Validity and Problem-Based Learning Research: A Review of Instruments Used to Assess Intended Learning Outcomes

Brian R. Belland, Brian F. French, and Peggy A. Ertmer

Abstract

Problem-based learning (PBL) spread from the medical school to other university and K-12 contexts due, in part, to the stated promise that PBL produces the target outcomes of deep content learning, increased problem-solving ability, and increased self-directed learning (Hmelo-Silver, 2004). However, research results have been unclear. This paper examines how the three target outcomes of PBL were measured in 33 empirical studies. Results indicate that few studies included 1) theoretical frameworks for the assessed variables and constructs, 2) rationales for how chosen assessments matched the constructs measured, or 3) other information required for readers to assess the validity of authors' interpretations. Implications for future research are discussed.

In problem-based learning (PBL), students generate and pursue learning issues to understand an ill-structured problem and develop a feasible solution (Hmelo-Silver, 2004). Initially developed to improve medical students' problem-solving and self-directed learning abilities (Barrows & Tamblyn, 1980), PBL has since spread to many levels of education (K-12, undergraduate, and graduate) and a variety of disciplines, ranging from language arts to biology (Barrows & Tamblyn; Chin & Chia, 2005; Gallagher, Stepien & Rosenthal, 1992; Reiter, Rasmann-Nuhlicek, Biernat, & Lawrence, 1994; Torp & Sage, 1998). This increase in PBL use has been due largely to PBL's stated promise to promote deep content learning (Hmelo-Silver) as well as students' problem-solving and self-directed learning abilities. While many authors have described the difficulty in achieving these outcomes (Colliver, 2000; Dochy, Segers, Van den Bossche, & Gijbels, 2003; Vernon & Blake, 1993), few have discussed the difficulty in operationalizing and measuring these outcomes. Even when researchers tackle this task, their efforts are not always as transparent as they could be, thus making it difficult if not impossible, for others to benefit from their work. The purpose of this paper is to examine how these intended outcomes have been measured and to determine how we might improve and benefit from work in this area.

The Interdisciplinary Journal of Problem-based Learning • volume 3, no. 1 (Spring 2009)

59–89

Reviews of Research Investigating the Impact of PBL

Most meta-analyses comparing PBL and conventional approaches (lecture and discussion) to medical instruction have indicated that PBL students outperformed conventional students on the National Board of Medical Examiners (NBME) exam,[1] part II (Albanese & Mitchell, 1993; Kalaian, Mullan, & Kasim, 1999; Vernon & Blake, 1993), a multiple choice test of clinical knowledge taken at the end of the third year of medical school (Federation of State Medical Boards [FSMB] & NBME, 2005b). However, conventional students outperformed PBL students on NBME part I (Albanese & Mitchell; Dochy et al., 2003; Kalaian et al.; Vernon & Blake), a multiple choice test of basic science knowledge taken at the end of the second year (FSMB & NBME, 2005a). Other meta-analyses provided contrasting findings. For example, PBL students outperformed conventional students on authentic knowledge application tasks, for example, open-ended questions about problems (Dochy et al.; Smits, Verbeek, & de Buisjonjé, 2002), and on understanding principles that link concepts (Gijbels, Dochy, Van den Bossche, & Segers, 2005), but did not differ from conventional students on either concept or application levels (Gijbels et al.). Other research reviews indicated no significant differences in performance on similar outcomes (Colliver, 2000; Dochy et al.; Vernon & Blake).

According to Berkson (1993), inconsistent findings may arise because available measures "are insensitive, incapable of capturing important areas of competence in which problem-based students [i.e., students who engage in PBL] excel, e.g., problem solving and self-directed learning" (p. S84). Cronbach noted, "no matter how satisfactory it is in other respects, a test that measures the wrong thing or that is wrongly interpreted is worthless" (1970, p. 121). To establish interpretability, one must establish test scores' construct validity—the degree to which test scores indicate the amount of an unobservable trait (construct) a test taker has (Anastasi & Urbina, 1997)—for a specified purpose (Messick, 1989). In this paper the term, "test" refers to any "systematic procedure for observing a person's behavior and describing it with the aid of a numerical scale or a category-system" (Cronbach, 1970, p. 26).

Validity and Reliability

Essential to constructing quality instruments or measures is gathering the required score reliability and validity evidence to support the instrument's scores, purpose, use, and interpretation. A common misconception is that tests can be valid. To the contrary, only specific test scores can be valid (Cronbach, 1970; Messick, 1989). To be clear, score validity and reliability are not a dichotomy: test scores can have different levels of construct validity for different purposes (Messick, 1989). Many forms of evidence, including the breadth of relevant content coverage of test items, the relationship between test scores and scores on other established tests that purport to measure the same construct, and the correlation

between test scores and levels of future performance, contribute to a body of evidence to support the construct validity of test scores for a given purpose (Messick). Interested readers are directed to a special issue on validity issues in *Educational Researcher* (2007). Score reliability is the extent to which variance in scores of a given test is reflective of variance in the trait measured by the test (Anastasi & Urbina, 1997). Test scores cannot have construct validity if they are not first reliable, or consistent between test-taking sessions (i.e., test-retest), between test items (i.e., internal consistency), or between forms (i.e., parallel forms) (Anastasi & Urbina).

Many researchers naïvely use measures that have been used previously and appear in publications. This use assumes that the previous user paid careful attention to the quality of the measure. However, this assumption may not always be true, and poor measures can have a direct influence on the results. It is important to note that no measurement made, especially in the social and behavioral sciences, is free of error. Present in all measures is random error, which in turn, influences validity. Take the simple example of correlating two variables of interest (a bivariate correlation). The correlation (r_{xy}) of variable 1 (X) with variable 2 (Y) will be constrained by the reliability of the variables ($r_{xx,}$ r_{yy}). That is, $r_{xy} = r^*_{xy} \sqrt{r_{xx} r_{yy}}$. Notice that r_{xy} (the observed correlation) will only equal r^*_{xy} (the correlation between true scores) when the two measures have perfect reliability (1.0). As noted previously, all measures have error, so this correlation (i.e., r_{xy}) will be biased downward, the lower reliability on either measure becomes. That is, poor reliability for either or both measures of X and Y will attenuate or weaken the correlation between variables X and Y. This can lead to false conclusions that no relationship exists between two variables, when in fact it does, but cannot be observed due to poor reliability. See Nunnally and Bernstein (1994) for corrections for this issue.

Unfortunately, poor measurement quality can do more than merely attenuate correlations and in some cases, can even result in correlations in the opposite direction of the true relationships (Fleiss & Shrout, 1977). Other examples are provided by Cochran (1968). The major concern is that this may lead to incorrect theory based on such false results. Measurement issues, particularly error, can have real and serious influences on many aspects of the research process (e.g., design, statistical analysis). See Pedhazur and Schmelkin (1991) for a lengthy discussion on the topic.

Often results and conclusions are used to build theory that are based on analyses using a total score for an instrument with little, if any, information about the psychometric properties of that score. Without such knowledge of a score's properties (e.g., validity), it is unknown how statistical analyses are influenced by those properties. However, if information is provided on such issues (e.g., score reliability), one can gain a sense of how the analyses may have been influenced. Continued improvement of measurements should be one of the highest priorities of social and behavioral scientists (Pedhazur & Schmelkin, 1991). As Pedhazur and Schmelkin noted, issues of measurement do not get the neces-

sary and deserved attention in research publications and make it difficult to judge if the measures used meet standards set by the field (e.g., *Standards for educational and psychological testing* [American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999]).

Selecting appropriate instruments. There are a series of questions to ask when selecting an instrument (e.g., Rudner, 1994) such as (a) what is the intended use of the measure, (b) is the sample used to norm the measure representative of the sample with which I am working, (c) are reliability estimates sufficient for the intended use, (d) how aligned is the content with the content I intend to assess, (e) is the theoretical framework (e.g., how behavior is said to be predicted) clearly articulated, and (f) has the instrument been examined for bias or differential validity (i.e., validity that differs among diverse populations). A thorough review of an instrument requires careful examination of the answers to these types of questions.

Reporting. To allow others to review the instruments used in your research, certain types of information must be reported. Researchers must report (a) evidence supporting the degree of construct validity and reliability of test scores used in a research study (AERA, APA, & NCME, 1999; AERA, 2006), and (b) the rationale for how and why the instruments explain and predict the target outcomes (i.e., the theoretical framework; AERA) to support the construct validity of test score use.

Definition of PBL's Intended Learning Outcomes

Construct definition precedes construct measurement (Anastasi & Urbina, 1997). As many authors disagree on operational definitions of the intended learning outcomes of PBL (Albanese & Mitchell, 1993; Berkson, 1993; Colliver, 2000; Vernon & Blake, 1993), we present them here.

Deep Content Learning

PBL supporters argue that PBL students remember more content over longer periods of time (i.e., 1-2 years or more) than conventional students who studied the same content (Gallagher, 1997; Hmelo & Ferrari, 1997). However, it is unclear how researchers identify that participants have learned content deeply. For deep content learning to occur, students must connect the new content meaningfully with already learned content (Albanese & Mitchell, 1993), an idea long present in educational thought (Ausubel, 1963). However, measures that assess the existing knowledge to which new content is linked are not currently available.

PBL researchers have often portrayed deep content learning as the ability to understand and apply content to new situations (Gallagher, 1997; Gallagher & Stepien, 1996; Hmelo & Ferrari, 1997). Students who understand a concept deeply should be able to describe it in their own words, recognize relationships between it and other concepts, and

determine the implications of statements using it (Bloom, 1956). For example, in order for content to be useful to middle school students learning chemistry, students must be able to understand and apply the content in relevant situations. Thus, when evaluating if PBL leads to deep content learning, researchers should evaluate if PBL students understand and are able to apply unit content to real-life situations (e.g., use information learned about chemical reactions when determining the chemical properties of different substances).

Problem-solving Ability

Another intended learning outcome of PBL is increased problem-solving ability. A problem exists when there is a discrepancy between what is and what ought to be (Jonassen, 2003). Specifically, PBL is designed to increase students' abilities to solve ill-structured problems (Gallagher et al., 1992). Ill-structured problems "have many alternative solutions, vaguely defined or unclear goals and unstated constraints, and multiple criteria for evaluating solutions" (Jonassen, p. 21). To solve a PBL problem, students must be able to deconstruct the problem into its constituent parts (e.g., stakeholders, relationships among them, impacts of the problem on them), define the problem in their own words (Bodner, 1991; Glaser, Raghavan, & Baxter, 1992; Scandura, 1977; Schoenfeld, 1985; Smith, 1991), determine resources to help them understand the problem, (Schoenfeld), determine and pursue learning issues (Hmelo-Silver, 2004), and develop and test a solution (Hmelo-Silver). To assess problem-solving ability, it is important to assess students' abilities to successfully and effectively complete each step of the process of solving an ill-structured problem. While experts in the field may be able to generate quality solutions using a more heuristic approach (Schank & Abelson, 1977), novice problem solvers need to learn (and be assessed) on their abilities to complete each step in the problem-solving process. Although it is possible, it is also unlikely, that students will be able to develop and test an effective solution to a problem if they haven't first deconstructed the problem, defined it in their own words, determined the necessary resources, and identified and pursued relevant learning issues.

Self-Directed Learning

Self-directed learning is "any increase in knowledge, skill, accomplishment, or personal development that an individual selects and brings about by his or her own efforts using any method in any circumstances at any time" (Gibbons, 2002, p. 2). PBL was specifically designed to increase students' abilities to direct their own learning. This was based on the fact that medical students would be required to stay abreast of developments in medical research after graduation (Barrows & Tamblyn, 1980). Self-directed learning is essential during the PBL process because students need to determine what they do and do not know, and then design and follow a path to gain the knowledge they need in order to find a viable solution to the problem (Hmelo et al., 1997). Given that PBL is designed to promote self-directed learning both during the unit and afterwards, researchers can as-

sess two levels of self-directed learning: during and after the unit. A similar outcome often measured in PBL research is self-regulated learning, or students' abilities to set goals for and engage independently in learning activities (Pajares, 2002).

Purpose

As Cronbach (1970) noted, "if a program is trying to produce a certain change in behavior, to evaluate its effectiveness, the tester needs to observe just that type of behavior" (pp. 122-123). What types of behavior have been observed and measured in PBL research? We reviewed assessments used in PBL research to help readers understand the theoretical and measurement considerations that have guided development of existing measures in PBL research. Ultimately we hope this article will help PBL researchers select and design appropriate instruments for future research and write reports that convey essential validity evidence.

Method

Criteria for Inclusion

To be included in our review, empirical studies (investigating any level of education) needed to examine the impact of PBL on students' attainment of one or more of the three intended learning outcomes: deep content learning, increased problem-solving ability, or increased self-directed learning. We did not constrict our review to specific years of publication. We reviewed 33 studies, of which 30 were quantitative, 2 used mixed methods, and one was qualitative.

Procedure

We followed recommendations from Gall, Gall, and Borg (2003) for conducting literature reviews. First, we searched preliminary sources (PsychInfo, Education Full Text, and Educational Resources Information Center, and Academic Search Premier) to identify studies using the following search terms: *problem-based learning, higher-order thinking, problem-solving, content,* and *self-directed learning*. In a second search to locate additional articles, we added the terms *problem-solving measures*, *self-directed learning measures, higher-order outcomes, deep content*, *university, middle school, elementary school,* and *high school*. Subsequently, we used secondary sources and examined references cited in each study to identify additional studies. Third, we classified all research articles (n=33) according to what the author claimed to be measuring. We also examined measure descriptions and, when possible, authors' descriptions of the theoretical frameworks behind, and psychometric properties of, the measures. Finally, to synthesize the literature, we discussed major findings in the results section and compiled the following information in a table:

- Author, year of publication
- Study context (e.g., middle school science classroom)
- Research questions (either quoted, or, when not stated explicitly by the authors, as gathered from our reading of the article)
- Measure classification (e.g., multiple choice questions, case: think aloud while defining and solving problem) and description
- Sufficiency of the reported reliability and validity evidence for each measure

We used established standards (AERA, APA, & NCME, 1999) to judge the sufficiency of reported reliability and validity evidence. According to the standards, empirical research reports should contain four essential elements—(a) theoretical definitions of the assessed constructs, (b) theoretical rationales (e.g., construct validity evidence, other authors who have used the measure to assess the identified construct) for measure use, (c) measure description and procedures (e.g., scoring procedures, types of questions), and (d) reliability (e.g., internal consistency, interrater) of the test scores used in the study.

Results

Deep Content Learning

Types of Assessment. Of the 33 studies, five assessed the impact of PBL on deep content learning (see Table 1) using the following measure types: multiple choice questions (Aaron et al., 1998), self-report surveys about approaches to studying (Newble & Clarke, 1986), and depth of understanding of terms (Dods, 1997), essay questions (Antepohl & Herzig, 1999; Finch, 1999), list of terms to define (Dods), and presentation of a case after which the next steps needed to be suggested (Aaron et al.).

Validity and Reliability Information

Survey. Authors who used surveys to assess the impact of PBL on deep content learning (n=2) described their surveys in detail and gave references for where the surveys could be found, but did not provide validity or reliability evidence for their use of the surveys (Dods, 1997; Newble & Clarke, 1986). Due to the lack of validity evidence, readers are faced with important questions regarding why the specific surveys were chosen rather than interviews or direct observation. A simple sentence or two stating what the survey test scores represent and how they represent the construct would have been helpful to readers. Similarly, the rationale behind the use of a self-assessment of depth of understanding of terms was not clear (Dods). One might ask if high school students can accurately self-evaluate the degree to which they know a term. It would have been helpful, for example, if the author had mentioned other studies where this technique has been used and any concurrent validity evidence that may have been collected.

Table 1. *Studies on the Impact of PBL on Deep Content Learning.*

Study	Context	Objective or RQs	Test Classification & Description	Insufficient reporting * on	
				Reliability	Validity
Aaron et al. (1998)	Medical school with conventional and PBL courses	"To determine if change to a PBL method in a single course influences students' performance on…[recall and knowledge application] examinations compared to courses taught to the same students in a lecture-based format" (p. 86).	*Multiple choice questions* Measured factual recall	X	X
			Case + Suggest next steps Given cases, students suggested the next "history-taking, physical examination, or management" steps (p. 87-88)	X	X
Antepohl & Herzig (1999)	Medical school with conventional and PBL tracks	"To find out whether such a change [from conventional to PBL curriculum] would be possible without disadvantages for the students' factual knowledge" (p. 107-108)	*Multiple choice questions* Measured factual recall.	X	X
			Essay questions Measured comprehension and application of concepts.	X	X
Dods (1997)	Gifted high school science	To compare "the efficacies of problem-based learning…[and] traditional lecture…to promote understanding and retention of the principle content of an elective science course" (p. 423)	*Self-report questionnaire & Short answer* Students rated depth of understanding of, and provided definitions of each term on a list. Then, researchers assigned (using a rubric) depth of understanding scores to definitions.	X	
Finch (1999)	Podiatric medicine school	"The goal of the study was to determine the effect of PBL on the academic performance of the students" (p. 413)	*Multiple choice questions* About "clinically relevant biomedical knowledge" (p. 413)	X	X
			Essay questions On "deeper understanding and the cognitive skills related to patient management" (p. 413)	X	X
Newble & Clarke (1986)	PBL and conventional medical schools	Are there differences in the learning approaches (deep or surface) between PBL and conventional medical students?	*Self-report questionnaire* Survey about approaches to studying	X	

* No study included content, criterion-related, or predictive validity evidence. To save space, this lack is not noted individually for each study.

Combination of multiple choice and essay questions. Four studies were found in which a combination of multiple choice and essay questions were used to assess the impact of PBL on deep content learning (Aaron et al., 1998; Antepohl & Herzig, 1999; Dods, 1997; Finch, 1999). Of these four, three explained their choice of measure type while one did not (Aaron et al.). One explained how the questions were scored, while three did not (Aaron et al.; Antepohl & Herzig; Finch). Three described the questions, while one did not (Antepohl & Herzig). Two gave validity information, while two did not (Antepohl & Herzig; Finch). One gave reliability information, while three did not (Antepohl & Herzig; Dods; Finch). This is problematic because any numerical score without this information is just a number. When no explanation for choice of measure type is given, readers cannot know why the measure used was appropriate. With no explanation of scoring, readers cannot know what scorers were looking for in responses. Without a description of questions, readers cannot really know if deep content learning or rote memorization, for example,

was being assessed. Readers cannot assess the accuracy of scores or the appropriateness of statistical calculations without an estimate of reliability of scores.

Problem-solving Ability

Types of Assessment. Of the 33 studies, 23 assessed the impact of PBL on problem-solving ability (see Table 2). In 18 studies, participants were presented with cases or simulated patients, after which they performed tasks including: answered questions about the problem (Arts, Gijselaers, & Segers, 2002; Goodman et al., 1991; Moore et al., 1994), outlined problem solution paths (Gallagher et al., 1992), examined the simulated patient and provided a diagnosis (Distlehorst & Robbs, 1998; Heale et al., 1998; Moore et al., 1990; Sanci et al., 2000; Schwartz & Burgett, 1997; Schmidt et al., 1996), wrote problem definitions (Barrows & Tamblyn, 1980), answered multiple choice questions about next steps (Zumbach, Kumpf, & Koch, 2004), generated learning issues (Pedersen & Liu, 2002-2003), engaged in a think aloud while solving the problem (Boshuizen, Schmidt, & Wassner, 1993; Segers, 1997), and defined the problem and generated learning issues (Hmelo, 1998). Given key features influencing the solution and a case, participants solved the problem (Doucet, Purdy, Kaufman, & Langille, 1998) or answered questions about the problem (Schuwirth et al., 1999). Other measures included clinical ratings (Distlehorst & Robbs; Lewis & Tamblyn, 1987; Moore et al., 1994; Richards et al., 1996; Santos-Gomez et al., 1990), project ratings (Lee & Kim, 2005), honors or remedial selection (Distlehorst & Robbs), and essay questions (Schwartz & Burgett).

Validity and Reliability Information

Case: solve problem. Six studies used performance testing, in which students were required to gather all the information required to solve the presented problem or diagnose a simulated patient's "ailment" (Distlehorst & Robbs, 1998; Heale et al, 1988; Moore et al., 1990; Moore et al., 1994; Sanci et al., 2000; Schwartz & Burgett, 1997). While readers of the forums in which these studies were published may not have expected an explicit rationale for the use of performance testing, greater detail about the scoring procedures could have allowed for the interpretability of results (AERA et al., 1999). Scoring procedures were described in one study (Heale et al.), but were not described in five studies (Distlehorst & Robbs; Moore et al., 1994; Moore et al., 1990; Sanci et al.; Schwartz & Burgett).

In two studies, students engaged in a think-aloud as they solved a problem (Boshuizen et al., 1993; Segers, 1997). No reliability information was presented in either study. While the theoretical framework was clear in Segers, the case presentation was not. The opposite was true for the study by Boshuizen et al. Thus, neither study reported a complete set of reliability information. Think-aloud protocols, in which participants are told to think aloud while solving a problem, can also be problematic, as Gilhooly (1990) noted, because people do not habitually say all they are thinking as they are thinking it, or do so when asked.

Table 2. *Studies on the Impact of PBL on Problem-solving Ability.*

Study	Context	Objective or RQs	Test Classification & Description	Insufficient reporting * on	
				Reliability	Validity
Arts et al. (2002)	Business school	"Does…[incorporating a database with authentic company artifacts (e.g., annual reports) to augment the case descriptions used in the described business school's PBL units] lead to a better application of knowledge in new and authentic problem-solving situations?" (p. 482)	*Case + Answer essay questions* Students were presented with ill-structured cases and had to answer essay questions		X
Barrows & Tamblyn (1976)	Medical school	Is there a difference in problem definition between conventional and PBL students?	*Simulated patients + Write problem definition* 2 simulated patient problems were presented to conventional and PBL students; students had to (1) define each problem, (2) answer a multiple choice test over content relevant to the problem. Then they could (1) study a database with relevant information, and (2) revise their problem definition or answers to the test questions.	X	X
Boshuizen et al. (1993)	PBL and conventional medical schools	"students in a problem-based medical school would show more indications of knowledge integration than students in a traditional school" (p. 35)	*Case + Think-aloud while defining and solving problem* Participants were taped as they "thought aloud" when solving a case	X	X
Distlehorst & Robbs (1998)	Medical school with conventional and PBL tracks	"To report on…performance outcomes [content knowledge and clinical performance] for those students at Southern Illinois University…School of Medicine who participated in the two curricular tracks during the first two years that the dual curricula were offered" (p. 131)	*Clinical rating* Performance ratings on third year clinical clerkships	X	X
			Simulated patients + Generate and pursue Learning Issues + Solve problem Simulated patients were presented to students, who were then required to examine the patient and make a final diagnosis	X	X
			Honors and remedial selection Number of students from each track given honors or recommended for remedial study	X	X
Doucet et al. (1998)	Continuing medical education	"Compared with a didactic approach, does family physician participation in a continuing medical education programme using…PBL… positively affect clinical reasoning skills?" (p. 591)	*Multiple brief cases & key features + Answer questions* Given cases and key features of the problem, students answered questions	X	X

* No study included content, criterion-related, predictive, or content validity evidence. To save space, this lack is not noted individually for each study.

Table 2 (cont.). *Studies on the Impact of PBL on Problem-solving Ability.*

Study	Context	Objective or RQs	Test Classification & Description	Insufficient reporting on	
				Reliability	Validity
Gallagher et al. (1992)	Gifted high school	"Significant improvement in problem-solving schemes would be observed in students in the experimental class; and… improvements observed in the experimental group would not be observed in the control group" (p. 196)	*Case + Outline a path to solution* Given a case, students outlined steps to problem solution. 1 point was assigned for each specific step prescribed by the authors	X	X
Goodman et al. (1991)	Medical school with conventional and PBL tracks	Are there differences in content learning and problem-solving ability gain between PBL and conventional medial students?	*Case or simulated patient + Answer questions* Students were given an oral exam in which they had to either read a case, or examine an SP, and then answer questions.	X	X
Heale et al. (1988)	Continuing medical education course	Is there a difference between conventional and PBL-based continuing medical education in diagnostic performance of physicians?	*Simulated patient + Diagnosis* 3 simulated patients "presenting" symptoms of the 3 illnesses covered in the course were sent unannounced to the students' practices at 3 and 7 months after the course. The SPs used a checklist to evaluate the physicians' performances.	X	
Hmelo (1998)	Two medical schools	"to examine the changes in …[clinical reasoning and learning strategies] that develop over the 1st year of medical school …and to compare the effect of different curricula (PBL and traditional) on this development" (p. 178)	*Case + Write problem definition and LIs* Given cases, students had to write out problem definition in terms of pathophysiological processes, and further information needed to solve the problem. Protocols were coded on criteria such as accuracy of diagnostic hypotheses and number of relational operators (such as because) in the explanations.		X
Lee & Kim (2005)	Online educational technology course	"Are there significant differences in problem solving *outcomes* when the CRST [collaborative representation supporting tool] is given in web-based collaborative PBL, in comparison to web bulletin boards without the CRST?" (p. 275)	*Project rating* Students created a project in which they designed "instruction based on their chosen ISD [instructional systems design] model" (p. 288). The projects "were assessed on the basis of (a) inquiry activities, (b) the qualities of outcome, (c) the degree of collaboration, and (d) creativity" (p. 288)	X	X

Table 2 (cont.). *Studies on the Impact of PBL on Problem-solving Ability.*

Study	Context	Objective or RQs	Test Classification & Description	Insufficient reporting on	
				Reliability	Validity
Lewis & Tamblyn (1987)	Class in nursing school with PBL and lecture sections	"There would be no difference between the experimental and the control groups in relation to their improvements in problem solving ability either a) overall or b) specific sub-skills (i.e., assess, plan, implement, and evaluate) as demonstrated in the clinical practice (hospital) setting" (p. 17)	*Clinical rating* Clinical instructors observed nursing students in hospital setting, and recorded observations on a standardized checklist	X	X
Moore et al. (1994)	Medical school with conventional and PBL tracks	"NP students [in the PBL track], when compared with the controls in their first two years, would…know better how to approach and solve problems" (p. 984)	*Case + Answer questions* Students were asked questions about cases	X	X
			Other tests Two different computer simulations of cases Another ill-described test	X	X
Moore et al. (1990)	Medical school with conventional and PBL tracks	"to determine the differences in their response to their educational experiences and to assess whether the two curricula were associated with any differences in outcomes" (p. 2)	*Simulated patients + Generate and pursue learning issues + Solve problem* During medical clerkship, students encountered simulated patients and had to diagnose the patient problem.	X	X
			Clinical rating (content analysis) Reviewer blind to student condition performed content analysis of clinical clerkship records.	X	X
Pedersen & Liu (2002-2003)	6th grade science class	"Do students transfer problem-solving strategies that are modeled for them during PBL [ask questions, pursue answers to questions, and write up solution rationale] to their work on a similar problem on an unrelated topic?" (p. 306)	*Case + Generate LIs* Students were presented with a written problem, and they had to write the kind of questions they would ask to help them gain information to solve the problem.		X
Richards et al. (1996)	Medical school with conventional and PBL tracks	"To compare clinical performances in a third-year internal medicine clerkship between students from a problem-based learning…curriculum and students from a … lecture-based… curriculum" (p. 187)	*Clinical rating* Clinical clerkship ratings from "approximately ten different interns, upper-level residents, and faculty" on the basis of "four scales: dealing with knowledge, history and physical examination, differential diagnosis, and communication" (p. 188)	X	X

Table 2 (cont.). *Studies on the Impact of PBL on Problem-solving Ability.*

Study	Context	Objective or RQs	Test Classification & Description	Insufficient reporting on	
				Reliability	Validity
Sanci et al. (2000)	Continuing medical education	What is the effect of the use of the PBL format in a continuing medical education course on the clinical skills of physicians?	*Simulated patient + Examination and diagnosis of patient problem* Physicians examined SPs 7 and 13 months after the intervention; the SPs evaluated the physician with a checklist and the examination was videotaped; independent observers rated the taped examinations.	X	X
Santos-Gomez et al. (1990)	Medical school with conventional and PBL tracks	To evaluate "the residency performance of New Mexico's parallel curricular tracks by analyzing three performance assessments representing different perspectives within the health care team" (p. 367).	*Clinical rating* Nurses and supervising physicians used a Likert-type checklist to rate the critical thinking ability and knowledge of residents who graduated from each track. Residents rated themselves using the same instrument.	X	X
Schmidt et al. (1996)	3 medical schools	"In the present study, we compared the diagnostic performances of 612 students from three Dutch medical schools: a problem-based school; a school with a systems-based but teacher-driven curriculum; and a school with a…lecture-based curriculum" (p. 660)	*Multiple cases + Diagnose patient problem* Students were presented with medical cases, and had to provide a diagnosis. They were provided 2 points if the diagnosis was correct, or 1 point if the student identified the right organ.		X
Schuwirth et al. (1999)	PBL and conventional medical schools	"it was expected to find differences between those students following an integrated curriculum and those following the more traditional curriculum" (p. 236)	*Multiple brief cases + Given key features to problem, answer questions* Students were presented with many brief medical cases, and then answered multiple choice and short answer questions	X	X
Schwartz & Burgett (1997)	Surgery clerkship with PBL and conventional tracks	Are there "differences in student achievement for those in a problem-based curriculum compared with those in a more traditional one"?	*Essay questions* Modified essay examination	X	X
			Standardized patient + Generate and pursue LIs + Solve problem Simulated patients were presented to students, who were then required to examine the patient and make a final diagnosis	X	X

Table 2 (cont.). *Studies on the Impact of PBL on Problem-solving Ability.*

Study	Context	Objective or RQs	Test Classification & Description	Insufficient reporting on	
				Reliability	Validity
Segers (1997)	Business school	"To what extent do the test scores [on the OverAll test] reflect students' ability to analyze and solve economics problems?" (p. 388)	*Case + Think-aloud while defining and solving problem* Given a case, students who scored in the top 27% on the OverAll test "thought aloud" while analyzing and then wrote a solution to the problem. Protocols analyzed for number of correct concepts, appropriate links between concepts, and correct solution.	X	X
Zumbach et al. (2004)	Elementary school	Compared the impacts on problem-solving ability of a conventional class and a PBL class designed to teach students about badgers	*Multiple brief cases + Answer multiple choice questions on next steps* Students read scenarios and then answered multiple choice questions about what to do next and "their certainty in answering each question" (p. 32).	X	X

Cases: key features. Two authors used the key features case approach (Doucet et al., 1998; Schuwirth et al., 1999). A central premise for this approach is that "the process by which physicians resolve clinical problems on paper mirrors their response when presented with the same clinical cases in practice" (Doucet et al., p. 591). However, this premise is problematic because, according to Doucet et al., it is not supported by evidence. Additionally, when physicians encounter clinical cases, someone else does not identify a specific key feature that most impacts the solution before the patient examination. One author attempted to justify the focus on providing treatment by stating, "in some medical cases . . . the actual diagnosis may not be the key element, but treatment or management may be more significant" (Schuwirth et al., p. 236). However, diagnosing patient problems may be as important to physician problem solving as providing treatment.

Other cases. None of the authors provided a rationale for using other case-related measures to assess the impact of PBL on problem-solving ability (Goodman et al., 1991; Moore et al., 1994; Schmidt et al., 1996; Zumbach et al., 2004). The finding in one study of no significant differences between the treatment and control groups was attributed to problems with the measure (Zumbach et al.), a likely reason given issues of measurement and instructional sensitivity in many areas of research, especially education (W. Popham, personal communication, November 3, 2007). An implicit rationale for the use of measures

of problem-solving ability in two studies could be that students' problem-solving abilities could be explained by giving them opportunities to investigate a problem, and then measuring what they learned from the investigation (Goodman et al., 1991; Moore et al., 1994). Measures of learning, of course, are critical to determining outcomes related to deep content learning, yet are not related directly to increases in students' problem solving abilities, which is how they were used in these studies. Furthermore, while measuring one part of the problem-solving process—generating and pursuing learning issues—may be a good place to begin, researchers also should consider measuring students' ability to complete the other steps of the problem-solving process, especially if the goal is to measure changes in students' problem solving skills.

Other measures included presenting cases and having students (a) outline how they would solve the problem (Gallagher et al., 1992), (b) answer essay questions (Arts et al., 2002), or (c) write problem definitions and learning issues (Hmelo, 1998) or just learning issues (Pedersen & Liu, 2002-2003). However, several questions remain. First, if students perform well outlining how they would solve the problem, are they demonstrating problem-solving skills or recall of the steps that they were encouraged to use (Gallagher et al.)? It also is not clear how essay questions measured the application of content knowledge in authentic problem-solving situations (Arts et al.). Additional information about the types of questions used would help the reader understand how they might be appropriate for this purpose. As noted earlier, it is not clear if defining a problem and generating but not pursuing learning issues is equivalent to solving a problem or if the number of relational operators measure coherence of an argument (Hmelo). After all, a pair in a leaky boat can define that their boat is leaking and generate a list of symptoms they need to examine and research, but until they actually examine and research the symptoms and fix the leak the problem is still there. While it is possible that all of these measures were appropriate for the researchers' purposes, without additional information, it is impossible to tell.

Clinical ratings. Clinical clerkships typically occur in the last two years of medical and nursing school, and provide opportunities for students to examine patients under supervision. As one measure of the impact of PBL on problem-solving ability, comparisons have been made between the supervisor-assigned clerkship ratings of medical and nursing students enrolled in PBL and traditional tracks (Distlehorst & Robbs, 1998; Lewis & Tamblyn, 1987; Richards et al., 1996). However, no evidence was provided that the ratings assessed problem-solving ability or even clinical performance (Distlehorst & Robbs; Lewis & Tamblyn; Richards et al.), making the conclusion problematic that "pre-clinical PBL curricula as found at the Bowman Gray School of Medicine may enhance third-year students' clinical performance" (Richards et al., p. 187). In addition, clinical clerkship GPA was found to only explain 7.8% of the variance in residency performance ratings (Hamdy et al., 2006).

Some authors noted that numerical clerkship ratings can suffer from the halo effect—that is, supervisors often give high ratings to all or most clerkship students (Cacamese, Elnicki, & Speer, 2007; Moore et al, 1990; Santos-Gomez et al., 1990). Due to this potential problem, Moore et al. used a content analysis of the clerkship ratings to compare PBL and conventional students. However, others (Santos-Gomez et al.) have continued to compare the numerical residency ratings given by supervising nurses and doctors to graduates of the PBL and conventional tracks of a medical school. When ratings were used, authors did not give explicit criteria for how the ratings were calculated (Moore et al.; Santos-Gomez et al.).

Project ratings. Another measure used to assess the impact of PBL on problem-solving ability was ratings on students' final projects in an educational technology course (Lee & Kim, 2005). Unfortunately, project assessment procedures were not explained (Lee & Kim). Though they noted that outcomes "were assessed on the basis of (a) inquiry activities, (b) the qualities of outcome, (c) the degree of collaboration, and (d) creativity" (p. 288), it is unclear exactly how such assessment occurred.

Essay questions. Essay questions were used in one study to measure problem-solving ability (Schwartz & Burgett, 1997). However, the questions were not described, making it difficult to evaluate the fit of the measure. In addition, the scoring procedures were not clear.

Honors and remedial selection. The impact of PBL versus conventional tracks on problem-solving ability was compared using percentages of students from PBL and conventional tracks whom faculty selected for honors or remedial instruction (Distlehorst & Robbs, 1998). A possible, but implicit, premise behind the use of such figures could be a perception that students who were better at problem solving would be selected for honors by the faculty, and those who were poorer would be selected for remediation (Distlehorst & Robbs). However, this is problematic for three reasons. First, just as grade inflation has been observed in higher education in general, it has also happened in medical schools (Cacamese et al., 2007). In a survey of medical schools, Cacamese et al. found that almost half of clinical clerkship students received honors, and over three fourths received high grades. Second, according to Hamdy et al. (2006), clerkship GPAs and Dean's letter rankings (rankings of medical students included in letters sent to residency locations) of medical school students have relatively small correlations (0.28 and 0.22, respectively) with supervisor ratings during residency. Thus, clerkship GPA and Dean's letters only explained 4.8 to 7.8% of the variance in supervisor ratings during residency. Finally, medical schools vary in their approaches to remediation. Some encourage students to take remedial courses and rarely have students drop out, while others tend to let students drop out who are having difficulty (Hughes, 2002).

Self-directed Learning

Types of Assessment. Of the 33 studies, seven assessed the impact of PBL on self-directed learning (see Table 3). Measures included self-report questionnaires (Blumberg & Michael, 1992; Chanlin & Chan, 2004; Lohman & Finkelstein, 2000), interviews (Chanlin & Chan; Evensen, Salisbury-Glennon, & Glenn, 2001), student reflections (Chanlin & Chan; Evensen et al.), scores on NBME I and II (Kaufman et al., 1989), clerkship ratings (Kaufman et al.), and library circulation data (Blumberg & Michael). In other studies that used case presentations, participants were required to define the problem, identify information they needed to know (learning issues), and either (a) determine how to address the learning issues (Hmelo et al., 1997) or (b) determine how to address learning issues, address learning issues, and define the problem (Barrows & Tamblyn, 1976).

Validity and Reliability Information

Self-report questionnaires. The strategy of asking students about the frequency of using various library resources appears to be reasonable (Blumberg & Michael, 1992). However, neither reliability information nor a clear rationale for how the measure assessed self-directed learning was presented (Blumberg & Michael). Given a definition of self-directed learning as "recognizing the need for new learning, setting one's own learning objectives, defining relevant questions for study, accessing relevant information, testing one's depth of understanding of what one has learned" (Blumberg & Michael, p. 3), asking students how often they used different library resources but not how or why they used the resources does not appear to be sufficient. Other authors provided limited or questionable validity evidence for the test scores (Chanlin & Chan, 2004; Lohman & Finkelstein, 2000). One noted that experts examined a questionnaire for construct validity (Chanlin & Chan). While experts can be involved in the process of assessing the construct validity, they can only perform part of the process: list what constructs might account for performance on the measure (Anastasi & Urbina, 1997; Cronbach, 1970; Kerlinger & Lee, 2000). Subsequently, empirical tests (e.g., factor analysis, multitrait-multimethod) must provide evidence that the suggested constructs do in fact account for test performance (Anastasi & Urbina; Cronbach; Kerlinger & Lee). Though not certain, it seems likely that the experts in Chanlin and Chan's study examined the questionnaire for content validity, or the extent to which it includes questions representative of the behavior domain. Content validity contributes to evidence of but is not the same as construct validity. It also is not appropriate to provide a citation for a measure and note that its "validity and reliability have been extensively documented" (Lohman & Finkelstein, p. 299), as test scores, and not tests, can be reliable, and validity refers to the goodness of fit of test scores to current research or assessment purposes (Messick, 1989).

Table 3. *Studies on the Impact of PBL on Self-Directed Learning.*

Study	Context	Objective or RQs	Test Classification & Description	Insufficient reporting* on	
				Reliability	Validity
Barrows & Tamblyn (1976)	Medical school	Is there a difference in problem definition and self-study skills between conventional and PBL students?	*SPs + Answer Questions + Write problem definition + Study + Rewrite Definition* Given SPs, students (1) defined problem, (2) took multiple choice test over relevant content, (3) studied a database with relevant information, and (4) revised problem definition or answers. Change between initial and final scores indicated self-directed learning.	X	X
Blumberg & Michael (1992)	Medical school with conventional and PBL tracks	"do students in a mixed PBL curriculum exhibit more…[SDL] than do students from the same school in a regular, traditional curriculum…that is largely teacher-directed?" (p. 4)	*Self-report questionnaire* On how often students used different types of library resources.	X	
			Library circulation data Number of books checked out to each student per semester.	X	X
Chanlin & Chan (2004)	Online course on dietetics	"To what extent can PBL foster a positive learning response in terms of…use of web resources, use of online-discussion, and students' involvement in the learning task?" (p. 438)	*Grade* Points awarded on students' final projects		
			Artifact The final project	X	X
			Self-report questionnaire & Interviews & Student reflections	X	X
Evensen et al. (2001)	Medical school	"in the present study we sought to investigate how and to what degree medical students in a PBL context self-regulated their learning" (p. 661)	*Student reflections* Students kept oral diaries about "what and how they were planning to study, how they actually proceeded, and how useful they found particular strategies and resources" (p. 662)		X
			Interviews & Observations Students observed during PBL activities and interviewed every other week.		
Hmelo et al. (1997)	Medical school: PBL elective, another elective	"to assess measures of specific cognitive changes purported to be affected by the PBL curriculum [in the integration of clinical and basic sciences, developing clinical reasoning, and becoming lifelong learners] " (p. 388)	*Case + Describe problem + Identify learning issues & way to address learning issues* Given case, students had to explain causal mechanisms, what they needed learn to make a full diagnosis, and how they would learn it. Students were deemed more self-directed if they wrote they would look up information in basic science texts.		X

* No study included content, criterion-related, predictive, or content validity evidence. To save space, this lack is not noted individually for each study.

Table 3 (cont.). *Studies on the Impact of PBL on Self-Directed Learning.*

Study	Context	Objective or RQs	Test Classification & Description	Insufficient reporting on	
				Reliability	Validity
Kaufman et al. (1989)	Medical school with conventional and PBL tracks	Given the goal of equipping "graduates with skills in self-directed, lifelong learning…how well has the experimental curriculum achieved its goals?" (p. 286)	*Clinical ratings* Clinical ratings during clerkships	X	X
			Scores on NBME Part I and II	X	X
Lohman & Finkelstein (2000)	Dental school	Will students "in small, medium, and large PBL groups develop significantly different levels of self-directedness?" (p. 295)	*Self-report questionnaire* Self-directed learning readiness scale	X	X

Interviews. When interviews were used (Chanlin & Chan, 2004; Evensen et al., 2001), the content of interview questions was not explained, leaving it difficult to determine how well they measured self-directed learning. Due to the paucity of information about such data sources, we could not assess the validity of, or the theoretical considerations behind, the interview measures.

Student reflections. The authors of one study gave a reasonable description of the content of entries in the oral learning logs students completed in their study (Evensen et al., 2001). Evensen et al. were investigating the outcome of self-regulated learning; as described earlier, self-regulated learning is a similar outcome to self-directed learning. Though they did not explicitly articulate their framework for self-regulated learning, we can deduce that they believed that self-regulated learning can be explained, at least in part, through students' self-reporting as they identify and address learning issues. The authors gave a sufficient account of how transcripts were coded and accuracy ensured through member checking.

Clinical clerkship ratings and scores on NBME II. It was not clear how clinical clerkship ratings and scores on the NBME II related to self-directed learning (Kaufman et al., 1989). Perhaps because the PBL students performed worse than conventional students on NBME I but better on NBME II and in their clerkships, the authors may have accepted this as evidence that PBL students engaged in self-directed learning between NBME I and II to address gaps in their knowledge. However, this was not stated (Kaufman et al.). In addition, clinical clerkship grades may be inflated, making it difficult to use as evidence to discriminate between students (Cacamese et al., 2007).

Library circulation data. A clear rationale for the use of library circulation data to compare the self-directed learning skills of PBL and conventional students was not provided

(Blumberg & Michael, 1992). Potential problems with this measure include (a) that PBL schools give specific workshops on how to use the library more often than conventional medical schools (Woodward, 1996), and (b) students in the PBL curricula presumably need to check out books more often to address learning issues during the preclinical years. During the clinical years perhaps students go to the library more often because they do not have textbooks from earlier courses to which to refer (textbooks and other books counted the same in the authors' measure).

Present problem and have students identify learning issues. An assumption behind the use of this measure type was that the nature of learning issues that students generate (disease-driven, data-driven, or basic science), and the type of resources they use (clinical text or basic science book) to address learning issues indicate how self-directed they are (Hmelo et al., 1997). But this assumption generates a few questions. For example, are students who research in clinical texts less self-directed than those who use basic science texts? Researching issues in basic science texts may be indicative of data-driven reasoning, which has been associated with expertise (Hmelo et al.). However, expertise and self-directed learning are different constructs.

Another assumption is that self-directed learning can be measured by examining the difference between scores on students' definitions of a problem and answers to a test on content before and after an extended study period (Barrows & Tamblyn, 1976). However, one may ask if this is assessing students' self-directed learning skills, or their researching skills. Also, though Barrows and Tamblyn assigned numerical scores to the problem definitions, they did not mention their criteria for scoring the problem definitions.

Summary of Included Validity and Reliability Information

Of the 33 reports reviewed, only four gave interpretable reliability and dependability coefficients (e.g., interrater reliability of 92% [Hmelo et al., 1997]) for all measures (Hmelo, 1998; Hmelo et al.; Lohman & Finkelstein, 2000; Pedersen & Liu, 2002-2003). Two gave coefficients for some but not all measures (Chanlin & Chan, 2004; Sanci et al., 2000). Three reports gave incomplete or uninterpretable reliability evidence (Aaron et al., 1998; Lee & Kim, 2005; Santos-Gomez et al., 1990). No report contained content or criterion-related (concurrent or predictive) score validity evidence (other evidence attesting to test scores' construct validity).

Discussion

Validity is "the degree to which evidence and theory support the interpretations of test scores" (AERA et al., 1999, p. 9; Embretson, 2007). A recurring problem was that the constructs under examination often were not defined. Of the 33 studies, only 3 gave a complete theoretical rationale for test score use (Doucet et al., 1998; Hmelo, 1998; Hmelo et al., 1997).

Like the problems used in PBL, the desired outcomes of PBL (increased self-directed learning, deep content learning, and increased problem-solving ability) are ill defined (Berkson, 1993; Neufeld, 1989; Scandura, 1977; Vernon & Blake, 1993; Woodward, 1996). Because the outcomes are ill defined, PBL researchers should ensure that their definitions of constructs are clear to readers. We urge writers to provide all necessary information so that readers can determine the potential applicability of the conclusions to new contexts (AERA, 2006; AERA et al.). Without a clear explanation of the theoretical frameworks that authors use to explain and predict the target outcomes, readers cannot evaluate the validity of test score uses.

As test scores cannot be valid for all purposes, it is insufficient to state that the measures used in the current study "had been developed and validated elsewhere" (Moore et al., 1994, p. 984), especially when no evidence is given that the context (population, purpose, etc.) of measure use in the current study is highly similar to the recommended use of the test scores. Authors should build a rationale for the validity of their test scores so that readers can make their own judgments. The rationale should include information about the constructs the test purports to measure, along with empirical data to support that the test measures the given constructs the researchers are studying with their particular sample populations.

Test scores cannot be valid unless they are reliable (Anastasi & Urbina, 1997). Of the 33 empirical papers we included in this review, only eight gave appropriate evidence of the reliability and dependability of test scores (Arts et al., 2002; Boshuizen et al., 1993; Doucet et al., 1998; Evensen et al., 2001; Hmelo, 1998; Hmelo et al., 1997; Pedersen & Liu, 2002-2003; Schmidt et al., 1996). Four gave partial accounts of test score reliability (Aaron et al., 1998; Chanlin & Chan, 2004; Sanci et al., 2000). Readers cannot assess score validity if they do not know the measure's accuracy, and thus are unable to estimate the standard error of the measurement (in the case of quantitative research), or the extent to which different researchers provided with the same data would come to the same conclusions (in the case of qualitative research).

To allow readers to assess the validity of test scores, authors also must clearly describe test procedures—how they were administered, what students did, and how their responses were scored (Messick, 1989). For example, if the measure involves cases, descriptive information about the cases, and what students had to do after reading the cases, should be included in the research report. Many authors did not give sufficient information about (a) how measures were administered (Boshuizen et al., 1993; Chanlin & Chan, 2004; Moore et al., 1994; Schuwirth et al., 1999), (b) the content of questions (Antepohl & Herzig, 1999; Doucet et al., 1998; Evensen et al., 2001; Finch, 1999; Schuwirth et al., 1999; Schwartz & Burgett, 1997; Zumbach et al., 2004), or (c) scoring procedures (Aaron et al., 1998; Barrows & Tamblyn, 1976; Chanlin & Chan; Doucet et al.; Goodman et al., 1991; Lee & Kim, 2005; Moore et al.; Sanci et al., 2000; Schuwirth et al.).

Future Directions of PBL Research

While several measures have been used in PBL research to assess intended learning outcomes, inconsistent information has been reported about those measures. Based on our review, many authors did not give sufficient information about how measures were (a) selected, (b) administered or (c) scored. The solution is not to search for perfect measures of problem-solving ability, deep content learning, and self-directed learning, as validity pertains to test score use, not tests (AERA et al., 1999). Rather, the solution is to report better on the selection, use, and psychometric properties of measures. Such information should lead to researchers realizing the shortcomings of measures and seeing the need to improve these measures for future use.

Shortcomings in PBL measurement reporting are not unique among social sciences research (Hamdy et al., 2006; Hogan & Agnello, 2004). For example, among 38 medical education papers attempting to correlate various measures taken during medical school and residency performance, only one reported the reliability of both predictor and outcome variables (Hamdy et al.). Only 55% of articles from a variety of leading education and psychology journals contained any validity evidence (Hogan & Agnello). Gaps in measurement reporting can happen due to journal length requirements, as when reviewers ask authors to add non-measurement information, but to keep a manuscript within the page limit (Hogan & Agnello).

So why is the lack of appropriate measurement reporting important? Simply stated, better measurement reporting is needed to move PBL research forward. A fundamental purpose of educational research is to improve educational practice. If PBL does, in fact, lead to increases in self-directed learning, problem-solving ability, and deep content learning, it should be more widely used, especially in K-12 schools where students need to develop stronger problem-solving (Hulse, 2006; Jonassen, 2003; Warner, 2004) and self-directed learning skills (Hulse). Resources such as the *Doing What Works* website (US Department of Education, n.d., a) exist to help teachers learn about these types of educational approaches. However, at present, if one were to search for strategies for increasing problem-solving ability among K-12 students, one would be advised to design coaching and mentoring programs. PBL's absence on the *Doing What Works* site could lead to less dissemination of PBL among K-12 teachers.

Doing What Works selects interventions based on studies that have demonstrated satisfactory research evidence (US Department of Education, n.d., b). That is, the *What Works Clearinghouse* employs a three-stage process to determine if a study was conducted appropriately for providing sound evidence. One such criterion is the quality of the outcome measure. If the measure is judged inadequate, the entire study is listed as "*Does Not Meet Evidence Screens"* and is eliminated at stage one of the review process. As standards for high quality research increase in the data-driven environment, it is paramount that all

instruments used to assess outcomes are (a) aligned with the goals of the study, (b) have appropriate psychometric evidence to support the intended use, and (c) are understood by the researchers employing the instruments. To our knowledge, no study reviewed in this paper would meet evidence screens. Thus, given the studies identified for this review, the likelihood of PBL being promoted by the *Doing What Works* website to increase deep content learning and problem-solving and self-directed learning abilities appears slight.

However, PBL researchers can increase the likelihood of their research being used by the *What Works Clearinghouse* and the *Doing What Works* site by following a simple process (see Figure 1) in the preparation, implementation, and reporting of their studies. To begin, researchers must define their study goal. Within the goal statement there will be a construct (e.g., problem-solving ability) for which they must provide an operational definition. Following this, they need to identify measures that purport to measure the same construct and select from among those measures that include appropriate reliability and validity evidence from past uses with similar populations to the researchers' target populations. They must then administer the measures, collect appropriate validity and reliability evidence, and report that evidence, along with past reliability and validity evidence associated with the measure, the study goal and operational definitions of the constructs being measured, and measure description and procedures. While we recognize that not every form of reliability and validity can be collected for every study, efforts should be made to collect as much reliability and validity evidence as possible.

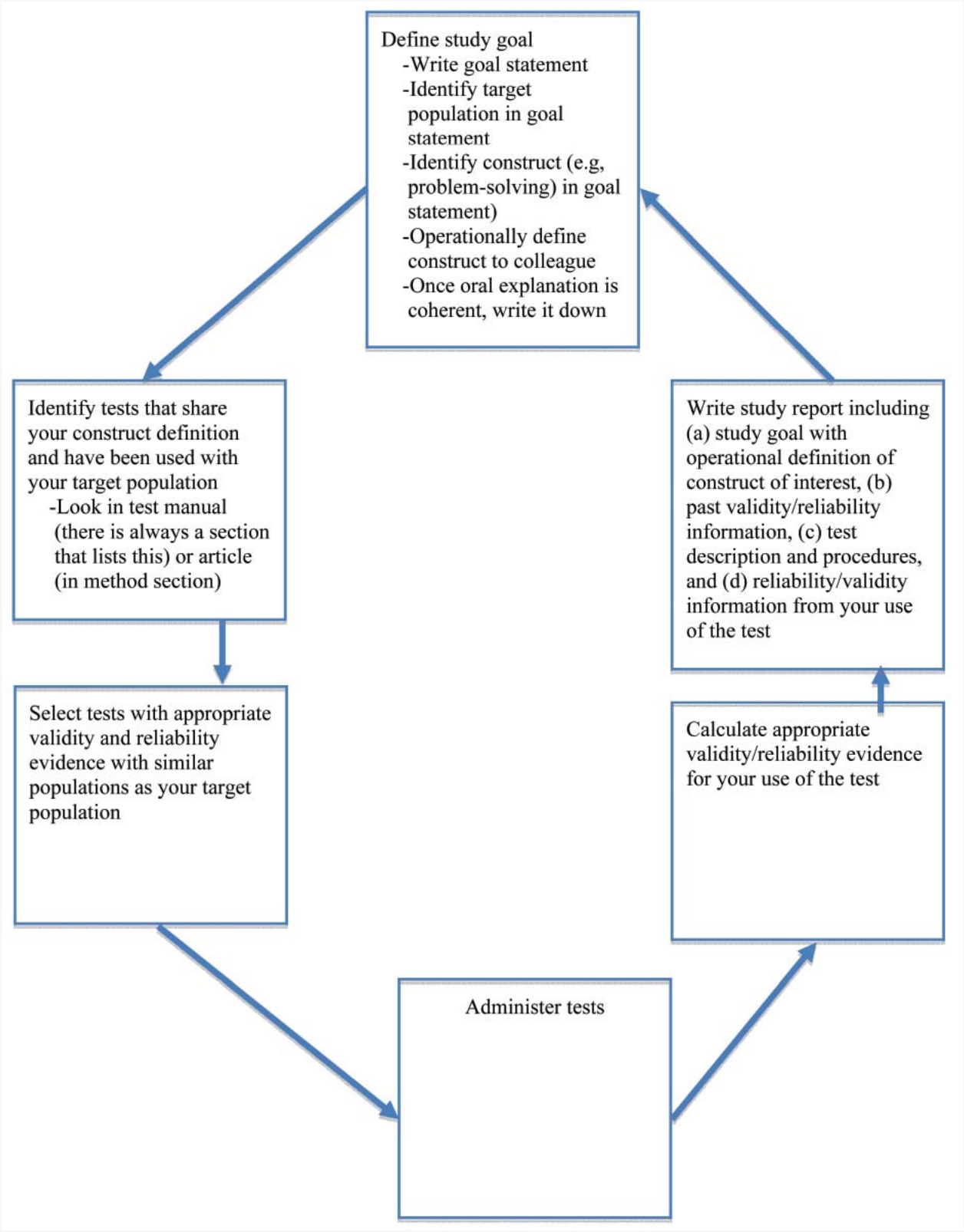
Limitations

The majority of research reviewed here was in the area of medical or allied medical education. The use of many specific examples such as simulated patient tests is most applicable to medical education. Though research on PBL in a variety of content areas and levels of education was sought, research into the effects of PBL on its desired outcomes has been more widespread in medical education (Gallagher, 1997) for several reasons. First, PBL was developed in medical schools and has thus been used in medical schools longer than in other contexts (Barrows & Tamblyn, 1980). Second, in medical school PBL is often chosen not only as a way to teach specific content, but also as a way to structure the curriculum (Albanese, 2000). As such, medical educators and deans of medical schools have been interested in seeing if PBL has a discernable impact on the desired competencies of future physicians (Albanese).

Implications

Measurement in problem-based learning (PBL) research suffers from many problems of validity and scant information about measurement procedures and accompanying theoretical frameworks. The reader and researcher should not be discouraged by this issue. First, some authors included in this review did a good job including much required

Figure 1. *Measurement reporting process.*



measurement information (Hmelo, 1998; Hmelo et al, 1997; Pedersen & Liu, 2002-2003). Second, there are many ways to increase the quality of the instruments employed in PBL research. Carefully employing the steps to create better measurements will allow the field to move forward in a very positive fashion. Such work may be difficult, if not painful at times, but the benefits will be evident in the long-term results of research agendas focusing on PBL issues.

The development of more psychometrically sound instruments will set high standards for PBL researchers. Taking these steps in a relatively young field of inquiry holds promise to serve as a model for other areas of research. Meeting such high standards is becoming critical. We encourage PBL researchers to pay full attention to measurement issues as they pursue their research agendas with the goal of producing the most accurate and defensible results possible.

Acknowledgements

The authors thank Krista Glazewski and Jennifer Richardson, as well as anonymous reviewers, for their helpful comments that were used to improve the paper.

Note

1. The name was changed to United States Medical Licensing Examination in the 1990s due to its adoption as a component of the medical licensing process. However, in this paper, versions of the exam before and after the name change will be referred to as NBME. Steps 1 and 2 of the USMLE follow largely the same format as NBME, except that in the former the clinical context of questions is more central (Albanese, 2000).

References

Aaron, S., Crocket, J., Morrish, D., Basualdo, C., Kovithavongs, T., Mielke, B., & Cook, D. (1998). Assessment of exam performance after change to problem-based learning: Differential effects by question type. *Teaching and Learning in Medicine, 10*(2), 86-91.

Albanese, M. (2000). Problem-based learning: Why curricula are likely to show little effect on knowledge and clinical skills. *Medical Education, 34,* 729-738.

Albanese, M. A., & Mitchell, S. (1993). Problem-based learning: A review of literature on its outcomes and implementation issues. *Academic Medicine, 68,* 52-81.

Amador, J. A., & Görres, J. H. (2004). A problem-based learning approach to teaching introductory soil science. *Journal of Natural Resources and Life Sciences Education, 33,* 21-27.

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher, 35*(6), 33-40.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, D. C.: American Educational Research Association.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Antepohl, W., & Herzig, S. (1999). Problem-based learning versus lecture-based learning in a course of basic pharmacology: A controlled, randomized study. *Medical Education, 33,* 106-113.

Arts, J. A. R., Gijselaers, W. H., & Segers, M. S. R. (2002). Cognitive effects of an authentic computer-supported, problem-based learning environment. *Instructional Science, 30,* 465-495.

Ausubel, D. P. (1963). *The psychology of meaningful verbal learning: An introduction to school learning.* New York: Grune & Stratton.

Barrows, H. S. (1985). *How to design a problem-based curriculum for the preclinical years.* New York: Springer.

Barrows, H. S., & Tamblyn, R. M. (1980). *Problem-based learning: An approach to medical education*. New York: Springer.

Barrows, H. S., & Tamblyn, R. M. (1976). An evaluation of problem-based learning in small groups utilizing a simulated patient. *Journal of Medical Education, 51,* 52-54.

Berkson, L. (1993). Problem-based learning: Have the expectations been met? *Academic Medicine, 68(10),* S79-S88.

Bloom, B. S. (Ed.) (1956). *Taxonomy of educational objectives: The classification of educational goals: Vol. 1. Cognitive domain.* New York: David McKay Co.

Blumberg, P., & Michael, J. A. (1992). Development of self-directed learning behaviors in a partially teacher-directed problem-based learning curriculum. *Teaching and Learning in Medicine, 4(1),* 3-8.

Bodner, G. M. (1991). A view from chemistry. In M. U. Smith (Ed.), *Toward a unified theory of problem solving: Views from the content domains* (pp. 21-33). Hillsdale, NJ: Lawrence Erlbaum.

Boshuizen, H. P. A., Schmidt, H. G., & Wassner, L. (1993). Curriculum style and the integration of biomedical and clinical knowledge. In P. A. J. Bouhuijs, H. G. Schmidt, & H. J. M. van Berkel (Eds.), *Problem-based learning as an educational strategy* (pp. 33-41). Maastricht, Netherlands: Network of Community-Oriented Educational Institutions for Health Sciences.

Bridgham, R., Solomon, D., & Haf, J. (1991). The effect of curriculum era on NBME part I outcomes in a problem-based versus a traditional curriculum track. *Academic Medicine, 66* (September Supplement), S82-S84.

Cacamese, S. M., Elnicki, M., & Speer, A. J. (2007). Grade inflation and the internal medicine subinternship: A national survey of clerkship directors. *Teaching and Learning in Medicine, 19,* 343-346.

Cochran, W. G. (1968). Errors of measurement in statistics. *Technometrics, 10*, 637-666.

Chanlin, L., & Chan, K. (2004). Assessment of PBL design approach in a dietetic web-based instruction. *Journal of Educational Computing Research, 31*, 437-452.

Chin, C., & Chia, L. (2005). Problem-based learning: Using ill-structured problems in biology project work. *Science Education, 90*(1), 44-67.

Coliver, J. A. (2000). Effectiveness of problem-based learning curricula: Research and theory. *Academic Medicine, 75,* 259-266.

Costa, A. L., & Kallick, B. (2004). *Assessment strategies for self-directed learning*. Thousand Oaks, CA: Corwin.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.; pp. 443-507). Washington: American Council on Education.

Cronbach, L. J. (1970). *Essentials of Psychological Testing* (3rd ed.). New York: Harper & Row.

Distlehorst, L. H., & Robbs, R. S. (1998). A comparison of problem-based learning and standard curriculum students: Three years of retrospective data. *Teaching and Learning in Medicine, 10,* 131-137.

Dochy, F., Segers, M., Van den Bossche, P., & Gijbels, D. (2003). Effects of problem-based learning: A meta-analysis. *Learning and Instruction, 13,* 533-568.

Dods, R. F. (1997). An action research study of the effectiveness of problem-based learning in promoting the acquisition and retention of knowledge. *Journal for the Education of the Gifted, 20,* 423-437.

Doucet, M. D., Purdy, R. A., Kaufman, D. M., & Langille, D. B. (1998). Comparison of problem-based learning and lecture format in continuing medical education on headache diagnosis and management. *Medical Education, 32,* 590-596.

Educational Researcher (2007). *Special issue on validity*, *36*(8).

Embretson, S. E. (2007). Construct validity: A universal validity system or just another test evaluation procedure, *Educational Researcher, 36,* 449-455.

Evensen, D. H., Salisbury-Glennon, J. D., & Glenn, J. (2001). A qualitative study of six medical students in a problem-based curriculum: Toward a situated model of self-regulation. *Journal of Educational Psychology, 93,* 659-676.

Federation of State Medical Boards & National Board of Medical Examiners. (2005a). *United States Medical Licensing Examination: Step 1 content outline*. Retrieved October 26, 2006, from http://www.usmle.org/step1/intro.htm.

Federation of State Medical Boards & National Board of Medical Examiners. (2005b). *United States Medical Licensing Examination: Step 2 content outline*. Retrieved October 26, 2006 from http://www.usmle.org/step2/intro.htm.

Finch, P. M. (1999). The effect of problem-based learning on the academic performance of students studying podiatric medicine in Ontario. *Medical Education, 33,* 411-417.

Fleiss, J. L., & Shrout, P. E. (1977). The effects of measurement errors on some multivariate procedures. *American Journal of Public Health, 67,* 1188-1191.

Gall, M. D., Gall, J. P., & Borg, W. R. (2003). *Educational research: An introduction* (7th ed.). Boston: Pearson Education.

Gallagher, S. A. (1997). Problem-based learning: Where did it come from, what does it do, and where is it going? *Journal for the Education of the Gifted, 20,* 332-362.

Gallagher, S. A., & Stepien, W. J. (1996). Content acquisition in problem-based learning: Depth versus breadth in American studies. *Journal for the Education of the Gifted, 19,* 257-275.

Gallagher, S. A., Stepien, W. J., & Rosenthal, H. (1992). The effects of problem-based learning on problem solving. *Gifted Child Quarterly, 36,* 195-200.

Gibbons, M. (2002). *The self-directed learning handbook*. San Francisco: Jossey-Bass.

Gijbels, D., Dochy, F., Van den Bossche, & Segers, M. (2005). Effects of problem-based learning: A meta-analysis from the angle of assessment. *Review of Educational Research, 75(1),* 27-61.

Gilhooly, K. J. (1990). Cognitive psychology and medical diagnosis. *Applied Cognitive Psychology, 4,* 261-272.

Glaser, R., Raghavan, K., & Baxter, G. P. (1992). *Cognitive theory as the basis for design of innovative assessment: Design characteristics of science assessments*. CSE Technical Report No. 349. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing. [Eric Document Reproduction Service No. ED357038].

Goodman, L. J., Brueschke, E. E., Bone, R. C., Rose, W. H., Williams, E. J., & Paul, H. A. (1991). An experiment in medical education: A critical analysis using traditional criteria. *Journal of the American Medical Association, 265,* 2373-2376.

Hamdy, H., Prasad, K., Anderson, M. B., Scherpbier, A., Williams, R., Zwierstra, R., et al. (2006). BEME systematic review: Predictive values of measurements obtained in medical schools and future performance in medical practice. *Medical Teacher, 28*(2), 103-116.

Heale, J., Davis, D., Norman, G., Woodward, C., Neufeld, V., & Dodd, P. (1988). A randomized trial assessing the impact of problem-based versus didactic teaching methods in CME. *Research in Medical Education; Proceedings of the Annual Conference, 27,* 72-77.

Hmelo, C. E. (1998). Problem-based learning: Effects on the early acquisition of cognitive skill in medicine. *Journal of the Learning Sciences, 7,* 173-208.

Hmelo, C. E., & Ferrari, M. (1997). The problem-based learning tutorial: Cultivating higher-order thinking skills. *Journal for the Education of the Gifted, 20,* 401-422.

Hmelo, C. E., Gotterer, G. S., & Bransford, J. D. (1997). A theory-driven approach to assessing the cognitive effects of PBL. *Instructional Science, 25,* 387-408.

Hmelo-Silver, C. E. (2004). Problem-based learning: What and how do students learn? *Educational Psychology Review*, *16,* 235-266.

Hogan, T. P, & Agnello, J. (2004). An empirical study of reporting practices regarding measurement validity. *Educational and Psychological Measurement, 64,* 802-812.

Hughes, P. (2002). Can we improve on how we select medical students? *Journal of the Royal Society of Medicine, 95(1),* 18-22. Retrieved October 30, 2007 from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1279142.

Hulse, R. A. (2006). Preparing K-12 students for the new interdisciplinary world of science. *Journal of Experimental Biology and Medicine, 231,* 1192-1196.

Jonassen, D. H. (2003). *Learning to solve problems with technology: A constructivist perspective* (2nd ed.). Upper Saddle River, NJ: Pearson.

Kalaian, H. A., Mullan, P. B., & Kasim, R. M. (1999). What can studies of problem-based learning tell us? Synthesizing and modeling PBL effects on National Board of Medical Examination performance: Hierarchical linear modeling meta-analytic approach. *Advances in Health Science Education, 4*, 209-221.

Kassierer, J. P., & Gorry, G. A. (1978). Clinical problem-solving: A behavioral analysis. *Annals of Internal Medicine, 89,* 245-255.

Kaufman, A., Mennin, S., Waterman, R., Duban, S., Hansbarger, C., Silverblatt, H., Obenshain, S., Kantrowitz, M., Becker, T., Samet, J., & Wiese, W. (1989). The New Mexico experiment: Educational innovation and institutional change. *Academic Medicine, 64,* 285-294.

Kerlinger, F. N., & Lee, H. B. (2000). *Foundations of behavioral research* (4th ed.). South Melbourne, Australia: Wadsworth.

Larkin, J., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. *Science, 208,* 1335-1342.

Lee, M., & Kim, D. (2005). The effects of the collaborative representation supporting tool on problem-solving processes and outcomes in web-based collaborative problem-based learning (PBL) environments. *Journal of Interactive Learning Research, 16,* 273-293.

Lewis, K. E., & Tamblyn, R. M. (1987). The problem-based learning approach in baccalaureate nursing education: How effective is it? *Nursing Papers/Perspectives en Nursing, 19(2),* 17-26.

Lohman, M. C., & Finkelstein, M. (2000). Designing groups in problem-based learning to promote problem-solving skill and self-directedness. *Instructional Science, 28,* 291-307.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.; pp. 13-103). New York: American Council on Education.

Moore, G., Block, S., & Mitchell, R. (1990). *A randomized controlled trial evaluating the impact of the New Pathway curriculum at Harvard Medical School.* Cambridge, MA: Harvard University. (ERIC Document Reproduction Service No. ED359866)

Moore, G. T., Block, S. D., Style, C. G., & Mitchell, R. (1994). The influence of the New Pathway curriculum on Harvard medical students. *Academic Medicine, 69,* 983-989.

Neufeld, V. (1989). Issues and guidelines for student and program evaluation. In H. G. Schmidt, M. Lipkin, M. W. de Vries, & J. M. Greep (Eds.), *New directions in medical education: Problem-based learning and community-oriented medical education* (pp. 196-205). New York: Springer-Verlag.

Neufeld, V., & Sibley, J. C. (1989). Evaluation of health sciences education programs: Program and student assessment at McMaster University. In H. G. Schmidt, M. Lipkin, M. W. de Vries, & J. M. Greep (Eds.), *New directions in medical education: Problem-based learning and community-oriented medical education* (pp. 165-179). New York: Springer-Verlag.

Newble, D. I., & Clarke, R. M. (1986). The approaches to learning of students in a traditional and in an innovative problem-based medical school. *Medical Education, 20,* 267-273.

Nunnally, I. H., & Bernstein, I. H. (1994). *Psychometric Theory*. New York: McGraw-Hill.

Pajares, F. (2002). Gender and perceived self-efficacy in self-regulated learning. *Theory into Practice, 41*(2), 116-125.

Pedersen, S., & Liu, M. (2002-2003). The transfer of problem-solving skills from a problem-based learning environment: The effect of modeling an expert's cognitive processes. *Journal of Research on Technology in Education, 35,* 303-320.

Pedhazur, E. J. & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Lawrence Erlbaum.

Reiter, S. A., Rasmann-Nuhlicek, D. N., Biernat, K., & Lawrence, S. L. (1994). Registered dieticians as problem-based learning facilitators in a nutrition curriculum for freshman medical students. *Journal of the American Dietetic Association, 94,* 652-654.

Richards, B. F., Ober, K. P., Cariaga-Lo, L., Camp, M. G., Philip, J., McFarlane, M., Rupp, R., & Zaccaro, D. J. (1996). Ratings of students' performances in a third-year internal medicine clerkship: A comparison between problem-based and lecture-based curricula. *Academic Medicine, 71,* 187-189.

Rudner, L. M. (1994). Questions to ask when evaluating tests. *Practical Assessment, Research & Evaluation*, 4(2). Retrieved July 29, 2008 from http://PAREonline.net/getvn.asp?v=4&n=2.

Sanci, L. A., Coffey, C. M. M., Veit, F. C. M., Carr-Gregg, M., Patton, G. C., Day, N., & Bowes, G. (2000). Evaluation of the effectiveness of an educational intervention for general practitioners in adolescent health care: Randomised controlled trial. *British Medical Journal, 320,* 224-229.

Santos-Gomez, L., Kalishman, S., Rezler, A., Skipper, B., & Mennin, S. P. (1990). Residency performance of graduates from a problem-based and a conventional curriculum. *Medical Education, 24,* 366-375.

Scandura, J. M. (1977). *Problem solving: A structural/process approach with instructional implications.* New York: Academic Press.

Schank, R. C., & Abelson, R. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schmidt, H. G., Machiels-Bongaerts, M., Hormans, H., ten Cate, T. J., Venekamp, R., & Boshuizen, H. P. A. (1996). The development of diagnostic competence: Comparison of problem-based, and integrated, and a conventional medical curriculum. *Academic Medicine, 71,* 658-664.

Schoenfeld, A. H. (1985). *Mathematical problem solving.* Orlando, FL: Academic Press.

Schuwirth, L. W. T, Verhoeven, B. H., Scherpbier, A. J. J. A., Mom, E. M. A., Cohen-Schotanus, J., Van Rossum, H. J. M., & van der Vleuten, C. P. M. (1999). An inter- and intra-university comparison with short case-based testing. *Advances in Health Sciences Education, 4,* 233-244.

Schwartz, R. W., & Burgett, J. E. (1997). Problem-based learning and performance-based testing: Effective alternatives for undergraduate surgical education and assessment of student performance [Electronic version]. *Medical Teacher, 19(1).*

Segers, M. S. R. (1997). An alternative for assessing problem-solving skills: The overall test. *Studies in Educational Evaluation, 23,* 373-398.

Smith, M. U. (1991). A view from biology. In M. U. Smith (Ed.), *Toward a unified theory of problem solving: Views from the content domains* (pp. 1-19). Hillsdale, NJ: Lawrence Erlbaum.

Smits, P. B. A., Verbeek, J. H. A. M., & de Buisjonjé, C. D. (2002). Problem-based learning in continuing medical education: A review of controlled evaluation studies. *British Medical Journal, 324,* 153-156.

Stepien, W. J. Gallagher, S. A., & Workman, D. (1993). Problem-based learning for traditional and interdisciplinary classrooms. *Journal for the Education of the Gifted, 16,* 338-357.

Sugrue, B. (1995). A theory-based framework for assessing domain specific problem-solving ability. *Educational Measurement: Issues and Practice, 14,* 29-36.

Sugrue, B. (1993). *Specifications for the design of problem-solving assessments in science.* Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing. [Eric Document Reproduction Service No. ED372081].

Torp, L., & Sage, S. (1998). *Problems as possibilities: Problem-based learning for K-12 education.* Alexandria, VA: Association for Supervision and Curriculum Development.

US Department of Education. (N. d. a). Doing What Works. Retrieved August 1, 2008 from http://dww.ed.gov/.

US Department of Education. (N. d. b). What Works Clearinghouse. Retrieved August 1, 2008 from http://ies.ed.gov/ncee/wwc/.

Vernon, D. T. A., & Blake, R. L. (1993). Does problem-based learning work? A meta-analysis of evaluative research. *Academic Medicine, 68,* 550-563.

Warner, I. M. (2004). Climbing Bloom's ladder. *Journal of Chemical Education, 81,* 1413

West, D. A., Umland, B. E., & Lucero, S. M. (1985). Evaluating student performance. In A. Kaufman (Ed.), *Implementing problem-based medical education: Lessons from successful innovations* (pp. 144-163). New York: Springer.

Woodward, C. A. (1996). Problem-based learning in medical education: Developing a research agenda. *Advances in Health Sciences Education, 1,* 83-94.

Yang, S. C. (2003). Reconceptualizing think-aloud methodology: Refining the encoding and categorizing techniques via contextualized perspectives. *Computers in Human Behavior, 19,* 95-115.

Zumbach, J., Kumpf, D., & Koch, S. C. (2004). Using multimedia to enhance problem-based learning in elementary school. *Information Technology in Childhood Education Annual, 2004(1),* 25-37.

Brian R. Belland is an assistant professor of instructional technology and learning sciences at Utah State University. His research interests center on the use of technology to support problem-solving and argumentation among middle school and university students, specifically during problem-based units. He also is interested in strategies to promote technology integration and the impact of measurement quality on research findings.

Brian F. French is an associate professor in the Department of Educational Leadership and Counseling Psychology in the area of Research, Evaluation, and Measurement at Washington State University. Dr. French's research focuses on applied educational and psychological measurement issues. Specifically, he is interested in test validity. He enjoys applying various psychometric methods to solve applied measurement problems.

Peggy A. Ertmer is a professor of educational technology at Purdue University. Her research interests relate to helping students become expert instructional designers, specifically through the use of case- and problem-based learning methods. She currently serves as the editor of *IJPBL*.