-Whitney U test was again used, but this time to compare levels of agreement within sections of the protocol. As in the previous comparison, condition one and condition two percentages of agreement across participants were first transformed into arcsin values in order to meet assumptions regarding the necessary level of measurement. No significance differences were found between condition one and condition two agreement within any section of the protocol (see Table 4). However, agreement in sections regarding the independent variable and experimental control were the closest to significance and should therefore be closely monitored.

Examination of Level, Trend, and Effect Sizes

Level. Across the two evaluation conditions, there were small increases in level (mean percentages) for the first two participants (+4% and +9%, respectively). The third participant showed the largest change in level from condition one to condition two, with an increase of 12% (see Figure 2 for level lines).

Trend. Trend lines were calculated for the condition one and condition two data sets of each participant, and R-squared values for each trend line were computed as an indicator of predictive reliability (see Figure 2 for trend lines and R^2 values). For two participants (Robert and Bill), condition one and condition two trend lines were parallel and increasing, suggesting that predicted change in agreement over time was similar in both conditions. However, for one participant (Lisa), the predicted condition one trend was decreasing, while the predicted condition two trend was maintaining (almost flat).

In interpreting these results, it is important to note that as an R-squared value more closely approaches 1, the corresponding trend line is more likely to be reliable. Only one R-squared value indicated high predictability within this analysis (that of Lisa's baseline trend: R^2 =

.91), mostly due to variance across the relatively few data points available for this type of analysis.

Effect sizes. Time-series data were examined with several nonparametric measures of effect size, none of which indicated notable differences between levels of condition one and condition two agreement. The C-statistic (Young, 1941; Tryon, 1982; Nourbakhsh & Ottenbacher, 1994) provides a method for examining small datasets for changes due to treatment effects and provides a *p* value for an initial dataset (no initial significance means proceed with the complete analysis) and the aggregation of initial and secondary datasets. Percentage of all non-overlapping data (or, PAND; Parker, Hagan-Burke, Vannest, 2007), percentage of data points exceeding the median (or, PEM; Ma, 2006), and percentage of non-overlapping data (or, PND; Mastropieri & Scruggs, 1985), are each systems of measurement that examine comparative data overlap in slightly different ways (see Table 5).

Treatment Integrity

Treatment integrity (TI) was measured by a doctoral student assistant using three documents: (a) a checklist of ordered procedures, (b) the full text of the standardized instructions for participants, and (c) a brief set of guidelines governing how types of participant questions should be handled (e.g., questions regarding the internal details of the studies, the definition of terms such as “internal validity,” and the selection of rating choices were to be answered in the following way: “I am sorry, but I cannot help you with that information, but if you have a question about the procedure that you will be going through today, I will be glad to help.”). Checklist results indicated 100% TI across conditions and participants, and no deviations from the expected delivery of instructions or the guidelines for participant questions were noted.

Discussion

Group Findings

The use of inferential statistics (i.e., the Man-Whitney) indicated that there were no significant differences in aggregated agreement between evaluation conditions, both within and across the eight EBP standards. This positive outcome can be broadly interpreted as an adequate demonstration that the doctoral students, viewed as a cohort group, could independently rate single-case research in terms of most EBP standards as well as they could with specific guidance in analysis and interpretation, as supplied through the use of the protocol. This general interpretation was also supported by chi-square and Fisher's exact results, which revealed no significant differences in all standards areas except one. In addition, Cohen's kappa revealed that, for two out of three participants, adequate levels of agreement with both experts was fairly stable across the two conditions, again suggesting that an applied understanding of EBP principles was in place prior to the start of the evaluation. However, this type of analysis can be likened to an initial group screening measure and therefore cannot be used to identify specific areas of individual need. Within the group analysis, specific areas identified as being in need of further, individualized analysis were: (a) the significant difference found within the standard, *Experimental Control*, and (b) agreement measures for the third participant, Bill. These and other areas evaluated were addressed in the level, trend, and effect size analyses conducted across individuals.

Results for Individuals

Lisa. The small, positive change in level between conditions indicates that, for Lisa, her relatively high level of mean agreement during independent performance was very comparable to that of her performance with specific guidance. This is further substantiated by the fact that none of the effect sizes computed across the four different nonparametric methods were found to be

significant. However, in comparing the trends across conditions, a decreasing trend in condition one was found, along with an almost flat, stable trend during condition two. Evaluated along with uniformly positive level change and effect size results, trend results suggest that Lisa's responding in the first condition was experiencing temporary drift (perhaps due to difficulty with the specific articles being rated, fatigue with the task, and/or simply being given too few opportunities to respond over time, as compared with other participants) that was initially refocused and then maintained throughout the longer guided performance condition. The recommendations for Lisa are therefore less comprehensive and include: (1) a brief follow-up prior to her graduation in regard to independent performance with EBP analysis of research, included as one subcomponent of an item within her comprehensive exams (to be incorporated and evaluated by her committee chairperson), and (2) an informal evaluation of the quality of the individual studies that she wishes to cite and reference in her dissertation research (to be supervised by the members of her doctoral committee).

Robert. The positive change in level between conditions for Robert (along with the fact that none of the effect sizes computed across the four different nonparametric methods were found to be significant) indicates that his quite high level of mean agreement during independent performance was very comparable to that of his performance with specific guidance. Although his trend lines have little predictive value, their directionality within both data sets shows almost parallel slopes of increase. Evaluated along with level change and effect size results, trend results suggest that Robert was the best prepared member of his cohort to evaluate single-case research as EBP. Therefore, no additional recommendations for follow up or remediation (beyond the routine comprehensive exam and dissertation procedures already in place) were deemed necessary for him.

Bill. The increasing change in level between conditions for Bill (+12%) was the largest among participants, indicating that he was comparatively less prepared than his cohorts for independent performance of the task. None of the effect sizes computed across the four different nonparametric methods were found to be significant, but PEM was close to significance (.67, with .70 as the point of significance), perhaps due to the fact that this computation is more sensitive to variance in the data rather than overlap, as with most of the other nonparametric measures in the analysis. Most of the significant difference between condition one and two proportions of agreement within the area of experimental control was due to Bill's 49% increase in agreement during the provision of specific guidance. Trends in both data sets were almost identical in slope and direction, both showing a barely increasing trend. Therefore, recommendations for Bill were as follows (to be implemented prior to his graduation): (1) direct instruction emphasizing the procedures and requirements for establishing experimental control within and across various types of SCRDs (to be delivered by an instructor during the upcoming semester, within a special topics course remaining in Bill's graduate program of study); (2) a comprehensive review of single-case research conventions as they relate to the eight EBP standards, including an oral exit examination (to be completed as a written assignment and interview in the previously mentioned special topics course); (3) more experience with rating single-case studies using expert guidance, as with the EBP Protocol (to be included as an entire item within his comprehensive exams, complete with a written overview of the process that was utilized, individual ratings of studies, and written justifications of ratings that cite specific article content).

Limitations

In condition one (independent performance), students were not asked to independently identify general areas of analysis for EBP, such as external and social validity, but were provided with these areas in the outline of the rating form, for purposes of a more valid comparison between evaluation conditions. In addition, all three participants had achieved relatively advanced progress in the doctoral program, including the successful completion of three research courses, one of which focused solely on SCRDs. It would have added an additionally informative layer to the evaluation to have measures that were taken earlier in their programs, for the purposes of comparison to later performance.

Conclusions and Implications for Practice

Within the current project, a research-based instrument and a systematically applied method of evaluation were successfully used to gauge the performance of a small cohort of special education doctoral students at both the group and individual level, to determine their readiness to evaluate SCRDs across eight EBP standards. The similar use of the *EBP Protocol for Single-Case Research* may therefore provide useful data for doctoral program faculty who wish to assess performance within this essential dimension of EBP for special educators.

The results obtained from an analysis of these data can be used in making subsequent decisions at both the programmatic and individual level. Group-level analyses from similar applications could reveal specific areas of doctoral student preparation (e.g., evaluating experimental control/internal validity within SCRD applications) within which a proportion of student ratings significantly differ from those of expert faculty, and, within applicable courses, a closer match could then be made between (a) specific EBP standards and (b) learning objectives and assignments. In addition, individualized results can provide valuable information leading to the type, extent, and delivery of remediation needed within the programs of targeted doctoral

students. For example, the proportional scope of ratings differences across all standards and indicators could be used to determine the extent of the remediation required, in terms of application across one, some, or all of a student's remaining programmatic components (compare results and subsequent recommendations across Robert, Bill, and Lisa). Likewise, the number of significant differences across individual standards could be used to determine the level of needed instruction (e.g., on a continuum from independent remediation with periodic monitoring to full instruction and intensive evaluation), thereby suggesting the most appropriate programmatic delivery modes (e.g., writing assignment, supervised practicum application, focused content within a course, comprehensive exam, or research proposal for dissertation). It is recommended that future projects of this type explore the relationship among evaluation results, extent and type of suggested remediation, and subsequent performance on relevant measures.

References

Altman, D. G. (1991). *Practical statistics for medical research*. London England: Chapman and Hall.

Benedict, K. M., Johnson, H., & Antia, S. D. (2011). Faculty needs, doctoral preparation, and the future of teacher preparation programs in the education of deaf and hard of hearing students. *American Annals of the Deaf, 156,* 35-46.

Browder, D. M., & Cooper-Duffy, K. (2003). Evidence-based practices for students with severe disabilities and the requirement for accountability in No Child Left Behind. *The Journal of Special Education, 37,* 157-163.

Council for Exceptional Children (2012). *CEC special education specialist advanced preparation standards.* Retrieved from http://www.cec.sped.org/~/media/Files/Standards/Professional%20Preparation%20Standards/Advanced%20Preparation%20Standards%20with%20Elaborations.pdf

Delucchi, K. L. (1983). The use and misuse of chi-square: Lewis and Burke revisited. *Psychological Bulletin, 94,* 166-176.

Detrich, R., & Lewis, T. (2013). A decade of evidence-based education: Where are we and where do we need to go? *Journal of Positive Behavior Interventions, 5,* 214-220.

Emanuel, D. C., Robinson, C. G., & Korczak, P. (2013). Development of a formative and summative assessment system for AuD education. *American Journal of Audiology, 22,* 14-25.

Hopkins, K. D., & Chappell, D. (1994). Quick power estimates for comparing proportions. *Educational and Psychological Measurement, 54,* 903-912.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., Wolery, M. (2005). The use of

single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71*, 165-179.

Individuals with Disabilities Education Act, Pub. L. 94-142 as amended, 20 U.S.C. § 1400 et seq. (2007).

Kazdin, A. E. (2010). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). New York: Oxford University Press.

Lewis, D., & Burke, C. J. (1949). The use and misuse of the chi-square test. *Psychological Bulletin, 46*, 433-489.

Ma, H. (2006). An alternative method for quantitative synthesis of single-subject researches: Percentage of data points exceeding the mean. *Behavior Modification, 30*, 598-617.

Mastropieri, M. A., & Scruggs, T. E. (1985). Early intervention for socially withdrawn children. *The Journal of Special Education, 19*, 429-441.

Mayton, M. R., Menendez, A. L., Wheeler, J. J., Carter, S. L., & Chitiyo, M. (2013). An analysis of Social Stories™ research using an evidence-based practice model. *Journal of Research in Special Educational Needs, 13*, 208-217.

Mayton, M. R., Wheeler, J. J., Menendez, A. L., & Zhang, J. (2010). An analysis of evidence-based practices in the education and treatment of learners with autism spectrum disorders. *Education and Training in Autism and Developmental Disabilities, 45*, 539-551.

No Child Left Behind Act of 2001, 20 U.S.C. 70 § 6301 et seq. (2002).

Nourbakhsh, M. R., & Ottenbacher, K. J. (1994). The statistical analysis of single-subject data: A comparative examination. *Physical Therapy, 74*, 768-776.

Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data (PAND): An alternative to PND. *The Journal of Special Education*, *40*(4), 194-204.

Paulsen, K. J. (2005). Infusing evidence-based practices into the special education curriculum. *Teacher Education and Special Education, 28,* 21-28.

Reed, F. D. D., & Reed, D. D. (2008). Towards an understanding of evidence-based practice. *Journal of Early and Intensive Behavior Intervention, 5*(2), 20-29.

Sabus, C. (2008). The effects of modeling evidence-based practice during clinical internship. *Journal of Physical Therapy Education, 22,* 74-84.

Smith, D. D., Robb, S. M., West, J., & Tyler, N. C. (2010). The changing education landscape: How special education leadership preparation can make a difference for teachers and their students with disabilities. *Teacher Education and Special Education, 33,* 25-43.

Tankersley, M., Harjusola-Webb, S., & Landrum, T. J. (2008). Using single-subject research to establish the evidence base of special education. *Intervention in School and Clinic, 44*(2), 83-90.

Tryon, W. W. (1982). A simplified time-series analysis for evaluating treatment interventions. *Journal of Applied Behavior Analysis, 15,* 423-429.

Washburn-Moses, L. (2008). Satisfaction among current doctoral students in special education. *Remedial and Special Education, 29,* 259-268.

Washburn-Moses, L., & Therrien, W. J. (2008). The impact of leadership personnel preparation grants on the doctoral student population in special education. *Teacher Education and Special Education, 31,* 65-76.

Young, L.C. (1941). On randomness in ordered sequences. *Annals of Mathematical Statistics, 12,* 153-162.

Zirkel, P. A. (2008). A legal roadmap of SBR, PRR, and related terms under the IDEA. *Focus on Exceptional Children, 40,* 1-4.

Table 1

Relevant Participant Characteristics

Participant	Age	Time Between Master's Degree and Start of Doctoral Program	Research Courses Completed	Doctoral Courses in Progress	Semesters Remaining in Program
Judy	35	11 years	2*	1	2
Lisa	46	20 years	2*	1	2
Robert	29	1 year	2*	2	1
Bill	31	1 year	2*	1	2

Note. *Included one course in single-case design.

Table 2

Order of Article Packet Presentation Across Participants

Participant	Order of Article Presentation									
Lisa	1	4	7	10	3	6	9	2	5	8
Robert	2	5	8	1	4	7	10	3	6	9
Bill	3	6	9	2	5	8	1	4	7	10

Note. Shaded cells represent articles rated during the independent condition (condition one).

Table 3

Cohen's Kappa Across Experimental Conditions, Participants, and Expert Raters

Participant	Condition 1: Expert 1	Condition 1: Expert 2	Condition 2: Expert 1	Condition 2: Expert 2	Change in kappa Coefficient
Lisa	0.34 (fair)	0.32 (fair)	0.42 (moderate)	0.38 (fair)	Expert 1: + 0.08 Expert 2: + 0.06
Robert	0.44 (moderate)	0.44 (moderate)	0.51 (moderate)	0.47 (moderate)	Expert 1: + 0.07 Expert 2: + 0.03
Bill	0.09 (poor)	0.04 (poor)	0.37 (fair)	0.13 (poor)	Expert 1: + 0.28 Expert 2: + 0.09

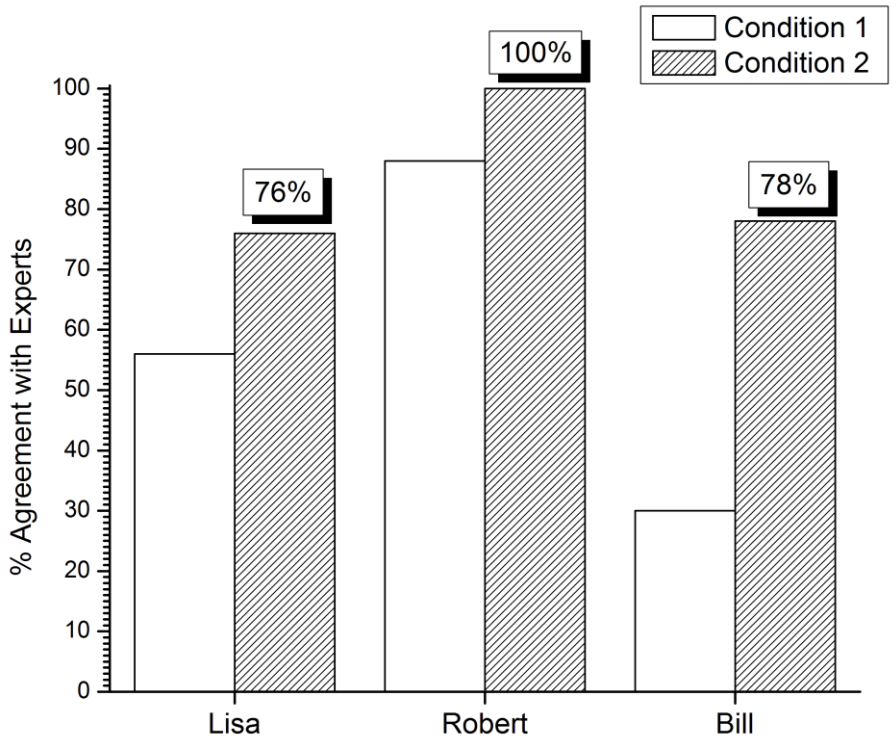

Figure 1. Level of agreement with expert ratings within section E (*Experimental Control*) of the EBP protocol across all studies, participants, and evaluation conditions.

Table 4

Mann-Whitney U Agreement Comparisons Within Protocol Sections

Section	Evaluation Standard	Significant difference between baseline & intervention agreement?
A	Participants and setting	No: U = 4.5; Z = .000; *p* = 1.0
B	Dependent variable	No: U = 4.0; Z = -.218; *p* = .83
C	Independent variable	No: U = 2.0; Z = -1.09; *p* = .28
D	Basline	No: U = 4.0; Z = -.225; *p* = .82
E	Experimental control (internal validity)	No: U = 2.0; Z = -1.09; *p* = .28
F	External validity	No: U = 2.5; Z = -.886; *p* = .38
G	Social validity	No: U = 3.0; Z = -.655; *p* = .51
H	Research questions	No: U = 4.0; Z = -.258; *p* = .80

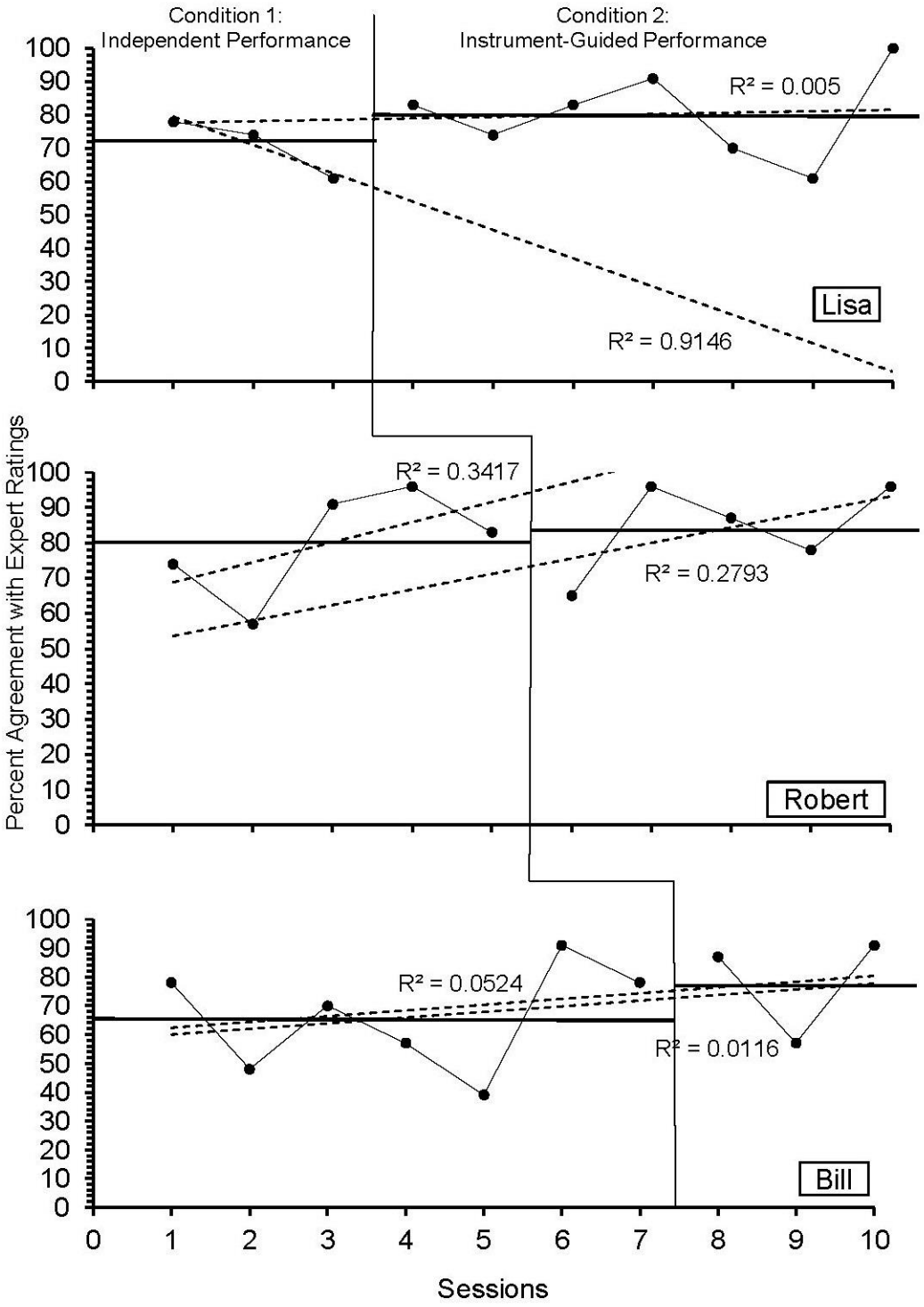

Figure 2. Percent agreement with expert ratings across sessions. Note that each participant rated one article per session and that multiple rating sessions occurred per day.

Table 5

Effect Sizes Across Four Nonparametric Methods

Method	Type	Effect Size	Score Type / Determination
C-statistic	logistic regression, non-parametric	<u>Condition 1</u> Lisa: .69 Robert: .58 Bill: .42 <u>Condition 1 & 2</u> Lisa: .75 Robert: .46 Bill: .42	*p* values / all values show no significance at the .01 level
PAND	non-regression based, non-parametric	Lisa: 100 Robert: 99.6 Bill: 99.8	percentage / either all or almost all of baseline data points overlapped intervention data points
PEM	non-regression based, non-parametric	Lisa: .57 Robert: .60 Bill: .67	score from 0 to 1 / scores less than .70 reflect a non-significant change
PND	non-regression based, non-parametric	Lisa: 57 Robert: 0 Bill: 0	percentage / scores at 50% or less reflect a non-significant change