Research in Developmental Disabilities 34 (2013) 2514–2522

Contents lists available at SciVerse ScienceDirect

Research in Developmental Disabilities

Review article

Evidence-based practice: A quality indicator analysis of peer-tutoring in adapted physical education

Laura Kalef [a], Greg Reid [a], Cathy MacDonald [b,*]

[a] Department of Kinesiology and Physical Education, McGill University, Montreal, QC, Canada H2W 1S4
[b] Department of Physical Education, State University of New York at Cortland, P.O. Box 2000, Cortland, NY 13045, United States

ARTICLE INFO

Article history:
Received 19 December 2012
Received in revised form 1 May 2013
Accepted 2 May 2013
Available online 7 June 2013

Keywords:
Quality indicators
Peer-tutoring
Adapted physical education
Evidence-based practice

ABSTRACT

The purpose of the research was to conduct a quality indicator analysis of studies investigating peer-tutoring for students with a disability in adapted physical education. An electronic search was conducted among English journals published from 1960 to November 2012. Databases included ERIC, PsycINFO, and SPORTDiscus. Fifteen research studies employing group-experimental (Gersten et al., 2005) or single-subject designs (Horner et al., 2005) met inclusion criteria. Each study was assessed for the presence and clarity of quality indicators. Group designs met an average of 62.5% essential and 69% desirable indicators. An average of 80% of indicators was present for single-subject designs. Results suggest claims of peer-tutoring being an evidence-based practice are premature. Recommendations for clarifying and applying the quality indicators are offered.

Contents

1. Introduction

Evidence-based practice (EBP) was first used in medicine in the late 1900s (Odom et al., 2005) and was defined as integrating individual clinical expertise with the best available evidence from systematic research (Balsor et al., 2000). Presently, EBP has been proposed in many areas including medicine (Sackett, 1997), rehabilitation (Cicerone et al., 2000),

* Corresponding author. Tel.: +1 607 753 4991.
E-mail addresses: catherine.macdonald@cortland.edu, greg.reid@mcgill.ca (C. MacDonald).

0891-4222/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ridd.2013.05.004

nursing (Melnyk, 2010), psychology (Kratochwill & Shernoff, 2004), and education (Pring & Thomas, 2004), including adapted physical education (Hutzler, 2011; Bouffard, & Reid, 2012 and Reid, Bouffard, and MacDonald (2012)). An emphasis has been placed on delivering services based on the best possible research evidence (Worrall, 2002). In accordance with policies such as the No Child Left Behind Act, many schools presently focus on improving the quality of education by implementing teaching practices that have been demonstrated to be effective by scientific evidence (Odom et al., 2005). Peer-tutoring involves one student enhancing the learning of another (Ehly & Larsen, 1976). Research suggests peer-tutoring has positive effects in physical education for students with a disability (Klavina, 2008; Lieberman, Newcomer, McCubbin, & Dalrymple, 1997) and therefore it might be a candidate for evidence-based practice.

In 2003, the Council for Exceptional Children's Division for Research established a task force to assess the quality of individual studies in special education and to identify effective practices. In the 2005 special issue of Exceptional Children (Odom et al., 2005), the task force identified four types of research methodologies used in special education; experimental group (Gersten et al., 2005), correlational (Thompson, Diamond, McWilliam, Snyder, & Snyder, 2005), single-subject (Horner et al., 2005) and qualitative designs (Brantlinger, Jimenez, Klingner, Pugach, & Richardson, 2005). For each methodology, indicators of research quality were presented to serve as standards for determining the strength of specific studies. In general, more quality indicators were assumed to be associated with research of high quality and therefore practices that might be considered evidence-based. For example, providing sufficient detail of participants for replicable precision is an indicator for single-subject articles (Horner et al., 2005). The quality indicators were developed by the Task Force which aimed to put forth clearly stated, understandable, and easily accessible guidelines to identify high-quality research in special education to serve researchers, reviewers, and practitioners who determine the usability of findings (Odom et al., 2005).

Gersten et al. (2005) outlined quality indicators for group-experimental and quasi-experimental research. They divided the quality indicators into two groups: *essential for quality* and *desirable for quality*. For a study to be considered "acceptable" it must: (a) meet all but one of the essential quality indicators and (b) at least one of the desirable quality indicators. For a report to be considered "high quality" it must: (a) meet all but one of the essential quality indicators and (b) at least four of the desirable quality indicators. Notably, the authors provided no clear rationale of those criteria. Essential quality indicators were arranged into four groups, coinciding with four areas of a group experimental research report. These included (a) describing participants, (b) implementation of the intervention and description of comparison conditions, (c) outcome measures, and (d) data analysis (Gersten et al., 2005). For example, one of the quality indicators for describing participants considers if sufficient information was provided to verify the participants' disabilities.

Horner et al. (2005) presented only a single list of essential quality indicators for single-subject research. They did not distinguish between essential and desirable indicators. Seven broad categories were proposed for single-subject research: (a) description of participants and settings, (b) dependent variable, (c) independent variable, (d) baseline, (e) experimental control/internal validity, (f) external validity, and (g) social validity. To attain a replicable status, a study must describe an intervention with sufficient clarity so that it may be duplicated. The current study aims to apply the aforementioned quality indicators for group-experimental and single-subject designs within the peer-tutoring literature to determine if this popular practice can be deemed evidence-based.

1.1. Peer-tutoring

Public Law 94-142, the Education for All Handicapped Children Act (1975), later renamed the Individuals with Disabilities Education Improvement Act (2004) requires students with a disability to receive physical education. Peer-tutoring, also known as peer-teaching, is one of the instructional models that can be considered an innovation in the way physical education subject matter is taught (Metzler, 2000) and might be particularly beneficial for students with a disability. Peer-tutoring can be categorized as peer-assisted learning and includes other forms such as class-wide peer-tutoring (Ward & Ayvazo, 2006) and peer-assessment (Block, Conatser, Montgomery, & Flynn, 2001). These instructional approaches use tutors to directly teach their peers, with minor variations. Class-wide peer-tutoring involves the class working in reciprocal, rotating roles of the tutor and the tutee (Ward & Ayvazo, 2006), whereas peer-tutoring occurs when students are arranged in established pairs (Ward & Lee, 2005). Peer assessment encompasses either method, with a specific focus on the outcome of students assessing others (Ward & Lee, 2005). In physical education, it is argued that peer-tutoring supports the inclusion of students with a disability (Block & Obrusnikova, 2007), provides social and health benefits (Block & Oberweiser, 1995), and enhances student learning and participation. For example, Lieberman et al. (1997) found that peer-tutoring can help students with Down syndrome become more focused and involved. Similarly, Ward and Ayvazo (2006) determined that peer-tutoring facilitated inclusion of students with autism and an enabled them to perform an increased number of correct ball catches. Peer-tutoring has been used among students with varying disabilities, including visual impairments (Wiskochil, Lieberman, Houston-Wilson, & Petersen, 2007), autism (Vansteenkiste, Lens, & Deci, 2006), and Down syndrome (Lieberman et al., 1997). Several studies have examined the effects of peer-tutoring on individuals with specific disabilities whereas others have investigated its effects on individuals with general disabilities (Turlington, 2009) or severe multiple disabilities (Klavina & Block, 2008). Peer-tutoring has been a popular research topic in adapted physical education and research supports it as a favorable practice, but according to CEC criteria, is the research evidence of sufficient quality to claim peer-tutoring as an evidence-based practice?

The current research will assess individual studies on peer-tutoring in adapted physical education using the quality indicators presented by The Council for Exceptional Children's Division for Research task force in 2005 (Odom et al., 2005). It

will also evaluate the clarity and ease of interpretation of the quality indicators themselves (Cook, Tankersley, & Landrum, 2009). Do they highlight all the necessary areas of the research studies? Are all indicators clearly identified and easy to interpret? Thus, the purpose of the current research is to assess individual studies using a quality indicator analysis to evaluate the effectiveness of peer-tutoring in adapted physical education. The clarity of the quality indicators was also evaluated to enhance their practicality.

2. Methods

2.1. Criteria for inclusion

Studies selected were required to meet several inclusion criteria. First, the primary purpose of the study was to investigate the use of peer-tutoring among students with a disability. Second, the authors had to specifically state that the intervention was taking place in physical education as opposed to a classroom. Third, the study adopted a single-subject or group-experimental design. Fourth, studies were published in peer-reviewed journals or included in PhD Dissertations and theses between 1960 and November 2012. Finally, all studies were data-based and written in English.

2.2. Search procedure

An electronic search was conducted among peer-reviewed English language journals published from 1960 to November 2012. Databases included ERIC, PsycINFO, Health & Psychosocial Instruments, Current Contents/All Editions, ProQuest Dissertations & Theses: Full Text, and SPORTDiscus. Specific search terms included ''physical activity'', ''physical education'', ''exercise'', ''sport'' and ''recreation''. These terms were used in conjunction with ''peer-tutoring'', ''class-wide peer-tutoring'', as well as ''disability'', or ''handicap''. To search for relevant studies that may have been overlooked, the footnote chasing approach was also used in which pertinent studies were identified in the reference lists of the articles.

2.3. Data compiling and analysis

Five hundred eighty-four studies were identified using the search terms. Notably, over 500 of these studies were doctoral dissertations. The large number of hits was likely attributed to the nature of the search in Dissertations & Theses: Full Text. For instance, articles from education, as opposed to physical education were identified. Only fifteen studies met the inclusion criteria. Each study was classified as single-subject or group experimental designs according to the definitions provided by the Exceptional Children task force in 2005 (Gersten et al., 2005; Horner et al., 2005). Each study was assessed using the quality indicator analysis identified for the respective research method: group experimental (Gersten et al., 2005) or single-subject (Horner et al., 2005). For group design articles, a distinction between acceptable quality and high quality was made by Gersten et al. (2005) based on the total number of indicators present. No such distinction was presented for single-subject design, however, the more indicators were assumed to be proportional to the strength of the claim for EBP (Horner et al., 2005).

To assess the individual studies, a four-step process was followed. First, the authors familiarized themselves with the quality indicators and posed any clarification questions they had. Second, each study was read carefully. Third, the studies were reviewed a second time to examine the presence of the quality indicators. Fourth, quality indicators that were considered vague were identified and critiqued individually among the authors.

Almost all of the studies involved students working in pairs, with peers teaching a student with a disability. One study (Ward & Ayvazo, 2006) applied class-wide peer-tutoring which required to work as tutor-tutee, but in small groups. Notably, only one study (Vivo, 2002) reported a reciprocal peer-teaching style, in which the student with a disability also taught his/her peer.

2.4. Inter-observer agreement

In addition to the primary author, an additional author, and an undergraduate student performed coding procedures to obtain an indication of inter-observer agreement. All observers were trained to reach an intra-observer reliability score of at least 80% prior to assessing the articles. Then, they randomly assessed 12 of the 15 studies, including both group and single-subject designs. Intra- and inter-observer agreements were assessed using percent agreement. The number of agreements between observers were summed and divided by the number of agreements plus disagreements between observers, and multiplied by 100. The average inter-reliability score for the group experimental design analysis was 97% and for single-subject designs was 96%. Because the agreement among observers was high, the results obtained by the primary author were used for analysis.

3. Results

Table 1 presents the number of quality indicators for the five studies with group-experimental designs (Gersten et al., 2005). Table 2 is a condensed version of Table 1, showing the total number of quality indicators present in each study. On

Table 1
Essential and desirable indicators for group experimental and quasi-experimental studies.

Quality indicators	Author, year of publication				
	Block et al. (2001)	DeLuzio (2009)	DePaepe (1985)	Halle and Gabler-Halle (1989)	Turlington (2009)
Essential					
Describing participants					
1. Was sufficient information provided to determine/confirm whether the participants demonstrated the disability(ies) or difficulties presented?	O	X	O	O	O
2. Were appropriate procedures used to increase the likelihood that relevant characteristics of participants in the sample were comparable across conditions?	O	X	X	O	O
3. Was sufficient information given characterizing the interventionists or teachers provided?	X	X	O	O	X
3.1 Did it indicate whether they were comparable across conditions?	O	X	O	O	O
Implementation of intervention and description of comparison conditions					
1. Was the intervention clearly described and specified?	O	X	X	X	X
2. Was the fidelity of intervention described and assessed?	O	X	O	X	O
3. Description of services provided in comparison conditions	O	X	X	X	X
Outcome measures					
1. Were multiple measures used to provide an appropriate balance between measures closely aligned with the intervention and measures of generalized performance?	X	X	X	X	O
2. Were outcomes for capturing the intervention's effect measured at appropriate times?	X	X	X	X	X
Data analysis					
1. Were the data analysis techniques appropriately linked to key research questions and hypotheses?	X	X	X	X	X
1.1 Were the data analysis techniques appropriately linked to unit of analysis?	X	X	X	X	X
2. Did the research report include not only inferential statistics but also effect size calculation?	O	X	O	O	O
Desirable					
1. Was data available on attrition rates among intervention samples?	X	X	X	X	X
1.1. Was severe overall attrition documented?	X	X	X	X	X
1.2. If so, is attrition comparable across samples?	n/a	n/a	n/a	X	n/a
1.3. Overall attrition less than 30%?	n/a	n/a	n/a	X	n/a
2. Did the study provide not only internal consistency reliability but also test-retest reliability and interrater reliability (when appropriate) for outcome measures?	O	X	X	O	X
2.1. Were data collectors and/or scorers blind to study conditions and equally (un)familiar to examinees across study conditions?	O	X	O	O	O
3. Were outcomes for capturing the intervention's effect measured beyond an immediate post-test?	O	O	O	X	O
4. Was evidence of the criterion-related validity and construct validity of the measures provided?	O	X	O	O	O
5. Did the research ream asses not only surface features of fidelity implementation, but also examine quality of implementation?	O	X	O	X	O
6. Was any documentation of the nature of instruction or series provided in comparison conditions?	O	X	X	X	O
Did the research report include audio or videotape excerpts that capture the nature of the intervention	O	X	O	O	O
8. Were the results presented in clear, coherent fashion?	X	X	X	X	X

Note: An "X" indicates the presence of a quality indicator. An "O" suggests the quality indicator was absent.

average, the studies included 62.5% of the essential indicators and 69% of the desirable indicators. In some studies the two attrition rate indicators were not applicable making the number of desirable indicators 10 rather than 12. The total number of essential quality indicators present in each study ranged from 8/22 (36%) to 20/22 (91%). Only one study (DeLuzio, 2009) was considered acceptable and high quality according to the criteria described by Gersten et al. (2005). The other studies failed to meet criteria to be considered acceptable.

Table 3 depicts the quality indicators present in the 10 single-subject reports. The number of quality indicators ranged from 0% (0/10) to 100% (19/10), with an average of 80% (8/10). As seen in Table 3, 40% (8/20) of the quality indicators were present in all studies. Several quality indicators were commonly absent and some were more frequent than others. For example, quality indicators addressing social and external validity were present in every study while only three of the studies included sufficient information in regards to *describing the setting of the interventions with replicable precision*.

Table 2
Total number of quality indicators within group-experimental studies.

Quality indicators	Author, year of publication				
	Block et al. (2001)	DeLuzio (2009)	DePaepe (1985)	Halle and Gabler-Halle (1989)	Turlington (2009)
Essential	5/12	12/12	7/12	7/12	6/12
Desirable	3/10	8/10	5/10	8/12	4/10
Total	8/22	20/22	12/22	15/24	10/22

Notably, one quality indicator was absent in all 10 articles, *controlling for common threats to internal validity*. Some quality indicators, such as *common threats to internal validity*, lacked clarity and were therefore difficult to assess.

4. Discussion

Only one group-experimental study by DeLuzio (2009) was considered acceptable and high quality according to the criteria described by Gersten et al. (2005). The other research reports failed to meet criteria to be considered acceptable. Essential indicators for describing *participants and the interventionist/teacher* were frequently omitted from the studies. According the Gersten et al. (2005), it is not sufficient to merely indicate the disability when describing participants; a diagnosis also needs to be documented. For example, Turlington (2009) stated that students had severe cognitive disabilities, physical and visual impairments, and/or used forms of mobility assistance, but did not further discuss the diagnoses or the specifics of each disability. More specific information such as indicating an individual with a physical disability has an amputated leg would enhance replicability of the study. It is possible the specifics of the disability were unavailable to the researcher or unable to be publically documented for confidentiality reasons. More detailed descriptions of background and capabilities of the individual or teacher implementing the intervention is also critical for replication of the research.

Four of five group design articles studies failed to include the essential indicator, *measurement of effect size*. Although the calculation was usually absent, several authors indicated that a small sample size may have contributed to the lack of significant findings. For example, DePaepe (1985) stated that his sample size was too limited for a power analysis, but indicated it should be calculated in future studies.

The final area of weakness in the *essential* group design quality indicators was the *description and assessment of the fidelity of the implementation*. Fidelity is critical to EBP because proper implementation of an intervention might increase the generalizability of the results (Protheroe, 2009). This indicator was only present in two of studies. DeLuzio (2009), for example, addressed fidelity by videotaping the peer-interactions among children with and without hearing loss, then analyzed quantity and quality of the interactions. Other studies involved peer tutor training to ensure the research was carried out in the way it was intended, but did not explicitly assess for fidelity (e.g., DePaepe, 1985). Since Horner and colleagues did not provide specific measures for fidelity, it was a challenge to assess and was identified as an area of discrepancy among raters.

Attrition rates were present in all five studies. The indicators related to attrition were presented in more general terms than most of the essential indicators. For example, the first indicator asked *if data was available on attrition rates among samples* and, therefore, to present this information, explicit documentation and description of attrition rates were unnecessary, as long as the information was available. In review of the five articles, there was a range of observations for this criteria; some explicitly mentioned attrition (Halle & Gabler-Halle, 1989) whereas tables of results were consulted in other studies. For example, DeLuzio (2009) provides tables with "n" values for both groups that are consistent with the initial number of participants, showing that no one dropped out of the study. Once the presence of this indicator was established, the criteria further asked *if severe attrition was documented and if so, if it is comparable across samples and if it is less than 30%*. These last two indicators were only relevant in one article, where severe attrition was presented (Halle & Gabler-Halle, 1989). In general, the presence of the indicators on the checklist is a positive attribution of the associated study, corresponding to high quality. This is not the case when reporting attrition, as indicating severe attrition would provide less effective results and negatively influence the quality of the paper. Using a grading scale may be helpful in this area as well, in order to assess quality. It may be true that a study is higher quality if there is explicit documentation of attrition information, as opposed to the information just being available. A scale would also incorporate the four attrition indicators into a combined criterion, and provide a positive result when indicating its presence. Alternatively, four indicators may be unnecessary for this topic and reducing to one or two at most may be more effective.

All five articles included the desirable indicator, *presented results in a clear, coherent fashion*, which might not be surprising as researchers often follow strict writing guidelines from the American Psychological Association and from journal editors in preparation for publishing.

The least frequently observed desirable indicators relate to measurement during intervention; *data collectors blind to the study conditions and equally unfamiliar to examinees across conditions*. It is common in peer-tutoring research that peer tutors are the data collectors (e.g., Block et al., 2001; DePaepe, 1985), or the researcher (e.g., DeLuzio, 2009), as opposed to a blind data collector. A data collector who is unaware of the purposes of the study poses some challenge, as the students are more likely to perform better as they become more comfortable with an individual as a relationship evolves. Although this may

Table 3
Number of quality indicators within single-subject research.

Quality indicators	Donder and Nietupski (1981)	Vivo (2002)	Houston-Wilson et al. (1997)	Klavina (2008)	Klavina and Block (2008)	Lieberman et al. (1997)	Lieberman et al. (2000)	Ward and Ayvazo (2006)	Webster (1987)	Wiskochil et al. (2007)
Descriptions of participants and settings										
Sufficient detail	X	X	O	X	X	X	X	X	X	X
Participant selection process described with replicable precision	O	X	O	X	X	O	X	X	X	X
Critical features of physical setting described with replicable precision	O	O	O	X	O	O	X	X	O	O
Dependent variables										
Described with operational precision	X	X	X	X	X	X	X	X	O	O
Each measured with procedure that generate quantifiable index	X	X	X	X	X	X	X	X	X	X
Valid measurement described with operational precision	X	X	X	X	X	X	X	X	X	X
Measured repeatedly over time	X	X	X	X	X	X	X	X	X	X
Data on reliability and inter-observer agreement associated with each dependent variable		X	X	X	X	X	X	X	X	X
Independent variable										
Described with replicable precision	O	X	X	X	X	X	X	X	O	X
Systematically manipulated, under control of experimenter	X	X	X	X	X	X	X	X	X	X
Overt measurement of fidelity of implementation	O	X	X	X	X	O	X	X	X	X
Baseline										
Baseline conditions described with replicable precision	X	X	X	O	O	X	X	X	O	X
Experimental control/internal validity										
At least 3 demonstrations of experimental effects at 3 different times	X	X	X	O	X	X	X	X	X	X
Common threats to internal validity controlled	O	O	O	O	O	O	O	O	O	O
Results document pattern that demonstrates experimental control	X	X	X	X	X	X	X	X	X	X
External validity										
Experimental effects replicated across participants, settings, materials, to establish external validity		X	X	X	X	X	X	X	X	X
Social validity										
Dependent variable is socially important	X	X	X	X	X	X	X	X	X	X
Magnitude of change in dependent variable resulting from independent variable is socially important	X	X	X	X	X	X	X	X	X	X
Independent variable implementation is practical and cost effective	X	X	X	X	X	X	X	X	X	X
Implementation of independent variable over extended time periods, by typical agents, in typical physical/social contexts	X	X	X	X	X	X	X	X	X	X

Note: An "X" indicates the presence of a quality indicator. An "O" suggests the quality indicator was absent.

result in bias, it may be beneficial from a health and safety standpoint to have an individual who is familiar with the participants.

Desirable indicators, measurements beyond an immediate post-test, and evidence of criterion and construct validity of measures used were frequently absent in the group designs. The one study that presented measurements beyond an immediate post-test had a group design study that looked specifically at each individual in the group over a period of 13 weeks (Halle & Gabler-Halle, 1989). The presentation of the measurements provided the reader with more concrete evidence of sustainability of each individual than would be seen with only pre- and post-testing. This difference makes it clear to the reader why this indicator is desirable for a high quality paper, yet is more frequently seen in single-subject design articles that follow a multiple baseline approach (Horner et al., 2005). Evidence of criterion and construct validity also provides the reader with information regarding the credibility of the implementation, which is important for replication. But such validity was largely absent, with only DeLuzio (2009) addressing the constructs.

The final desirable quality indicator missing from four of the five group-design studies was *inclusion of audio or video excerpts*, which are meant to capture the nature of the intervention. In the studies, it was common to use videotaping as a form of data collection, yet only one study, which was a thesis with appendices, provided excerpts for the reader (DeLuzio, 2009). While documented evidence would enhance research quality, publishing restrictions such as page limitations frequently prevent researchers from including this information. Perhaps such information could be placed in digital format at a site associated with the journal or the author. However, there may also be restrictions of public sharing of such data for ethical reasons in some jurisdictions.

Ten single-subject studies were evaluated using the criteria outlined by Horner et al. (2005). Notably, no specific number of criteria has been proposed (Horner et al., 2005), to deem a study as evidence-based. Thus, it might be assumed that a greater number of quality indicators is indicative of a higher quality study. The average number of indicators present in the 10 reviewed articles was 85% (16.9/20). Four indicators were most commonly absent, including *description of the method of selecting participants*, *description of setting*, *baseline described with replicable precision*, and *controlling for internal validity*. *External validity* and *social validity* were present in all studies.

While some studies described the participants with replicable precision, including the sampling design (Klavina & Block, 2008), several described only who was selected, as opposed to how they were selected (e.g., Houston-Wilson, Dunn, Mars, & McCubbin, 1997). Most studies incorporated descriptions of the participants' disability, including a diagnosis, which was less frequently present in the group designs. Setting descriptions ranged from providing general details, such as indicating the research took place in physical education (e.g., Lieberman et al., 1997), but few provided detailed accounts of the environment including the characteristics of the facility (but see Ward & Ayvazo, 2006 for some excellent specific examples). The studies were typically conducted in schools, therefore, purposeful rather than random sampling was frequently used. Since the indicators describing participants and settings are open to the assessor's interpretation, perhaps a rubric could be created and implemented to assess the quality of each, as opposed to a checklist. A detailed rubric would provide a clear indication of study quality in terms of meeting insufficient, sufficient, or exceptional standards, as seen in recent studies that have assessed quality indicators (Baker, Chard, Ketterlin-Geller, Apichatabutra, & Doabler, 2009; Chard, Ketterlin-Geller, Baker, Doabler, & Apichatabutra, 2009).

The quality indicator *controlled for common threats to internal validity by permitting elimination of rival hypotheses* was usually absent. Rival hypotheses were commonly discussed in the studies; however, they were not necessarily controlled to promote validity. For example, Lieberman,. Dunn, van de Mars, and McCubbin, (2000) described measurement effects that may have contributed to changes in student activity levels, yet did not control for them. Understandably, rival hypotheses will be presented as no study is perfect.

Every study included the quality indicators *external validity and social validity*. External validity is commonly demonstrated in single-subject designs by demonstrating effects using multiple participants (e.g. Vivo, 2002), settings (e.g. Webster, 1987), materials and/or behaviors (e.g. Donder & Nietupski, 1981). There may be a bias involved in terms of the frequency of the social validity indicators. The evaluators agreed that research on peer-tutoring has social importance as each study demonstrated effective intervention strategies that are feasible and easy to implement. However, since there is no clear definition of what 'socially important' means, personal bias has to be considered when evaluating this indicator.

The present research has some limitations. First, the research was limited to full-text articles only and foreign-language studies were excluded. Research on peer-tutoring presented at conferences was also excluded. Inclusion of these studies may have affected the results. Second, although every effort was made to conduct an extensive search of the literature using various search engines, some articles might have been missed.

This study might serve as a reminder to practitioners to embrace empirically based practices in adapted physical education, when possible. While some interventions have supporting research evidence (e.g. Sugden & Henderson, 2007) many are guided by 'best practices', or actions that have the authority and legitimacy of experienced practitioners or simply abide by common sense. Practitioners are encouraged to use empirical evidence in addition to their professional skills, competence, and expertise (Davies, 1999) when applying interventions such as peer-tutoring.

Researchers conducting future studies should make efforts to meet quality indicators. Much attention and care must be given to research design given that many criteria rely heavily on its quality. For example, an essential indicator for group designs considers the appropriateness of the data techniques and their link to key research questions and hypothesis. According to this review, some indicators can be met rather easily. For instance, all single-subject studies included social and external validity. Researchers should develop studies that are practical and include a sufficient number of participants. For

these same studies, more indicators would easily have been met if researchers simply described the setting, participants, and baseline intervention in more detail to warrant replication. While it is recognized that these indicators will be met with varying degrees of precision, efforts should be made to achieve a degree of methodological rigor.

5. Conclusion

Fifteen studies that examined the effects of peer-tutoring in physical education were divided into two groups: group experimental and single-subject designs. Five group-experimental designs and 10 single-subject studies were evaluated individually using the quality indicators developed for their respective methodologies (Gersten et al., 2005; Horner et al., 2005). Only one group-experimental study by DeLuzio (2009) was considered acceptable and high quality according to the criteria described by Gersten et al. (2005). The other studies failed to meet criteria to be considered acceptable. None of the single-subject designs (Horner et al., 2005) provided all 20 quality indicators but overall they possessed 85% of the indicators. In general however, both the group and single-subject design studies showed positive effects for peer-tutoring. In conclusion, results suggest claims of peer-tutoring for students with a disability in physical education being an EBP are promising but premature according to quality indicators presently available.

The quality indicators were also evaluated for clarity. They provide a systematic way to evaluate the merits of research, but do have some limitations. Continued testing and subsequent revisions of the quality indicators is necessary, as difficulties arose in understanding indicator definitions. Several recommendations can be proposed for future research. First, whereas Gersten and colleagues propose specific criteria for evaluating high versus acceptable quality research for group designs, Horner and colleagues do not provide such criteria. Rather, it must be assumed that a greater number of quality indicators are indicative of a greater quality of research, but at what level of quality must a study reach to claim evidence-based practice. It might be beneficial to develop more specific criteria for evaluating research reports, perhaps using the 'essential' and 'desirable' categories similar to those put forth by Gersten and colleagues. Presenting such criteria would make the quality indicators more 'user friendly' and promote consistency for evaluating research methodologies. Second, several of the group and single-subject quality indicators (e.g., *fidelity*, *social validity*, *presented results in a clear* and *coherent manner*) require a clearer definition of their exact meaning. For example, Gersten et al. (2005) did not provide clarifications of the quality indicator *results are presented in a clear and coherent way*, therefore it is suggested that when the indicators are revised, this one is given a precise meaning that can be interpreted in one way. Similarly, Horner et al. (2005) did not include criteria for exactly what needs to be measured for fidelity. Is there a minimum level of fidelity that is acceptable? More specific details as to how it should be evaluated would enhance consistency among researchers. Elimination of subjective terminology within descriptions of the indicators such as "sufficient detail", and "critical features" might be helpful in clarifying the meaning of the indicators. Exemplars of each quality indicator might also enhance clarity. Ultimately, obtaining a measure of inter-observer reliability, as in the current study, is important in determining the most effective way to present the quality indicators to researchers. Third, it would be beneficial to produce a meta-analysis reviewing the peer-tutoring studies to compare these quantitative results with a quality indicator analysis. Fourth, investigators should address the applicability of the quality indicators in other areas of research, such as correlational (Thompson et al., 2005) and qualitative (Brantlinger et al., 2005). Finally, with specific reference to adapted physical education, more research should be conducted to determine other practices which might be validated as evidence-based.

Acknowledgement

Our thanks to Melanie Kasner for assisting with the reliability analysis.

References

Baker, S., Chard, D., Ketterlin-Geller, L., Apichatabutra, C., & Doabler, C. (2009). Teaching writing to at-risk students: The quality of evidence for self-regulated strategy development. *Exceptional Children, 75*, 303–318.

Balsor, B., Beattie, P., Berk, A., Binkely, J., Brenneman, S., & Brett, L. (2000). Against the myth of evidence-based practice. *Journal of Orthopaedic and Sports Physical Therapy, 30*, 98–99.

Block, M., Conatser, P., Montgomery, R., & Flynn, L. (2001). Effects of middle school-aged partners on the motor and affective behaviors of students with severe disabilities. *Palaestra, 17*, 34–39.

Block, M., & Oberweiser, B. (1995). Using classwide peer-tutoring to facilitate inclusion of students with disabilities in regular physical education. *Physical Educator, 52*, 47–56.

Block, M., & Obrusnikova, I. (2007). Inclusion in physical education: A review of the literature from 1995–2005. *Adapted Physical Activity Quarterly, 24*, 103–124.

Brantlinger, E., Jimenez, R., Klingner, J., Pugach, M., & Richardson, V. (2005). Qualitative studies in special education. *Exceptional Children, 71*, 195–207.

Bouffard, M., & Reid, G. (2012). The good, the bad, and the ugly of evidence-based practice. *Adapted Physical Activity Quarterly, 29*, 1–24.

Chard, D., Ketterlin-Geller, L., Baker, S., Doabler, C., & Apichatabutra, C. (2009). Repeated reading interventions for students with learning disabilities: Status of the evidence. *Exceptional Children, 75*, 263–281.

Cicerone, K. D., Dahlberg, C., Kalmar, K., Langenbahn, D. M., Malec, J. F., Bergquist, T. F., et al. (2000). Evidence-based cognitive rehabilitation: Recommendations for clinical practice. *Archives of Physical Medicine and Rehabilitation, 81*, 1596–1615.

Cook, B. G., Tankersley, M., & Landrum, T. J. (2009). Determining evidence-based practices in special education. *Exceptional Children, 75*, 365–383.

Davies, P. (1999). What is evidence-based education? *British Journal of Educational Studies, 47*, 108–121.

DeLuzio, J. M. (2009). *Peer interactions of preschool children with and without hearing loss*. Toronto, ON: University of Toronto (unpublished doctoral dissertation).

DePaepe, J. (1985). The influence of three least restrictive environments on the content motor-ALT and performance of moderately mentally retarded students. *Journal of Teaching in Physical Education, 5*, 34–41.

Donder, D., & Nietupski, J. (1981). Nonhandicapped adolescents teaching playground skills to their mentally retarded peers: Toward a less restrictive middle school environment. *Education and Training of the Mentally Retarded,* 271–276.
Education for All Handicapped Children Act. (1975). *Public Law No. 94-142*.
Ehly, S., & Larsen, S. (1976). Peer-tutoring to individualize instruction. *The Elementary School Journal,* 475–480.
Gersten, R., Fuchs, L., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children, 71*, 149–164.
Halle, J., & Gabler-Halle, D. (1989). Effects of a peer-mediated aerobic conditioning program on fitness measures with children who have moderate and severe disabilities. *Journal of the Association for Persons with Severe Handicaps, 14*, 33–47.
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.
Houston-Wilson, C., Dunn, J., Mars, H. v. d, & McCubbin, J. (1997). The effect of peer tutors on motor performance in integrated physical education classes. *Adapted Physical Activity Quarterly, 14*, 298–313.
Individuals With Disabilities Education Improvement Act. (2004). *Public Law No. 108-446, 118 Stat. 2647*http://idea.ed.gov/.
Hutzler, S. (2011). Evidence-based practice and research: A challenge to adapted physical activity. *Adapted Physical Activity Quarterly, 28*, 189–209.
Klavina, A. (2008). Using peer-mediated instructions for students with severe and multiple disabilities in inclusive physical education: A multiple case study. *European Journal of Adapted Physical Activity, 1*, 7–19.
Klavina, A., & Block, M. (2008). The effect of peer-tutoring on interaction behaviours in inclusive physical education. *Adapted Physical Activity Quarterly, 25*, 132–158.
Kratochwill, T. R., & Shernoff, E. S. (2004). Evidence-based practice: Promoting evidence-based interventions in school psychology. *School Psychology Review, 33*, 34–48.
Lieberman, L., Newcomer, J., McCubbin, J., & Dalrymple, N. (1997). The effects of cross-aged peer tutors on the academic learning time of students with disabilities in inclusive elementary physical education classes. *Brazilian International Journal of Adapted Physical Education Research, 4*, 15–32.
Lieberman, L., Dunn, J., van de Mars, H., & McCubbin, J. (2000). Peer tutors effects on activity Levels of deaf students in inclusive elementary physical education. *Adapted Physical Activity Quarterly, 17*, 20–39.
Melnyk, B. M. (2010). *Evidence-based practice in nursing and health care: A guide to best practice*. Ambler, PA: Lippincott Williams & Wilkins.
Metzler, M. (2000). *Instructional models for physical education*. Boston, MA: Allyn and Bacon.
Odom, S., Brantlinger, E., Gersten, R., Horner, R., Thompson, B., & Harris, K. (2005). Research in special education: Scientific methods and evidence-based practices. *Exceptional Children, 71*, 137–148.
Pring, R., & Thomas, G. (Eds.). (2004). *Evidence-based practice in education*. New York, NY: McGraw-Hill International.
Protheroe, N. (2009). Fidelity of implementation. *Principal's Research Review, 4*, 1–8.
Reid, G., Bouffard, M., & MacDonald, C. (2012). Creating evidence-based practice in adapted physical activity. *Adapted Physical Activity Quarterly, 29*, 115–131.
Sackett, D. L. (1997). Evidence-based medicine. *Seminars in Perinatology, 21*, 3–5.
Sugden, D. A., & Henderson, S. E. (2007). *Ecological intervention for children with movement difficulties*. London: Pearson.
Thompson, B., Diamond, K. E., McWilliam, R., Snyder, P., & Snyder, S. W. (2005). Evaluating the quality of evidence from correlational research for evidence-based practice. *Exceptional Children, 71*, 181–194.
Turlington, H. L. (2009). *The attitude change of second-grade peer tutors working with students with severe disabilities through Laban's movement analysis*. Cullowhee, NC: Western Carolina University (unpublished master's thesis).
Vansteenkiste, M., Lens, W., & Deci, E. L. (2006). Intrinsic versus extrinsic goal contents in self-determination theory: Another look at the quality of academic motivation. *Educational Psychologist, 41*, 19–31.
Vivo, M. F. (2002). *The effects of peer-tutoring on the academic learning time in physical education of elementary school students with visual impairments in inclusive physical education classes*. Tallahassee, FL: Florida State University (unpublished doctoral dissertation).
Ward, P., & Ayvazo, S. (2006). Classwide peer-tutoring in physical education: Assessing its effects with kindergardeners with autism. *Adapted Physical Activity Quarterly, 23*, 233–244.
Ward, P., & Lee, M. A. (2005). Peer- assisted learning in physical education: A review of theory and research. *Journal of Teaching in Physical Education, 24*, 205–225.
Webster, G. (1987). Influence of peer tutors upon academic learning time-physical education of mentally handicapped students. *Journal of Teaching in Physical Education, 6*, 393–403.
Wiskochil, B., Lieberman, L., Houston-Wilson, C., & Petersen, S. (2007). The effects of trained peer tutors on the physical education of children who are visually impaired. *Journal of Visual Impairment & Blindness, 101*, 339–350.
Worrall, J. (2002). What evidence in evidence-based medicine? *Philosophy of Science, 69*, 316–330.