1

The Measurement and Development of Professional Performance: An Introduction to the Topic and a Background to the Design and Origin of This Book

K. ANDERS ERICSSON, RAY S. PEREZ, DAVID W. ECCLES, LAURA LANG, EVA L. BAKER, JOHN D. BRANSFORD, KURT VANLEHN, AND PAUL WARD

Developments in technology and software engineering are making many types of traditional jobs, such as bookkeeping, accounting, routine design, and document indexing, virtually obsolete (Rasmussen, 2000). The rapid improvements in technology and automated work methods challenge even the traditional idea of stable job competence, as well as the ability to predict the length and the nature of current or future professional careers. Today's work conditions require ongoing adaptations by employees and entrepreneurs to new demands and competitive opportunities through continuing education and training. Technological innovations, such as the World Wide Web, broadband communication, and highly portable communication and work devices, have reduced the constraints of geography on work. Today, many services can be provided with an equivalent level of quality irrespective of whether the provider is in the office next door or on a different continent. It is, indeed, becoming an age of global consumerism in which one can "work with anyone, anytime, anywhere." Additionally, many specialized skills previously performed by human beings are now the purview of automated systems, and can often be conducted anywhere in the world at a fraction of the cost if carried out in Western Europe and North America. This technological revolution suggests that the competitive advantage of any country aspiring to economic prosperity is increasingly dependent on the capability of both its research and development groups and its skilled workforce, not only to create and develop new and improved products that are at the cutting edge, but also to quickly react and adapt to market forces. The shift from the industrial to the technological age clearly motivates increased efforts to support the development of existing and future professionals with these skill sets; to identify existing

Author Note: Generous support to the authors for the preparation of this article and the Development of Professional Performance conference described within it was provided by the US Office of Naval Research, grant # N00014-05-1-0785.

experts and high-performers; and to provide suitable learning environments that can position and maintain companies and nations ahead of the curve.

There are many recent books that address professional development and the need to foster and develop a company's personnel and skilled professionals. What makes this book unique and distinctive is its focus on measurable performance in representative activities that capture expertise in the associated domain, and its study of factors that promote the acquisition and development of superior professional performance. The commitment to focus on measurable performance comes from a desire to study superior professional achievement scientifically and is based on recent advances in the study of expert performance in more traditional domains of expertise, such as chess, music, sports, and medicine (Ericsson, Charness, Feltovich, & Hoffman, 2006). Recent research on reproducible superior performance has successfully challenged the popular myth that expertise and superior performance develop as inevitable, naturally emerging consequences of many years of experience in a domain. There is now ample evidence from many different domains that the number of years of experience is a poor predictor of objective professional performance (for more recent extensive reviews and a meta analysis supporting this claim, see Choudhrey, Fletcher, & Soumerai, 2005; Ericsson, 2004, 2006a; Ericsson, Whyte, & Ward, 2007). In fact, there is even evidence showing that the objective performance of medical professionals decreases as the number of years since graduation from initial professional education increases (Choudhry et al., 2005; Ericsson et al., 2007). For example, years of experience and age has been found to be negatively related to adherence of accepted standards for medical treatment. Even more importantly, survival of patients treated for heart problems has been found to decrease with number of years since graduation from medical school of the treating physician when other relevant variables were statistically controlled (Choudhry et al., 2005).

During the last five years an impressive number of scholarly books have been published on the topics of expertise, expert performance, high levels of skill, and excellence (Boshuizen, Bromme, & Gruber, 2004; Chaffin, Imreh, & Crawford, 2002; Ericsson, Charness, Feltovich, & Hoffman, 2006; Feist, 2006; Ferrari, 2002; Hoffman, 2007; Kurz-Milcke & Gigenrenzer, 2004; Montgomery, Lipshitz, & Brehmer, 2005; Runco, 2007; Simonton, 2004; Starkes & Ericsson, 2003; Sternberg & Grigorenko, 2003; Tetlock, 2005; Tsui, 2003; Weisberg, 2007; Williamon, 2005; Williams & Hodges, 2004). These books describe a wide range of methods used to study the structure and acquisition of high levels of achievement across a wide range of different domains of expertise, such as music, teaching, chess, sports, business, and medicine.

The study of expertise and expert performance has been conducted with several different approaches, but two approaches have been particularly dominant. The original theory of human expertise was developed by de Groot (1946/1978) and Simon and Chase (1973) and emphasized the importance

of extended professional experience for the attainment of the expert level of achievement. Given the difficulties of measuring objective performance in most domains, this approach focused on how less accomplished individuals, such as novices and beginners, differed from experts, who were defined as individuals with extensive professional experience (typically over 10 years), or nominated by their peers as particularly accomplished professionals, or both (Chi, 2006). As mentioned above, experience was later found to be a poor predictor of objective performance and this finding led to proposals of an alternative approach, namely the *expert performance approach* (Ericsson & Lehmann, 1996; Ericsson & Smith, 1991). This approach focuses on objectively measurable superior performance on representative tasks that capture expertise in the domain (Ericsson, 2006a, 2006b; Ericsson & Smith, 1991). This approach to the measurement of expert performance avoids the problem of using questionable criteria that is based on professional experience and peer nomination to identify reproducibly superior performance. Throughout this book, we have encouraged contributors to cite research that used objectively measured performance to support claims about antecedents to increases in professional performance and, thus, professional development. Examinations of the changes in the nature of performance over extended periods of development have uncovered effective methods for enhancing many different aspects of performance, such as deliberate practice (e.g., Ericsson, 2006a; Ericsson, Krampe, & Tesch-Römer, 1993).

The unique perspective of this book on the study of professional development comes from the mission of a grant from the Office of Naval Research (Contract # N00014-05-1-0785) to convene a conference on the possibility of applying the expert performance approach to the development and maintenance of skilled and expert performance with Laura Hassler Lang as Principle Investigator and David Eccles and Anders Ericsson as co-Principle Investigators, and with Ray Perez as Contract Officer. The focus was on producing a review and an evaluation of state-of-the-art knowledge about instruction and training in the development and maintenance of professional skills, searching for research that emphasized *measurable objective performance.* The goal was to develop a synthesis of the knowledge of the structure and acquisition of expert performance in traditional domains, such as chess, music, and sports, recently summarized in the *Cambridge Handbook of Expertise and Expert Performance* (Ericsson et al., 2006) and relate it to knowledge about development of performance in a broader range of domains, namely professional performance. Consequently, the focus of this book is on research that has examined the performance of personnel involved in actual professional settings, including medicine, industry, and the military.

The primary focus of this book is on individual performance. While this will involve performance by individual members of teams, the invited contributors were encouraged not to emphasize team performance, which has received a great deal of research attention elsewhere (Bowers, 2006; Salas &

Fiore, 2004) and has been very difficult to measure with objective methods. Before we give more details about how the contributions to this book were produced and later revised, we will consider the history of research on measurement of professional performance and how earlier approaches differ from the expert performance approach and the invited contributions to this book.

A BRIEF HISTORY OF RESEARCH ON OBJECTIVE MEASUREMENT OF PROFESSIONAL PERFORMANCE

The 20th century saw impressive advances in psychometric theories and procedures for the development of tests of general and basic abilities. In contrast, the development of theories of the structure of professional performance and its associated measurement were largely neglected. In the early 1990s, Wigdor and Green (1991a) published a book commissioned by the National Research Council in which some of the most outstanding researchers in applied psychology summarized the past research as well as a new project on the development of "Performance Assessment for the Workplace." Wigdor and Green (1991a) argued that it has been much easier to develop sophisticated test instruments and associated statistical techniques "than to find adequate measures of performance to use as criteria in judging the relevance of the tests. For the most part, industrial and organizational psychologists and their institutional clients have used measures of convenience, such as training grades or supervisor ratings as a surrogate for job performance" (p. 22). Wigdor and Green exemplify these concerns by an extended description of a report by Captain John Jenkins on the selection of pilots during World War II. This report showed that psychometric tests were able to predict in advance which of the candidates would be eliminated from the accelerated training program due to poor flying performance, fear, or their own requests. However, an analysis of actual combat performance of successfully graduated pilots showed that none of the tests "gave evidence of predicting the combat criterion measures to any marked degree" (Jenkins's report, cited by Wigdor & Green, 1991a, p. 25). Their review concluded that prior use of selection tests had been validated against criteria based on successfully completing training, such as multiple-choice tests of knowledge at the end of the course, rather than actual job performance attained after some period of on-the-job experience. By designing the selection tests to predict performance *during* training rather than predicting subsequent performance on the job, the selection tests would be likely to screen out many individuals who would actually have very successful careers. The use of written selection tests that focus on predicting performance *during* the initial schoolhouse training resulted in the rejection of large segments of the population with low scores on the traditional types of psychometric tests. With the change into a volunteer military service in the United States, military recruiters were

faced with the problem of what to do when there were not enough applicants who scored well on the selection tests used for recruitment. To what extent would individuals scoring below average on the selection tests be able to develop into average or above average performers on their military jobs? The first step to study this issue scientifically would require that applicants with a wide range of scores on the selection tests be admitted to training and their subsequent performance on the job evaluated by fair objective criteria. Instead of the typical measures of job performance, such as supervisor ratings or knowledge tests or both, Wigdor and Green (1991a) recommended "the criterion measure that comes closest to actual job performance, the hands-on job-sample test" (p. 30).

A group of scientists led by Wigdor and Green (1991a) gave scientific oversight to a massive project in the U.S. Department of Defense, namely the Joint-Service Job Performance Measurement/Enlistment Standards (JPM) project. In this project, several thousands of enlisted men and women were given selection tests at entry, such as variants of the Armed Services Vocational Aptitude Battery (ASVAB), and applicants with a wide range of scores on the selection tests were accepted for training. Later on, after several months of working, their performance was evaluated with various measures and tests of performance. The most innovative aspects of this project concerned the development of hands-on work-sample tests that would capture the actual behaviors necessary for executing particular job responsibilities for soldiers with a given occupational category, rather than testing soldiers by asking them to verbally describe procedure or answer multiple-choice questions about job-related knowledge. Wigdor and Green (1991a) define a hands-on *work-sample* as "an actual part of a job, chosen for its representativeness and importance to success on the job" (p. 59). For example, a work-sample test for a secretary might involve "a word-processing task, a filing task, and a form completion task" (p. 59). The work-sample methodology involved transforming these job activities into standardized hands-on tasks where all tested individuals can perform action sequences that can be checked for accuracy by trained observers.

Many of the findings from the JPM project with data from over 5,000 military personnel have been reported by Wigdor and Green (1991a, 1991b). One of the their general conclusions was that "it is possible to develop hands-on measures of job performance for a wide range of military jobs" (Wigdor & Green, 1991a, p. 183) and that hands-on performance on the work-sample tasks did not increase very much overall as function of length of work experience beyond the first year. Furthermore, the individual differences in hands-on performance attributable to cognitive abilities were reduced after the first year of service (Wigdor & Green, 1991a, p. 164). A part of the JPM study, referred to as Project A, was extended by the U.S. Army to include measurement of job performance immediately after training, and during the first and second 3-year tours of duty (Campbell, 2001; Campbell & Knapp, 2001). The focus

of Project A (Campbell & Knapp, 2001) was not on identifying high levels of performance, but on developing tests for selecting recruits and assigning them to training for different military occupations where they would be able to exhibit an *acceptable* level of performance. In the work on identifying the latent variables underlying the large body of tests of different types of abilities as well as ratings and objective measures of job performance, the JPM project and Project A found that some of the different measures of job performance had low inter-correlations (for an extended discussion, see Knapp, Campbell, Borman, Pulakos, & Hanson, 2001). For example, observed correlations between the most valid measure of job performance, namely scores on the hands-on work-sample tests, correlated poorly with the job ratings of the soldiers' supervisors, with an average correlation coefficient of around 0.2 (Campbell, McHenry, & Wise, 1990; Wigdor & Green, 1991a). In the concluding chapter in a book on Project A, Campbell (2001) argued: "The efforts also convincingly show that a clear focus on the latent structure will illuminate the gaps as well as the strengths of our research knowledge. For example, the lack of research attention to the latent structure of 'job-relevant knowledge' and 'job-relevant skill' became painfully obvious" (p. 588). More generally, these studies on the selection of personnel uncovered several remaining obstacles for ultimate validation of the standardized objective tests for selecting and identifying individuals who would ultimately develop reproducibly superior performance under the target condition, namely performance under combat conditions. The developed tests of job performance focused primarily on assessment of reliable execution of standard procedures, rather than on assessment of skilled performance and high fidelity simulations of representative situations under combat conditions and real-time constraints. Even the most recent books (Bennett, Lance, & Woehr, 2006a) on performance measurement continue to argue that "the criterion problem continues to be one of the most vexing issues facing organization researchers and practitioners today" (Bennett, Lance, & Woehr, 2006b, p. 1). In their review of the progress on the development of criteria for job performance, Austin and Crespin (2006) described the emerging knowledge about motivational factors that are correlated with average productive performance across days and weeks. These factors included the effects of counterproductive work behavior (Miles, Borman, Spector, & Fox, 2002), the importance of contextual behaviors that support the social and motivational work situation (Motowidlo, Borman, & Schmit, 1997), and the differences between maximal performance during a test and the actual average performance measured on the job (Sackett, Zedeck, & Fogli, 1988). However, there has been less progress on the development of objective measures for job performance.

Is it possible to develop a methodology that can accurately capture and measure superior professional performance in critical situations, such as emergencies and other stressful task conditions? In the next section we will

describe how the methods of the expert performance approach can be adapted toward the achievement of this goal.

THE EXPERT PERFORMANCE APPROACH TO THE STUDY OF SUPERIOR PERFORMANCE

The JPM project focused on basic performance and also found modest improvements in hands-on performance as a function of the length of experience on the job, suggesting that experience on the job may not dramatically improve this aspect of measured job performance. (See Mayberry & Carey, 1997, for an exceptionally large improvement in performance of helicopter repair technicians as a function of more experience even beyond the first year.) As noted earlier, experience on the job has not been found to relate closely to improvement in performance on representative tasks in domains such as the financial investment of clients' funds, treatment of patients with psychotherapy, and decision making in a wide range of situations involving prediction of behavior and events (Ericsson, 1996, 2004; Ericsson & Lehmann, 1996; Ericsson, Whyte, & Ward, 2007).

In the expert performance approach (Ericsson, 2006a, 2006b; Ericsson & Smith, 1991), investigators identify those individuals (expert performers) who exhibit superior performance on tasks that capture the essence of expertise in the critical domain. These studies encourage the identification of superior performers and specialists who are able to successfully deal with challenging and non-routine cases.

Once tasks with superior performance have been identified in everyday life, then the next step in the expert performance approach involves the design of tasks that can reproduce the superior expert performance in the laboratory. Repeated elicitation of the superior performance on representative tasks permits the application of standard cognitive methods to analyze the mechanisms that mediate experts' superior performance. The general paradigm pioneered by de Groot (1946/1978; Ericsson & Smith, 1991) started with an analysis of naturally occurring behavior, such as games between chess masters. He then identified key chess positions, where a chess move needs to be made, and where the best move can be determined after the fact. More generally, the expert performance approach involves the identification of critical situations, where an immediate action needs to taken, and where the correct action can be assessed after the fact. These critical situations can then be presented, for instance, as videos or simulations, with the requirement of immediate action to experts and less skilled performers to let them generate their best action. By presenting a sequence of these representative tasks and recording the speed and accuracy of generated actions, it has been possible to capture objective performance in different domains, such as chess, music, and the board game Scrabble, which is closely related to performance in tournaments and

competitions (Tuffiash, Roring, & Ericsson, 2007; for a review, see Ericsson, 2006b). For example, in chess it is possible to take measures of these abilities with 10 to 20 minutes of testing that approach the validity of measures based on outcomes of chess games lasting 50 to 200 hours during tournament play (Ericsson & Williams, 2007; van der Maas & Wagenmakers, 2005).

Once the superior performance of experts can be repeatedly reproduced with representative tasks in the laboratory, it is possible to apply the entire toolbox of cognitive psychology and trace performance with process measures, such as latencies, eye movements, and concurrent or retrospective reports, and to design experiments to test hypotheses about the nature and structure of the mediating mechanisms (Ericsson, 2006b). Research on expertise, especially expert memory performance, has shown how protocols can identify complex mechanisms that can later be confirmed by specially designed experiments (Ericsson, 2006b). The expert performance approach, with its identification of mechanisms mediating consistently superior performance, has now been successfully applied to a wide range of activities, such as medical diagnosis, surgical procedures, music performance, writing, painting, Scrabble, darts, ballet, soccer, running, field hockey, volleyball, rhythmic gymnastics, and tennis (Ericsson, 2006a). The most interesting and exciting discovery from studying the superior performance of experts is that it has been directly linked to complex representations that are specific to the domain of expertise and, consequently, were developed as a result of extended exposure and practice (Ericsson, 2006b). For example, chess masters develop the ability to explore consequences of long sequences of chess moves mentally and are, in fact, able to play blindfold chess; that is, to play without seeing a physical chess board and pieces. Similarly, elite athletes, such as tennis and squash players, develop superior ability to anticipate the trajectory of future shots, as revealed by successful predictions of ball landing locations generated even before the opponent player has hit the ball with his/her racquet.

More experience does not automatically lead to increased engagement in dedicated and focused practice to reach the highest level of performance, such as wining international competitions (Ericsson, 2006a; Simon & Chase, 1973). More generally, diaries and retrospective estimates of weekly engagement in particular activities have demonstrated that not all domain-related activities are correlated with increases in performance. Ericsson, Krampe, and Tesch-Römer (1993) found that the total amount of domain-related activities for musicians was not associated with differences in attained levels of performance. The activity most closely related to level of performance was the amount of engagement in solitary practice as reflected by diaries and retrospective estimates. During solitary practice, musicians work on clear practice goals recommended by their teachers using methods designed to improve specific aspects of their individual performance. The improvements in performance are due to changes in performance linked to repetitions

and refinements of processes with problem solving in response to feedback (deliberate practice). For example, piano students successfully master their assigned pieces of music by practicing and working on difficult sections by re-fingering transitions, repetitions, and speed work. Several researchers have reported a consistent association between the amount and quality of solitary activities meeting the criteria of deliberate practice and performance in different domains of expertise, such as chess (Gobet & Charness, 2006), darts (Duffy, Baluch, & Ericsson, 2004), music (Lehmann & Gruber, 2006), many types of sports (Ward, Hodges, Starkes, & Williams 2007; Williams, Ericsson, Ward, & Eccles, 2008), Scrabble (Tuffiash et al., 2007), and several other diverse domains (Ericsson, 2006a).

In sum, research within the expert performance framework has shown that individual differences in sustained activity and accumulated deliberate practice are correlated with attained performance in a wide range of domains of expertise.

THE PROCESS OF GENERATING THIS BOOK

Most of the authors of this chapter met in the spring of 2006 to design a conference on the objective measurement of professional performance and its development and acquisition in response to training and deliberate practice. We all agreed that we knew of no similar effort to organize a conference or produce an edited book on this topic.

We decided on a general approach that would maximize the chances that we would be able to identify published research on objective measurement of professional performance, and to stimulate discussion about related issues and their relation to training. We invited the most prominent and exciting researchers who had studied relevant issues to prepare chapters on how their domains of research related to the objective measurement of professional performance, and grouped them into four sections. To distill the most interesting ideas, we invited five eminent cognitive psychologists and educators to serve as discussants of the presentations within each section. The mere knowledge that one's chapter would be publicly discussed by eminent scientists should have motivated each group of authors to do their very best job. We selected eminent scientists who have served as editors for major journals and many books, expecting them to be able to critically review the presented material as well as extract and induce the key issues and fundamental empirical findings and connections to general theories in psychology and education. Hence, it was clear that the chapters and the resulting book would have to go through several iterations until we reached the final published form.

The first step was to generate a list of eminent scientists as well as key researchers who could present findings related to the key set of issues. We were grateful to find that everyone that we contacted was intrigued and

willing to participate. By distributing abstracts for all of the presentations in the fall of 2006, we started a process for increased connections between chapters, exchange of relevant information, and greater integration of the contributions. In the spring of 2007, we planned a conference, organized under the leadership of Eccles, where the invited presenters gave brief summaries of their previously circulated chapters, followed by questions and most importantly a discussion by our invited eminent scientists. The conference, chaired by Eccles and Ward, was structured around four groups of presentations with their associated discussant, and was concluded by presentations from several individuals with broad perspectives on professional training. These individuals presented overviews focusing on the implications and applications of the presented ideas for future research and development of training devices. After the conference, held at the Westin Grand Bohemian Hotel on March 2–4, 2007, in Orlando, Florida, U.S.A., the plan then was that all presenters would revise their chapters, which were to be given to each of the group discussants, who then were to finalize the written version of their commentaries. Finally, all the written materials (chapters and written commentaries) were handed over to two general discussants to allow them to finalize their contribution.

While organizing a conference is always challenging, we encountered few unexpected complications with this conference and the subsequent completion of this edited volume. The fact that all invited participants made plans to attend our conference and viewed our project with interest and excitement certainly reflected positively on the experience, although not everything went entirely to plan. One presenter was snowed in for a couple of days in a large city in North America and was unable to catch a flight in time to attend even part of the conference. Another participant was forced at the last minute to stay home for a family emergency. Unexpected problems led to a few changes, but the published book is a refined and distilled version of the original plan generated in the spring of 2006, and thus the ideas and information presented at our conference in Orlando in 2007. In the next section we provide a summary of the content of the subsequent chapters of this book, with a focus on our goal to find and develop objective measures of professional performance and to identify the training and practice activities that lead to improvements and maintenance of this type of performance.

THE OUTLINE OF THIS BOOK

The first section of the book is concerned with general overviews of the challenges of objective measurements and training of professional performance in some of the major domains, and a review of the progress toward objective measurement.