U.S. Army Research Institute
for the Behavioral and Social Sciences

Research Report 1900

Self Assessment: Review and Implications for Training

John T. Breidert
Western Kentucky University
Consortium Research Fellows Program

Jeffrey E. Fite
U.S. Army Research Institute

June 2009

Approved for public release; distribution is unlimited

AUTHENTICATED
U.S. GOVERNMENT
INFORMATION
GPO

U.S. Army Research Institute
for the Behavioral and Social Sciences

A Directorate of the Department of the Army
Deputy Chief of Staff, G1

Authorized and approved for distribution:

MICHELLE SAMS
Director

Technical review by

John Barnett, U.S. Army Research Institute
Christopher L. Vowels, U.S. Army Research Institute

NOTICES

DISTRIBUTION: Primary distribution of this Research Report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, Attn: DAPC-ARI-ZXM, 2511 Jefferson Davis highway, Arlington, Virginia 22202-3926.

FINAL DISPOSITION: This Research Report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this Research Report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy)	2. REPORT TYPE	3. DATES COVERED (from. . . to)
June 2009	Final	January 2008 to September 2008

4. TITLE AND SUBTITLE	5a. CONTRACT OR GRANT NUMBER
Self-Assessment: Review and Implications for Training	
	5b. PROGRAM ELEMENT NUMBER 622785
6. AUTHOR(S) John T. Breidert (Western Kentucky University), and Jeffrey E. Fite (USARI)	5c. PROJECT NUMBER A790
	5d. TASK NUMBER 331
	5e. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) U.S. Army Research Institute for the Behavioral and Social Sciences ATTN: DAPE-ARI-IK 121 Morande Street Fort Knox, KY 40121-4141	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Research Institute for the Behavioral and Social Sciences 2511 Jefferson Davis Highway Arlington, VA 22202-3926	10. MONITOR ACRONYM ARI
	11. MONITOR REPORT NUMBER Research Report 1900

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES

Subject Matter POC: John T. Breidert

14. ABSTRACT *(Maximum 200 words)*

Across the spectrum of self-assessment research, a topic of debate concerns the accuracy by which individuals can evaluate their own performance. While some research has found self-assessment to be an effective measure, the majority typically found it to be an under- or over-estimation of actual performance. Although the accuracy of self-assessment has seen skepticism, benefits have been well documented. The current review is the result of examination concerning self-assessment accuracy and utility. The literature was searched to evaluate the ability of trainees/job incumbents/students to accurately report their level of ability or performance. Upon examination of the self-assessment accuracy literature, problems arose concerning terminology and differential utilization of self-assessment. This review reports that self-assessment, as currently used, is generally inaccurate; but given appropriate consideration of the moderating variables and clarification of terminology, self-assessment accuracy could increase. The Army should utilize a continuum of self-assessment, considering domain and skill level as determinant factors. Self-grading could be useful for the introduction and training of new skills. Self-impression may be useful for assessing Soldiers' confidence, self-perception of personality or traits, and continuous performance appraisal. Implementation of the continuum has potential to improve training quality and skill retention throughout the Army.

15. SUBJECT TERMS

Self-Assessment, Assessment, Training, Performance Appraisal, Confidence Assessment, Metacognition, Self-Grading

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON
16. REPORT	17. ABSTRACT	18. THIS PAGE			
Unclassified	Unclassified	Unclassified	Unlimited	36	Ellen Kinzer Technical Publications Specialist (703) 602-8049

Research Report 1900

Self Assessment: Review and Implications for Training

John T. Breidert
Western Kentucky University
Consortium Research Fellows Program

Jeffrey E. Fite
U.S. Army Research Institute

ARI-Fort Knox Research Unit
James W. Lussier, Chief

U.S. Army Research Institute for the Behavioral and Social Sciences
2511 Jefferson Davis Highway, Arlington, Virginia 22202-3926

May 2009

Army Project Number
622785A790

Personnel, Performance and Training Technology

Approved for public release; distribution is unlimited.

SELF-ASSESSMENT: REVIEW AND IMPLICATIONS FOR TRAINING

EXECUTIVE SUMMARY

Research Requirement:

The research addressed in this report focuses on the accuracy by which trainees are capable of accurately self-assessing their own abilities or performance. Self-directed training and ongoing evaluation of performance could prove beneficial to the Army at a time of prolonged deployments and less time at home for schoolhouse training. This leads to an increased need for self-directed recognition of training needs. Additionally, training in important domains such as adaptive thinking skills, with programs such as Think Like a Commander (TLAC), has necessitated live instructor assessment and feedback—a requirement that can be impractical and cumbersome at best. This report reviews the literature concerning self-assessment accuracy and utility and poses recommendations for harnessing accurate self-assessments to be utilized during training and performance appraisal during deployment.

Procedure:

The current review reveals the disparity in the literature concerning self-assessment accuracy and utility. The objective was to discover whether self-assessment is accurate enough to be utilized for Army training purposes. Military, psychology, education, business, and health literature was searched in order to evaluate the ability of trainees/job incumbents/students to accurately report their level of ability or performance and to determine if there were common moderating factors that contribute to inaccuracy.

Findings:

This review supports the assertion that self-assessment, as currently used, is generally inaccurate. Although some research has found self-assessment to be an efficient measure of one's ability or performance, the majority of research typically found it to be an under- or over-estimation of one's actual performance. However, with appropriate consideration of the moderating variables and clarification of terminology, self-assessment could be accurately utilized. There are five general variables affecting the self-assessment accuracy of participants. The five variables are ambiguity, skill level, accuracy learned, individual differences, and methodologies. The five variables are central to the effective utilization and training of self-assessment. If not taken into account, we would have to assume that self-assessment should be considered inaccurate in general.

A new and clearer way of discussing self-assessment is proposed as a continuum. The self-assessment continuum allows movement from end to end with regards to objectivity and specificity depending on the situational demands for type of assessment. On the most objective and specific end of the continuum lies self-grading; at the most subjective and ambiguous end of

the continuum lies self-impression. The continuum is an attempt to minimize and utilize the differing influences that moderating variables impose on the accuracy of self-assessment. The Army should implement training which involves and utilizes the continuum of self-assessment, between self-grading and self-impression.

Utilization and Dissemination of Findings:

The U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) is currently using the information in this report to produce self-assessments for battalion commanders before and after they complete End State: Commander Visualization at the Battalion Level training. A wide range of ARI and Army knowledge could be improved by a greater understanding of what makes self-assessment more or less accurate.

SELF-ASSESSMENT: REVIEW AND IMPLICATIONS FOR TRAINING

CONTENTS

List of Figures

SELF-ASSESSMENT: REVIEW AND IMPLICATIONS FOR TRAINING

Introduction

Many terms are used to describe the process of self-assessment. One could identify several, such as self-evaluation, self-grading, self-appraisal, confidence ratings, and ability judgments to name a few. Regardless of what term is used, self-assessment typically refers to an individual's evaluation of their own ability or standing on a given construct. The realms most commonly performing self-assessment research include the health, educational, business, psychology, and military fields. The U.S. Army has used self-assessment for indication of trainee progression through training packages such as Red Cape (Shadrick & Schaefer, 2007) and self-assessment based on after action review methodology (e.g., Mirabella & Love, 1998). Self-assessment can be used to monitor skill acquisition progress or necessity for further training, and can be used post-training to monitor shifts in performance or ability. Across the spectrum of self-assessment research, the topic of most debate concerns the accuracy by which individuals can evaluate their own performance. Although some research has found self-assessment to be an efficient measure of one's performance or standing on a given construct, the majority of research typically found it to be an under- or over-estimation.

Although the accuracy of self-assessment has seen much skepticism, the benefits have been well documented. A case study conducted by Strong, Davis and Hawks (2004) involved college students enrolled in two general education classes. The benefits of being allowed to self-assess their own grades were striking. Among those benefits were that the majority of students were more motivated to learn, understood material better, and that the class was rated as more enjoyable. Ulmer (2000) emphasized that critical thinking, which stems from reflective thinking, is a major determinant for transfer of knowledge. The general thoughts about self-assessment and its uses revolve around the idea that improvement can be actively pursued once individuals recognize their own weaknesses. If able to identify one's own weaknesses, reliance upon superiors in order to provide feedback becomes unnecessary. Research shows that feedback is an integral part of effective training and motivation (e.g., Arnold, 1976; Koka & Hein, 2003; Locke & Latham, 1990; Shoenfelt, 1996). In the military, feedback for leaders that is quick and accurate should translate into better leader performance, which will have an impact on unit performance and mission success. Providing accurate feedback starts with planned observation and accurate assessment (Reider, 2008). If an individual can accurately assess his/her performance, weaknesses are identified promptly and a trajectory toward resolution is set. Self-assessment during training can be of particular beneficence because not only does it teach the trainee to evaluate his/her own performance, but potentially frees the trainers/supervisors from the duty of evaluation.

The current review reveals the disparity in the literature concerning self-assessment accuracy and utility. The objective was to discover whether self-assessment is accurate enough to be utilized for Army training purposes. The literature was searched in order to evaluate the ability of trainees/job incumbents/students to accurately report their level of ability or performance. Typically, accuracy is determined by comparison of the participants' self-assessment with some external criterion or standard. Literature was not entirely consistent

regarding whether self-assessment was accurate, underestimated, or overestimated. Evidence suggests, however, that several moderating variables could be the underlying foundation for the apparent disparity.

There are five general variables affecting the self-assessment accuracy of participants throughout the literature. The five variables are ambiguity, skill level, accuracy learned, individual differences, and methodological problems/inconsistencies. The ambiguity of rating-items and/or criteria is the first variable discussed. The more specific and measurable the domain and criteria are, the more accurate the self-assessment is likely to be. The second variable is skill level of the participant. If the participant is an expert, they possess the ability to accurately self-assess because they are familiar enough with the domain to discern differences that novices would miss. The third variable relates to the skill level, but is more closely related to the training of the domain. It is the extent to which self-assessment has been taught as its own skill. Individual differences is the fourth variable that can be influential in the accuracy of self-assessment. Individual differences, identified in the literature and in this report, include metacognitive ability, self-monitoring, self-esteem, demographics, and gender. Methodological problems can be a crucial portion of the accuracy puzzle. The fifth variable concerns differences in methodology and analysis (such as the reliance on correlational analyses and percentage agreement along with the criterion problem) which present difficult hurdles to clear en route to criterion-related validity. The five variables are central to the effective utilization and training of self-assessment. If not taken into account, we argue that self-assessment should be considered an inaccurate form of assessment at the macro level.

Upon close examination of the literature concerning self-assessment accuracy, problems arose concerning terminology and differential utilization of self-assessment. Some studies regarded self-assessment an evaluation of one's own traits/personality/knowledge and general skills or abilities (e.g., Ackerman, Beier, & Bowen, 2002; Dunning, Meyerowitz, & Holzberg, 1989; Kruger, 1999; Vogt & Colvin, 2005). Others regarded self-assessment as a prediction of either how well one will perform, or how well one has performed in the past, in general terms (e.g., Barnsley et al., 2004; Castle, Garton, & Kenward, 2007; Dunning, Heath, & Suls, 2004; Kruger & Dunning, 1999, 2002; Krueger & Mueller, 2002; Metcalfe, 1998; Parker, Alford, & Passmore, 2004). Many of the aforementioned research studies referred to this type of self-assessment as confidence ratings. Still others utilized self-assessment as post-performance assessment of a particular job-task or training session (e.g., Brewster et al., 2008; Davis et al., 2006; Moreland, Miller, & Laucka, 1981; Sidhu, Vikis, Cheifetz, & Phang, 2006; Strong, Davis, & Hawks, 2004). Without a clear idea of what the process of self-assessment refers to, the discussion concerning the accuracy of that process is muddled.

Although the general conclusion is that self-assessment is inaccurate, different types of assessments yielded more promising findings. The findings of this review support the assertion that self-assessment as currently used is generally inaccurate; but given appropriate consideration of the moderating variables and clarification of terminology, self-assessment could become increasingly accurate. This report first outlines the differing views on the accuracy of self-assessment, followed by discussion of its moderating variables. The concluding recommendations are then discussed with reference to the moderating variables and a new perspective on terminology, with their implications for training.

The overwhelming majority of the self-assessment literature suggests that self-assessment overestimates true performance. However, some contend that self-assessment is accurate or it underestimates true performance. This section reviews the literature from each of the respective positions.

Self-Assessment Overestimates

According to an abundance of research, self-assessment overestimates actual performance. Undergraduate dental students overestimated their own competence when compared to instructor marks (Mattheos, Nattestad, Falk-Nilsson, & Attstrom, 2004). Carless and Roberts-Thompson (2001) examined self-, superior-, and peer-ratings of participants in a Royal Australian Airforce training course and found that self-ratings were more lenient than ratings by others. Sidhu, Vikis, Cheifetz, and Phang (2006) found that surgeons training for laparoscopic colectomy persistently overestimated their own performance as compared to trained raters. In a study that was meant to evaluate clinician's ability to assess their own competence of important terms in evidenced-based medicine, practitioners in Sydney, Australia also overestimated their performance (Young, Glasziou, & Ward, 2002). This was evident, even though at first glance one would assume the practitioners had self-assessed in a seemingly diffident manner. Even with a low estimate of competence to begin with, no participants showed a competence that exceeded their self-rating. Only one participant's self-rating met criteria for a positive predictive value of actual competence, which contrasts with the outward appearance of a modest self-assessment. The above are all examples of a participant or trainee giving more credence to their own performance than does the criterion measure, typically expert or supervisor ratings.

One way to describe the difference between what people think they know versus what they actually know is by use of the terms confidence and competence. The research regarding confidence versus competence typically shows that confidence overestimates competence. Confidence refers to the estimate of self-ability or skill. Confidence self-assessments can be administered pre- and post-training in order to evaluate an individual's perceptions about their performance. Competence refers to an individual's actual performance or standing on a given construct. This can be estimated by expert raters or by test scores. When confidence is rated using an instrument, it becomes a self-assessment of a given ability, skill, or performance. Castle, Garton, and Kenward (2007) compared confidence scores with competence scores for nurses, doctors, and health-care assistants. The participants completed a confidence questionnaire regarding their performance in basic life support. Their competence was measured by an algorithm for basic life support that was produced by subject-matter experts. Results indicated a significant difference between confidence and competence, generally due to over-confidence. In their conclusion, the authors stated that training and exposure could increase both confidence and competence. Barnsley et al. (2004) found similar results that showed an even more pronounced over-confidence effect. Their research involved Australian junior doctors and their assessments of confidence compared with actual competence for several medical competencies. For all competencies, confidence scores over-estimated actual competence. In

the context of the above studies, confidence as an over-estimation of competence is another way of saying self-assessment overestimates actual performance or standing on a given construct.

Dunning, Heath, and Suls (2004) reported in concurrence with the previous authors that self-assessment of skill level is typically overestimated. The researchers attributed this to the supposition that when people evaluate their own level of skill, they are overly optimistic about what they know and ignore what they don't know. That is because people are not aware of what they do not know. The ability to estimate what one knows, as well as monitor knowledge and learning is called metacognition (Everson & Tobias, 1998). The lack of metacognitive ability leads to estimations of above-average competence. Although thorough training will lead to automaticity, it can have positive and negative consequences. It does free cognitive resources for further interpretation of and reaction to stimuli, but it also inhibits the recognition of possible mistakes that could be critical if not addressed. Dunning, Heath, and Suls (2004) stated that even when people have the necessary information that would aid in accurate self-assessments, they will ignore or diminish it. That leads to over-estimation of skill. The authors outlined four "informational deficits" that cause overestimation, which are as follows:

1) Double curse of incompetence – people that are incompetent possess deficits of information that lead to errors and also block knowledge gained by these errors.

2) Unknown errors of omission – people are only cognizant of the solutions that they produce, not the solutions that they could have or should have produced.

3) Uncertain lessons from feedback – people assume they are adequate and inflate their perception of their own skill because feedback is often limited.

4) The ill-defined nature of competence – domains are often very general and vaguely outlined.

The fourth concept, discussed more thoroughly in a later section, concerns the ambiguity of task statements and provides support for the argument of specifically rated items within a self-grading assessment of training. The informational deficits give us insight as to why people are likely to overestimate their own performance as well as provide another instance of the massive literature claiming the inaccuracy and overstatement of self-assessment.

Self-Assessment is Accurate

There are a few studies that contend that there is nothing wrong with self-assessment in its current state. An example of this is an education study by Sullivan and Hall (1997) which observed very good agreement between lecturers and students ($r = .72$, $p < .01$). Students overestimated slightly more than underestimated their own grades. Seventy-seven percent of the self-assessments made by students in the study were within one grade level of the teacher's assessment. Students with higher grades tended to be responsible for any underestimations. The authors suggest that this effect could be due to regression toward the mean. Matthews and Beal (2002, p. 13) found that the Mission Awareness Rating Scale, a self-assessment for situation awareness, "may have general utility as an effective and user-acceptable measure of situation

awareness." A study of males participating in a military course by Fox and Dinur (1988) found that self-assessments were significantly related to commander and peer ratings. In that study, the experimental group was told that their scores would be compared with other data to test if this would increase accuracy. Although the experimental group did not rate with any statistically significant improvement, predictive and convergent validity was found for course success, commander ratings, and peer ratings. For all but one assessment, commanders' assessments of efficiency under pressure, differences between groups were insignificant ($t < 1.78$, $p > .05$). Another positive finding was that there was less of a halo effect for the self-assessments than for the peer ratings. While this literature presents an argument for the capacity to effectively and accurately utilize self assessment, other research suggests that this capacity is limited.

Self-Assessment Underestimates

A minority of studies have found that self-assessments underestimate performance. One found that medical students evaluated themselves more severely than the tutors grading them (Chur-Hansen, 2000). Another study found that trainers of educational registrars rate their own skills as lower than the ratings made by others (McKinstry, Peacock, & Blaney, 2003). Similarly, a dental study found that dentists rate their own work with more scrutiny than others (Milgrom, Weinstein, Ratener, Read, & Morrison, 1978). This effect may be explained by the effect of skill level and the difficulty or complexity of the task (see *Skill Level* section below). To preview, as skill level increases, accuracy increases. This is true until the skill level is well above average, where the trainee becomes more critical and less aware of his/her own expertise. Also, as difficulty and/or complexity of the task increases, the likelihood of underestimating performance of the task also increases (e.g., Barnsley et al., 2004; Kruger, 1999). Medical students, educational registrar trainers, and dentists are all highly skilled positions that most likely involve difficult, complex tasks. Self-assessment may be applied across various situations, and accuracy may be influenced according to different aspects of the situations. A number of variables that change across situations are likely reasons for differential determination of the accuracy of self-assessment. The variables are discussed in detail in the following section.

Variables Affecting Self-Assessment Accuracy

Self-assessment has a clear tendency to mis-evaluate true performance scores. Considerations must be made, however, before arriving to any clear conclusion. Ambiguity, skill level, and training effects are major elements which deserve attention. Other concerns, such as methodological issues and individual differences are also factors that could provide more insight as to why self-assessment accuracy is so elusive. The moderating factors may alter the perception of self-assessment as an inaccurate form of assessment.

Ambiguity

Ambiguity of statements refers to the level of specificity of the domain and/or items being rated by self-assessors. Evidence points toward the level of ambiguity as a determinant for how accurate self-assessments can be. Hayes and Dunning (1997) found that when the domain of possible traits is defined ambiguously, college students self-assess more generously than their roommates rate them. Dunning, Meyerowitz, and Holzberg (1989) found that when the traits

were given specific definitions, the ratings tended not to be so generous. It was further noted that self-assessments show more concurrent validity with other-ratings when the traits are well defined (Hayes & Dunning, 1997; Story, 2003). Presumably, specificity leads to decreased inflation of self-assessment by eliminating much uncertainty that stems from poor metacognitive ability. It gives the rater a concrete referent by which objectivity takes the place of speculative subjectivity. Consider the item, "Did you play a good game today?" The question sounds easy to answer, but the individual answering it may not consider every aspect of the game in making the decision. The metacognitive ability of the individual determines how well he/she answers the question. However, consider the item, "did you hit the ball-carrier with your inside shoulder with your head up, wrap up, and take him down within one yard of collision every time you were in position to do so?" The question leaves most metacognitive insight out of the picture and relies solely on matching performance with criteria. In this way, ambiguous items allow for poor metacognitive ability to obstruct accurate self-assessments, but specific items do not. Strong, Davis, and Hawks (2004) suggest that to decrease the amount of inflation in assessments, thus making them more accurate, students should be given written objectives used as a template or standard in order to determine final grades. The apparent inference in these cases is that in order to improve accuracy, the domain must be clearly and specifically defined and assessed using comparative standards for self-assessment items.

Again, studies show that clarity of the domain and standards of satisfactory performance aid in obtaining accurate self-assessment. Ackerman, Beier, and Bowen (2002) stated that when assessing one's own abilities, the likelihood of under- or over-estimation is dependent upon familiarity with the ability and how broad the domain is. If the individual is unfamiliar with the ability in question, they are likely to infer from other information how they would perform. For example, if an individual were unfamiliar with the domain of visualization in military operations, they may draw from other information that they assume to be relevant in order to self-assess. Consider the item, "I could learn to synchronize visualization across relevant external players." Because the ability in question is unfamiliar to most people, the respondent is likely to base their evaluation on assumed and irrelevant information, such as their ability to play videogames. If the domain is too broad, such as leadership, the inaccuracy of estimation is increased. The solution to this problem proposed by Ackerman, Beier, and Bowen is to utilize more specific measures during the self-report. The authors found that for the knowledge domains of science, civics, and humanities, broad items along with an unspecified comparison group resulted in over-estimation of ability compared with the more accurate estimation of ability when specific stimuli and absolute scales were used. For the knowledge domain of business management, however, the estimation of ability did not fluctuate as more specific stimuli and absolute scales were used.

Dunning, Meyerowitz, and Holzberg (1989) described the over-confidence effect as a product of self-serving assessments of ability. The underlying theme of their four-study series is that given a certain skill, ability, or characteristic, there are usually many definitions of what constitutes a high standing or good performance. An example would be the vague domain of leadership potential. If a person were asked to evaluate their own leadership potential, they would be able to conjure definitions that range from compulsive and task oriented to deliberative and people oriented, depending on the skills they themselves possess. The individual is apt to see a range of behaviors possibly linked to the evaluation item and select the most self-serving combination to provide their rating. In the first two studies, participants produced these self-

serving assessments when the trait being rated was ambiguous, or subject to interpretation given a wide domain of behaviors. In the third study, it was found that as more criteria were given to produce an evaluation, the participants identified more with both positive and negative characteristics. The fourth study demonstrated that using a list of specific criteria created by an outside source tended to decrease the participants' self-serving tendencies. The problems with unclear domains and standards of performance seemed to be common issues in these studies that could be resolved by using objective domains and standards.

In similar fashion to Dunning, Meyerowitz, and Holzberg (1989), Metcalfe (1998) studied over-confidence and used the term "cognitive optimism" in order to explain it. According to Metcalfe, over-confidence in self-reflection is due to an individual's thought process that takes incorrect information and treats it as if it were a correct predictor of performance. Metcalfe outlined seven metacognitive phenomena that lead to over-confidence:

1) People think they will be able to solve problems when they won't.

2) People are highly confident that they are on the verge of producing the correct answer when they are, in fact, about to produce a mistake.

3) People think they know the answers to questions when they don't.

4) People think the answer is on the tip of their tongue when there is no answer, or the answer is wrong.

5) People think, even when given contradictory feedback, that they produced the correct answer and that they knew it all along.

6) People believe they have mastered learning material when they haven't.

7) People think they have understood, although they are demonstrably still in the dark.

Metcalfe explained these phenomena as a product of self-deception, and memory-based processing heuristics. In self-deception, people are aware on some level that their answers are or could be incorrect, but convince themselves that they are correct. Because most of the time people have no evidence to negate their correctness, most over-confidence is not seen as self-deception. The other explanation was related to memory-based processing heuristics in which people base decisions about judgments upon information retrieved from memory and information at hand that is not entirely accurate, but is treated as if it were. A person assumes that the first decision arrived upon, assembled from memory and information at hand, is correct. An example would be if one were watching Jeopardy and confidently blurted out an answer based on recollection of WWII history. The answer may have been very close due to a decent amount of knowledge of the subject, but the person answering had no capability to decipher between the assumed correct answer and the actual correct answer until it was revealed. Evidence for the heuristic view comes from Oskamp (1965) who studied psychiatrists and psychiatric residents that were either given a small amount of information or were given a large amount of information regarding a patient in a hypothetical situation. The information was

irrelevant to any diagnosable situation. As the amount of irrelevant information increased, so did confidence in their diagnoses. This occurred even though the participants were correct only by chance. The conclusion was that the irrelevant information was the cause of an illusion of knowledge. Ambiguity in self-assessment might allow memory to create irrelevant information that yields a feeling of false confidence in the assessed domain. The illusion of knowledge could be minimized by using comparative standards for self-assessment that cut down on ambiguity, thus allowing for confidence that matches actual performance or knowledge within the domain, and increasing accuracy.

Ambiguity was addressed by Dunning, Heath, and Suls (2004) when they wrote that one of the informational deficits that cause overestimations during self assessment is the ill-defined nature of competence. For example, if the domain is chemistry, the solutions and knowledge are specific and either right or wrong. If the domain is essay writing, the solutions are many and structured knowledge negotiable. When the domain is ill-defined, people overestimate their skill; but if the domain is narrowed or specific, people are likely to estimate their skill with more accuracy.

The current state of the literature regarding ambiguity of domains and comparative standards says that the lower the ambiguity, the more likely a participant or trainee will be to give an accurate self-assessment. Familiarity with ability and domain breadth, self-serving assessments, and cognitive optimism outlined by metacognitive phenomena are all reasons given for why, but the central issue at this juncture is that low ambiguity yields accurate self-assessment. Because ambiguity of a domain is not under the control of instrument creators, it would be important to control the ambiguity of individual items. This would narrow the domain into one small portion (e.g., individual steps of a sequence, instead of the sequence as a whole). Of course ambiguity, or any other single factor, is not to blame for the entirety of self-assessment inaccuracy, but to give it proper attention could augment the utility of self-assessment.

Skill Level

Studies to be discussed in this section report the accuracy of self-assessment as moderated by skill level; people who are low in actual competency rate themselves above average. Also, people that are exceptionally gifted in some competency rate themselves lower than exceptional, but more accurately than those that are less competent. Numerous studies have found that self-assessment accuracy is determined by skill level, or domain expertise. Zakay and Glicksohn (1992) found support for their hypothesis that participants who were over-confident tended to produce more wrong answers in multiple-choice tests. Thus, the people who were over-confident were actually lower performers than those who were not so. Higher skilled participants, however, were able to produce more accurate predictions of their performance. Over-confidence can be the result of inexperience. As one faces the challenges that actually occur within a domain, they adjust their confidence accordingly. This lowers their confidence, while increasing their competence. Davis et al. (2006) studied the ability of physicians to self-assess, as it has been deemed a necessary skill for the ongoing education demanded in the field. A literature search that included self-assessment of ability and some performance measure for physicians found 20 studies that directly compared self-assessment with external performance measures. Analyses revealed that only seven showed positive correlations. Many of the studies

found that the worst self-assessments came from the lowest performers, who also rated themselves highest. Parker, Alford, and Passmore (2004) used an In-training Examination (ITE) in order to test whether family medical residents had the capability to self-assess. The residents completed an estimation of their performance prior to being assessed with the ITE. The results showed that the residents did not predict their skills with accuracy. Results further showed that the bottom and top quartile performers showed the biggest deficits with regard to accuracy. Top performers tended to underestimate, while bottom performers overestimated. It can be argued that over-confidence and under-confidence are both detrimental to performance and ongoing self-improvement. The ideal scenario for performance is perfect accuracy. However, to be over-confident leaves a person vulnerable to and ignorant of mistakes; to be under-confident motivates a person to increase their ability through training and self-improvement.

When a person becomes more skilled, they also become less over-confident, and although sometimes under-confident, they are more accurate. Randal, Ferguson, and Patterson (2000) defined self-assessment "when a candidate or job incumbent makes some evaluation of their own performance." The researchers tested the relationship between participants' ability to accurately self-assess and their achievement at an assessment center. Participants who were accepted (offered the job, meaning they performed better) rated themselves more accurately than those that were not accepted. Furthermore, unsuccessful participants over-estimated their performance on all exercises designed to assess competence in areas crucial to the job. The authors proposed that the successful participants possessed some ability to assess their own performance accurately, such as recognizing evaluation criteria, which allowed them to assess their own performance similarly to assessors.

An example of military research evidence in support of the increased accuracy effect of skill level comes from the development and validation of the crisis response training package called Red Cape (Shadrick & Schaefer, 2007). At the beginning of training, the authors reported that the participants were prone to overconfidence (or inflated self-assessments). As the training continued the participants reported, according to the completed Likert scales, decreases in their self-assessments. As their self-assessments steadily decreased, their performance did the opposite. This convergence from overestimation of ability to actual ability comes from training. The authors suggested the reason for the increased awareness of actual ability in the training situation came from, "increased awareness of the complexities involved in large scale interagency efforts." Increase in skill level in this case seems to increase accuracy by means of increased awareness of the domain of interest. Awareness of the domain of interest may be attainable by using specific, measurable, micro-components of the overall training objective. The individual behavioral components would convey complexity from the very first rating. The trainees would become experts in the individual components necessary for success and train specifically for weak areas.

Kruger and Dunning (1999, 2002) explained the relationship between skill level, (i.e., competence) and estimation of skill (i.e., confidence) using metacognition as the foundation for accuracy. As people become more competent, they are more accurately confident as opposed to over-confident. As described previously, metacognition is the process of knowing what one knows (Dunning, Heath, & Suls, 2004). This could be extended to include a person's ability to reflect upon one's own knowledge or behavior and continually assess it. As one becomes more

competent, one can more accurately assess performance because what good performance should look like is now known. This effect actually leads to slight under-estimation of performance for top performers because they don't necessarily feel as if their performance is as good as it should be. Krueger and Mueller (2002) contested Kruger and Dunning (1999), noting that the asymmetric errors due to differences in metacognitive ability is merely a statistical regression effect combined with the better-than-average effect. The better-than-average effect is that given a trait, most will think they are better than 50% of the population. This is a statistical impossibility because more than half of a population cannot out-perform half of the same population. Kruger and Dunning (2002) responded that the regression artifact cannot explain the entirety of the effect found in their original study, and that Krueger and Mueller used unreliable data to form their own conclusions. Because participants tended toward more accurate self-assessment as their skill level increased, there is merit in the idea that more skilled trainees are more likely to be accurate self-assessors due to some metacognitive abilities that are produced through training.

Metacognitive ability might explain the disparity between highly skilled people and unskilled people with regard to their self-assessment accuracy. Everson and Tobias (1998) supported the necessity of metacognition as a tool for learning new skills. Conducting studies of students' abilities to monitor their own knowledge and how this relates to subsequent grade point average (GPA), the two researchers found that achievement, especially in English, was better predicted by knowledge monitoring ability than by raw scores in the subject of interest. This is to say that the relationship of test A to test B, given an interval, was moderated by the variable of knowledge monitoring ability (i.e., Everson & Tobias' measure of metacognition). People with high knowledge monitoring ability increased scores significantly more from test A to test B than those low in knowledge monitoring ability. They also found that knowledge monitoring ability scores separated the students with high GPAs from the students with low GPAs. This gives support for the idea that the more skilled an individual is, the more likely they are good at gauging their own abilities.

A study by Hodges, Regehr, and Martin (2001) investigated the viewing of professional physician performance and its effect on novice physicians' ability to self-assess. The novice physicians performed, then rated themselves on job relevant tasks. The novice physicians were rated by experts and categorized into three ranked groups. All three groups' self-evaluations were significantly different. The middle group was generally accurate when their scores were compared to scores made by experts. The bottom group self-rated much higher than experts rated them. The top group self-rated lower than experts rated them. After viewing video a benchmark comparison physician, each group rated themselves again. The middle group remained accurate, the top group corrected their scores upward toward the scores of the experts, and the bottom group showed erratic corrections to their scores. Once again, higher skill was a better predictor of self-assessment ability. This also shows that self-assessment ability may be trainable in itself as a skill.

In a somewhat contradictory study, Moreland, Miller, and Laucka (1981) discovered that low-achieving students may have the capability to follow grading criteria in an objective manner, but do not apply the skill when grading themselves. This may validate the proposed self-serving assessment ability discussed by Dunning, Meyerowitz, and Holzberg (1989). Moreland, Miller,

and Laucka (1981) instructed students to grade their own work and the work of their peers on two tests, including a midterm and a final examination, and to describe the instructor's grading criteria. As predicted, low-achieving students did not grade their own work with accuracy, while high-achieving students did so. What was unexpected was that poor students graded their peers accurately and demonstrated an understanding of the criteria by which the instructor was grading. This may be an indication of a student's inability to grade his/her own work is due to metacognitive shortcomings, or it may be the mere fact that students grade with a self-serving bias. In other words, the students may be biased about their own grades even though they have the capability of grading objectively. If the criteria with which the trainees are grading themselves are specific and objective, there may be less room for these biases to take effect. In the next section, training of self-assessment as a skill is described as one way to boost accuracy in assessment and enhance training effectiveness.

Accuracy Learned

We have seen studies showing that as expertise or skill level increases, the ability to accurately self-assess does as well. The ability to self-assess may be applicable in assessments across domain boundaries. Literature suggests that the ability to assess one's performance can be taught as a skill. Without consideration to ambiguity of statements or the ability of the participant, there are many studies that report a general inclination toward more accurate assessment as trials proceed. Murphy (2008) regarded performance appraisal, saying that it is necessary to motivate raters to willingly rate themselves. This may be extended even further to training and self-assessment. The inaccuracy could be merely a result of motivational issues. If the assessor knows that the training is absolutely necessary for performance in the field, and that their accurate assessment of progress throughout is absolutely necessary for training, it would be more likely that the assessments would be more accurate. Other issues remain, but motivation is a central factor. In an influential meta-analysis of self-assessment in higher education, Falchikov and Boud (1989) identified several problems with self-assessment, including the nature of self-assessment itself and methodological issues. Their conclusion, however was that self-assessment can be developed as a skill. They state that no matter the rater or the rating situation, good assessment practice includes training of the assessors. Falchikov and Boud also found that those students in advanced courses were more accurate at self-assessment than those in introductory courses. This is evidence for skill level as important for accurate self-assessment, especially as a function of training. In another review of self-assessment in health professions training, Gordon (1991) found 14 studies which compared a self-assessment of ability with an external assessment. The study found that those high in ability tended to underestimate their performance and those low in ability tended to overestimate their performance. However, five training courses were identified within the studies that showed that self-assessment improved when it was conveyed as a skill which merited training with goals and feedback. This gives testament to the necessity and worth of training self-assessment in order to increase accuracy.

Self-assessment could be considered as generalizable across domains and trained independently as a skill. Fitzgerald, Gruppen, and White (2000) examined the accuracy of self-assessment across different task formats. Their study is of particular interest because they found that self-assessment accuracy was constant regardless of the different task formats. One format was performance-based, and the other was cognitive-based. This study offers support for the

idea that self-assessment is a skill to be learned during training, and that training of accurate self-assessment in one area can generalize to other areas as well. Also, students were accurate assessors of their own performance when they estimated against an objective standard. This objective standard offers credit to the assumption that self-assessment is more accurate when graded against specific, objective, measurable criteria. Schraw (1997) also found what he described as "generalized metacognitive knowledge" regarding students' performance and confidence judgments on tests. In a repeated measures design, Schraw gave four separate tests to students and contrasted correlations between confidences in test responses with correlations between actual scores on the tests. Four tests measuring four separate constructs, each with a confidence scale, were used. While performance on the test was not consistent, confidence judgments were consistent. This suggested a general metacognitive process that extends across specific domains. This strengthens support for the idea that self-assessment is a skill in itself and is measurable and trainable independent of specific domains.

A fair portion of literature suggests that self-assessment can be trained and that feedback and objective standards are the key to that training. In a review of the reliability, validity, and utility of self-assessment, Ross (2006) concluded that the greatest impact on the utility of self-assessment comes from the training of students in how to assess their own work. MacDonald, Williams, and Rogers (2003) found that surgical trainees estimated their time and errors made during a simulation trainer with more accuracy as repetitions increased, learning both how to self-assess and how to perform the task. Taras (2001) found that feedback from tutors was a critical part of the process of self-assessment, and that "learning to learn" is an essential portion of education. Andrade (2003) suggested that the way to improve student essays is to provide a rubric and two self-assessment sessions. Edwards (2007) found that after grading their first exam, students improved their grading ability for the second exam. A contention could be made for regression toward the mean, but this is still one more study augmenting support for training self-assessment. Students reported enjoying the self-grading procedure because they were able to learn from their own mistakes and obtained very timely feedback. The students liked that they had a key that gave them exactly what was expected of them and that they were given the trust and responsibility for their own learning. Thus, feedback through objective comparison with standards has proven a valuable asset to accuracy of self-assessment, and has merit as a motivational force.

A sizable self-assessment accuracy gain comes from the correction of common self-serving tendencies through training. It is by the extinction of these tendencies that one can truly become an objective rater. One study found that medical residents' self-assessment improved merely after watching a videotape of their own performance (Ward et al., 2003). Perhaps the videotape allowed a more objective view of performance outside of the biases that are inherent in reflection of the self. The videotape in this instance provided feedback necessary to increase accuracy. Another study involving medical students showed that, although self-assessments were overestimates of skill throughout training sessions, accuracy increased throughout (Das, Mpofu, Dunn, & Lanphear, 1998). The increase in accuracy was attributed to tutor feedback and communication that was geared toward addressing the differences between the tutor evaluations and the self-assessments. Because self-serving biases are habitual, and possibly instinctual, they would be expected to return in time following their extinction. It is because of this that remedial

training for self-assessment would likely be beneficial; just as other skills are continually maintained and developed through refresher training.

The difficulty of the task also seems to play a major role in the accuracy of self-assessment. Barnsley et al., (2004) found that for every competency measured for Australian junior doctors, confidence was higher than actual competence. It was further noted that this over-estimation was particularly strong for tasks that, for doctors, are considered simple. People disregard how skilled their compared peers are when judging how they measure against them. Kruger (1999) found that for easy domains, people thought they would do much better than their peers because they themselves perform well. The participants did not take into account that most people can perform well in this domain. On the other hand, for difficult domains, people underestimated themselves regarding their relative ability to their peers. They did not tend to account for the fact that it was a difficult domain and that most people would perform accordingly. To inform participants through frame of reference training as to what constitutes good performance would be a remedy for participants' inability to distinguish their relative abilities. Also, education concerning the self-serving bias and/or attribution errors could be beneficial.

Self-assessment becomes more accurate with more practice, just like most trainable skills in most domains. It can be generalized across domains, increasing in its own right without regard to the domain being assessed. People learn to become objective observers of the self through practice and comparison to some feedback-generating standard. This feedback could be a videotape of themselves, peer or expert rating, or standard rating criteria. The efficiency of self-assessment as a feedback tool can reduce the necessity for videotape, or peer/expert-rating. The feedback garnered through reflective comparison with standard rating criteria is the best means for accurate self-assessment, and can be used for the simultaneous training of self-assessment skill and the domain of interest. In order to train for self-assessment, individual differences that could affect that training should be taken into account.

Individual Differences

The general consensus of literature reports self-assessment as being biased toward inaccuracy. Criteria or domain ambiguity, skill level, and task difficulty have been described as contributors to the inaccuracy dilemma, but there are others to examine. Individual differences may be at least partially to blame for miscalibration of self-assessment. Self-assessment skill itself may in fact be an individual difference inherent in some trainees. Some individual differences reported in the literature to date include metacognitive ability (Pallier et al., 2002), self-monitoring (Cutler & Wolfe, 1989), self-esteem (Ehrlinger & Dunning, 2003; Wells & Sweeney, 1986), demographics (Brutus, Fleenor, & McCauley, 1999), and gender (Hassmén & Hunt, 1994; Pallier, 2003). Although we will explore the individual differences that may influence inaccuracy of self-assessment, one must keep in mind that individual differences such as self-monitoring, self-esteem, demographics, or gender are very difficult, if not impossible, to change. Therefore, as far as recommendations are concerned, we should keep the individual differences in mind when interpreting self-assessments, but it is very difficult to eliminate them outside of research settings. If self-assessment can be trained, then it can be trained for everyone. The results will be that as training progresses, self-assessment will become more

accurate regardless of any individual differences. Trainees will improve most if they possess the individual differences leading to inaccurate self-assessment. However, the purpose of training would be to improve all participants; from bad to good, and good to expert.

Individual differences approaches toward self-assessment accuracy (and inaccuracy) say that some individuals have an overall tendency to either over-estimate or under-estimate their knowledge, skill, or ability regardless of the given domain (Pallier et al., 2002). Soll (1996) investigated confidence judgments for the probabilistic mental model, a questionnaire that obtains a response and confidence in that choice within the same instrument. He found that while some individuals tend to be over-confident, others tend to be under-confident. One of the groups involved in the study had a mean bias score (inaccuracy score) of 30%, while another group had a mean bias score that was less than 10%. Another study by Stankov and Crawford (1996) found that the mean bias rating for the entire sample can show an effect of over- or under-confidence, and up to 30% of the participants in that sample can have an opposite effect. Pallier et al. (2002) found over two experiments that there was a general confidence factor that occurred without regard to which cognitive ability test was administered or how well the participants performed. The researchers did not find personality traits or cognitive ability as determinants of self-assessment. Their conclusion decision was that there are many causes for inaccuracy for self-assessment, and that models of inaccuracy do not account for them. Self-monitoring is another individual difference that should be regarded as contributing to inaccuracy of self-assessment.

Self-monitoring is a construct related to personality that may be an individual difference necessary for understanding inaccuracy. Self-monitoring is the extent to which a person is concerned with presenting themselves as confident, and socially desirable (Cutler & Wolfe, 1989). Cutler and Wolfe (1989) hypothesized that because low scorers on self-monitoring would be less confident about their answers, their accuracy would be better. Results showed that people who were higher on self-monitoring did have higher confidence scores. The people with low self-monitoring were more accurate in their confidence scores, but not significantly more accurate than high self-monitors. Results suggested that high self-monitors, when given a social cognition task, are highly confident, but may be less accurate in their self-assessments. In short, low self-monitors are more accurate at self-assessment. While these results are overly specific to this domain, they do signal that individual differences deserve at least some merit concerning accuracy in self-assessment. Another individual difference deserving some attention is self-esteem.

Self-esteem reflects a person's overall evaluation or appraisal of his or her own worth. If an individual holds a low amount of self-esteem, that individual would assess his/herself in a conservative manner. The opposite could be said about those individuals with high self-esteem; they would assess themselves in an excessively optimistic manner. Results obtained by Wells and Sweeney (1986) support this claim. Their study was of 1,508 male high school students that completed ability tests, self-assessments of those abilities, and self-esteem and self-esteem stability measures. Results showed that there was a positive relationship between self-esteem and self-assessment. When using self-esteem in this context, the authors separate self-esteem from task confidence. Self-esteem in this sense refers to the baseline measurement of a person’s evaluation of his or her own worth. According to Wells and Sweeney, the self-consistency

theory says people tend to rate themselves in a way that is consistent, and rely upon their normal level of self-esteem to determine the assessment. Thus, self-esteem and the varying levels of it that different trainees possess could be a determinant of how accurate they are.

Almost parallel to self-esteem is self-view. Ehrlinger and Dunning (2003) describe self-view as "evaluations . . . individuals chronically hold about their abilities, ones they hold before they even begin a task." The authors offer that instead of feedback that constantly changes the self-view, the feedback is translated to conform to the previously held self-view. They cite Ditto and Lopez (1992) stating that people tend to maintain a positive self-view, making the generation of negative self-generated feedback difficult. Positive feedback is accepted without hesitation, whereas negative feedback is examined with skepticism. Thus, negative feedback is much more predisposed to being discarded. Ehrlinger and Dunning set out to discover if self-views were important determinants of self-assessment accuracy, and whether changing the self-view would in turn change the self-assessment. What the authors found was that self-view did correlate well with participants' normative self-assessments, but not with their actual performance ranking. Normative self-assessments are those that involve self-assessment of performance in relation to others. They also found that when a person's self-view was altered, their self-assessment changed in the direction of the self-view alteration without changes in performance. An example provided was self-view of geography knowledge. When participants were manipulated to hold a more positive self-view, they thought they did better on the geography test as compared to those who were manipulated in the opposite direction even when performance stayed constant. Ehrlinger and Dunning also found gender differences for self-view. Men possess higher self-views than women in the field of science, even though they performed equally. Gender, discussed next, is an individual difference that could relate to some discrepancy regarding accuracy of self-assessment.

Demographic characteristics, specifically gender, have been a topic for discussion. Regarding multi-source ratings that included the self-assessment, peer assessment, and subordinate assessment, several intriguing results were found. Brutus, Fleenor, and McCauley (1999) found that female managers' self-assessments were accurate according to peers and subordinates, while male managers provided overestimates of performance, also compared to peers and subordinates. According to their report, another trend was for older managers, more than younger managers, to overestimate their performance compared to supervisor ratings. Another interesting finding involved levels within an organizational hierarchy. When managers fell into lower organizational levels, they were more likely to underestimate their performance compared to subordinate and peer ratings. When they fell into higher organizational levels, they were more likely to overestimate their performance compared to subordinate and peer ratings. Gender yielded differential self-assessment accuracy in young undergraduates (Pallier, 2003). For all cognitive tests administered, men recorded higher confidence scores than women. The cognitive tests were measures of visualization tasks, crystallized intelligence tasks, and fluid intelligence tasks. Men and women did not have significant cognitive score differences. Because of this, men overestimated ability for the tests of fluid intelligence and crystallized intelligence. According to Pallier, visualization often produces underestimation of performance. The men were better calibrated for the tasks that involved visualization because these tasks produced higher scores. The women in this case were likely to underestimate their performance, while men were accurate. When implementing self-assessment, we should note the inherent

differences that accompany gender and implement enough training in self assessment that inequalities can be minimized. The individuals poor at self-assessment will be trained to be more accurate, and those that are already proficient self-assessors will be assisted in fine-tuning their skills. In this way, regardless of pre-training self-assessment levels or individual differences, the performance gap should be narrowed. This is a case for gender as an individual difference to consider when implementing self-assessment.

The individual differences discussed here may influence the accuracy of self-assessment. With the knowledge that individual differences could determine, at least partially, the value of a self-assessment, it is important to minimize these differences through training and methodology such as counterbalancing. Counterbalancing would not necessarily aid the individual learner with self-assessment skill, but would allow the researchers and training designers to recognize the efficacy of their current practices. We have already seen that self-assessment can be trained as a skill. Differences in self-assessment skill can be minimized through increases in accuracy for each individual, regardless of metacognitive ability, self-esteem, age, or gender. The training of self-assessment as a skill using specific, behavioral criteria in a self-grading manner is discussed further in the concluding proposal; but it is possible that individual differences could be reduced to insignificance if sufficient attention to training self-assessment is given.

Methodological Problems

Although we have seen that self-assessment does not generally appear to be accurate, we must look into the problems inherent with measuring its accuracy. Problems involve methodology, statistical analysis, and theoretical issues. Ward, Gruppen, and Regehr (2002) found so much methodological inconsistency with regard to the measurement of self-assessment accuracy that they doubted the literature's ability to validate conclusions. Their first issue was that most studies included in their analysis relied on correlational analysis and percent agreement. This effect can be skewed by any alteration of the rating scale. Agreement is measured differently when it could mean the same score on a 100 point scale, or it could mean the same grade on a five point grading scale. The best data is that which could be considered at least interval as opposed to the ordinal nature of the A-F grading scale. Regardless, percent agreement should be used sparingly as it has a tendency to skew the results. There exists no great alternative in self-assessment methodology to the use of correlational analyses. At the very least, the adept statistician should examine the effect of outliers on results in order to interpret results accordingly.

When the external standard being utilized to compare with self-assessment is an expert rating, the criterion problem is intensified. In this case, the criterion is fraught with unreliability. Osterman (2007) stated that any research into the validity of intervention that relies on performance assessments of any kind as the criterion may not yield results worth interpretation. Clearly, progress is necessary on the assessment front. Already noted is that self-assessment is more accurate when specific, behaviorally observable criterion are used for items to be rated. The same has been found true for expert raters, and likely for peer raters. When they rate short, structured, simple tasks, experts are more likely to demonstrate agreement (Martin, Regehr, Hodges, & McNaughton, 1998; Regehr, Hodges, Tiberius, & Lofchy, 1996). When examining self-assessment accuracy using expert-rating as a criterion, the unreliability of the criterion

should be addressed. Regehr et al., (1996) proposed using the 'correction for attenuation' formula in order to account for the unreliability of the criterion (i.e., expert ratings) when finding the correlation between self-ratings (i.e., non-expert ratings) and expert-ratings. They advocated the use of the single correction for attenuation. It is the raw self-expert correlation divided by the square root of the expert inter-rater reliability (Muchinsky, 1996). Muchinsky's single correction for attenuation formula is:

$$\rho_{xy} = \frac{r_{xy}}{\sqrt{r_{yy}}}$$

where ρ_{xy} is the corrected validity coefficient, r_{xy} is the obtained validity coefficient, and r_{yy} is the reliability of the criterion. In their study, Regehr et al., (1996) found that when expert ratings were corrected for attenuation, the self-expert correlation increased from .43 to .58. This is a mathematical solution, however, to a fundamental problem that could be corrected for by decreasing ambiguity.

Individual self-assessments are often thought of as one set of scores to be interpreted at the group level. Ward, Gruppen, and Regehr (2002) point out that in order to assume similarity of scores, we must also assume that each trainee is assessing the same exact construct. If group-level means are compared to an external criterion such as expert ratings or test scores, the interpretation has to be at the group level. At this level, individual correlations cannot be validly interpreted. Group-level examination poses a problem because outliers throw off the entire sample's assessment accuracy. The best interpretation of data is at the individual level, which can be aggregated if necessary to prove a more broad argument. The problem, again, is that we cannot assume all individuals assess the same constructs. So, we either relinquish individual interpretation of scores, or we assume the same construct is measured by each individual. This is a problem that has not met a viable solution, and should be addressed in future studies.

Further research shows that expert raters may not be trained well enough to serve as the criterion for self-assessment validity. Vogt and Colvin (2005) found that correlations between self-assessors and their parents were stronger than correlations between self-assessors and trained behavioral coders. This is a demonstration that either the "experts" were rating incorrectly, or the rating scale was not explained or taught well enough to the parents and self-assessors, so that they were making similar mistakes in their assessments. In order to make sure the same construct is rated, and scales are used in a standard way, precise anchors could be assigned for the assessment criteria (Ward et al., 2002). Also, frame of reference training could be provided to make sure the trainees are attuned to appropriate scores for corresponding levels of performance. The frame of reference should be from the perspective of the organization (e.g., military doctrine), formulated by research that points to a level of satisfactory if not optimal performance. A recommendation made by Eva, Cunningham, Reiter, Keane, and Norman (2004) to further enhance the correlation between self- and other-assessments, is to utilize constructs that are easily tested (e.g., factual knowledge) as compared to constructs that are difficult to test (e.g., conceptual knowledge). This is not always feasible as training does not always fit easily tested arenas, but is more related to difficult concepts such as visualization and situation awareness. With the right frame of reference and clear operational definitions for

measurement, self-assessments should show a stronger correlation with the expert assessments than previous literature suggests.

Researchers will continue to disagree as to the best utilization of analyses and methodologies. The criterion problem concerning expert raters can be approached with the same remedy as proposed in this report. When specific criteria are being used, and those criteria are behaviorally measurable, the experts become more reliable. The experts should indeed be trained in assessment before being considered reliable. The suggestion by Eva et al. (2004) to use constructs that are not difficult to test such as conceptual knowledge avoids the necessity of defining a clear domain. To train a certain skill such as visualization or situation awareness, it must be broken into trainable components. Thus, if the domain is trainable, it is inherently assessable by use of the same componential behavioral criteria that is used for that training. Individual behavioral criteria in conjunction with frame of reference training and explicit anchors could prove to increase the accuracy of self-assessment along with the expert assessment, diminishing the criterion problem as well as providing more effective feedback for trainees.

Concluding Recommendations

In this report, we have encountered cases where self-assessment can be accurate and useful, as well as cases where self-assessment was inaccurate as an under- or over-estimate of performance. The majority of cases reported that self-assessment was inaccurate, and, of those, most reported over-estimation of performance. This inaccuracy can be explained by ambiguity, skill level, training effects, individual differences, and methodological problems. Although not the entire causal foundation for inaccuracy of self-assessment, inferences about improving assessment skill can be made. From these inferences we have compiled recommendations regarding the utilization of self-assessment for training purposes. By utilizing objective, reflective, highly-specific criteria generated to train and improve the ability of a trainee in both the domain in question and self-assessment as a skill, the moderating variables that relate to inaccuracy can be either suppressed or used as a basis for sound training. Here we will outline the recommendations for a new, clearer approach to terminology regarding self-assessment, followed by recommendations for each moderating variable, and a brief conclusion.

Utilize Appropriate Terminology

The self-assessment literature uses a litter of terminology that confuses a concise and functional definition of self-assessment. In order to clarify this confusion surrounding self-assessment, we have narrowed the field of terms to three. It is appropriate here to distinguish between self-assessment, self-grading, and self-impression. Most literature reports self-assessment as an estimate of how skilled/competent one is regarding a particular skill, ability, or characteristic. We propose two new terms with regard to their utilization. Self-grading is the assessment of one's own performance according to some objective scale. Self-grading can be the grading of one's own responses to items on a test or learning check, or grading of past performance. Self-impression is the overall intuitive judgment of how skilled/competent an individual feels regarding a construct. The disparity here is whether the evaluation is subjective or objective. Regarding self-impression, the evaluation can be very subjective because the individual rates his/her own perceived knowledge or standing on the construct in question. These ratings are sometimes called confidence scores (Leopold et al., 2005). In contrast, the

evaluation for self-grading is less subjective because the individual rates actual performance with regard to some objective measurement system based on standards of performance set by an organization, hopefully using non-arbitrary metrics. Self-grading is the most useful form of self-assessment for new trainees because of its ability to garner accurate self-knowledge. It is only after self-grading improves that we should see a concurrent improvement in self-impression. Self-impression may be more useful for refresher training or trainees that have advanced in training, as well as in the environment in which the training will be ultimately put to test. This is because self-impression is an intuitive judgment, much more practical than a continually evaluated list of individual criteria. Self-impression serves as an accurate evaluation once skills and assessment have been engrained through self-grading. The final picture of self-assessment becomes a continuum with self-grading on one end, and self-impression on the other (see Figure 1). Along the continuum may be varying levels of how specific or ambiguous, objective or subjective the self-assessment may be.

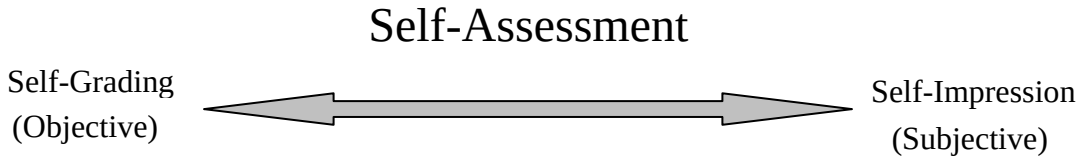

Figure 1. Self-Assessment Continuum.

Minimize Ambiguity

The section of this review concerning ambiguity in rating statements and criteria gives an explanation for why many individuals rate themselves in an inflated manner. It is due to incomparable standards and lack of behavioral criteria. This leads to self-impressions that are egocentric due to lack of restraint or guidelines. In order to correct for this effect of egocentric ability judgments, it is important to give the trainee an appropriate level of specificity. This will limit the ability of the trainee to exaggerate their performance. By using specific and behaviorally measurable criteria, the trainee would not be making subjective normative (compared to others) judgments, but would be making individual assessments about how well he/she performed each behavioral circumstance. Specific, behaviorally measurable criteria would include a series of items such as, “Did you hit the ball-carrier with your inside shoulder with your head up, wrap up, and take him down within one yard of collision every time you were in position to do so?” After making these micro-judgments for several iterations of training, the trainee will be able to make a blanket macro-judgment, allowing for increased ambiguity while maintaining accuracy. The number of iterations a trainee will need in order to do this is unknown and may rely upon natural metacognitive differences; further research is necessary. The individual criteria become engrained as a greater schema that makes up the overall assessment of performance. This is when the individual criteria check-lists or evaluations can become less intensive. The trainee will be able to make constant judgments of their overall performance that can be indicative of training needs, providing a more practical utilization of self-assessment. An example of a macro-judgment is the answer to the question, “How well did you play defensively?” The continuum between the self-grading and self-impression implies that at any time, the acuity by which one self-assesses can be adjusted according to how skilled they

remain. When their overall performance is deemed unsatisfactory, the trainee will be able to address the individual criteria when needed. The problem can then be identified and corrected. To minimize ambiguity for training purposes would increase the feedback accuracy which would, in turn, facilitate utilization of self-impression once mastery of a given skill has been achieved.

Consider Skill Level

Skill level is another relevant consideration when striving for improved self-assessment accuracy. The basic conclusion is that as an individual demonstrates more skill or aptitude for a certain behavior or domain than others, that individual is likely to be a more accurate self-assessor. The inference is that skilled individuals have metacognitive abilities that less skilled individuals do not. Compared with a highly skilled person, a less skilled person has less actual competence, but as much, if not more, confidence in the self. The difference between the two is a set of biases which inhibit the insight necessary to realize areas of weakness. Areas of weakness should be recognized and seen as a motivating force for further training. If these trainees were to use specific, objective criteria to obtain a self-grade, as opposed to self-impression, their ratings would reflect more accurately their actual performance as opposed to biases that fluctuate as a function of perceived competence. In this manner of self-grading, the ability to accurately rate oneself should not be strongly moderated by skill level.

The individual high in skill is likely more accurate at self-impression than the individual low in skill, but they should not differ with respect to accuracy when the domain is behaviorally broken down and compared to a set of clearly defined standards. Of course some domains are going to be harder to objectively break down due to their complexity and the inability to articulate skills. If, though, proper means of task analysis are used, the best approximation of behavioral and cognitive processes will be revealed. If the behavioral and cognitive processes can be articulated, then they can be individually rated. In this way, both experts and novices both have the tools available, and will be able to properly use them. After training occurs, the individual that began low in skill should be both more skilled, and better at self-impression. Increases in confidence should coincide with increases in accurate self-assessment. A valid self-assessment would be one that correlates highly with an external criterion. It is predicted that the validity of self-impression increases as expertise increases from low to high; there is a positive correlation. However, due to the specific, behavioral nature of self-grading, its validity should be as high for the individuals with low expertise as with high expertise (see Figure 2). The individual practicing self-grading should show high correlations with external measures of actual performance. Because skill level and self-assessment accuracy are positively correlated, we propose that the concurrent training of skill in the domain of interest and self-assessment (through the recurrent use of self-grading) would be an efficient training strategy.

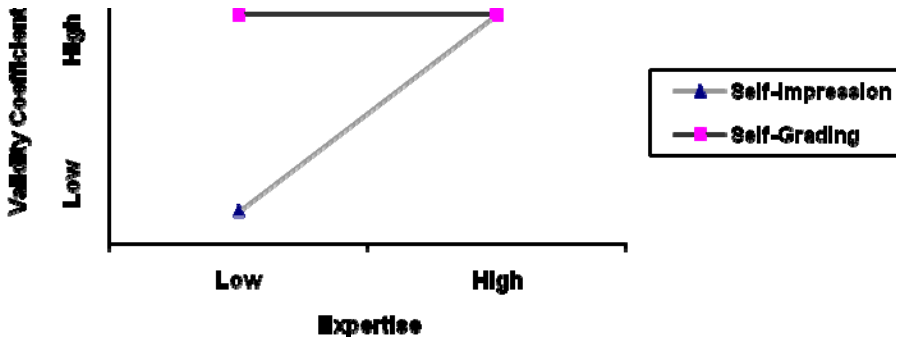

Figure 2. Hypothetical validity graph of anticipated results comparing self-impression to self-grading at differing levels of expertise.

Train Self-Assessment as a Skill

Similar to the argument that underlies skill level is that self-assessment is a trainable skill. The literature supports a positive relationship between self-assessment accuracy and number of trials (Andrade, 2003; Edwards, 2007; MacDonald, Williams, & Rogers, 2003; Ross, 2006; Taras, 2001). That self-assessment is a skill that is transferable and independent of the domains being trained (Fitzgerald, Gruppen, & White, 2000; Schraw, 1997) gives support to the idea that care must be taken to instill this skill actively through the use of self-grading. Self-assessment of any domain, however, does presuppose that criteria be given to the trainee in order to objectively apply such assessment skill. We postulate that the instinctual post-performance self-impression can only be highly valid if the participant is both trained in self-assessment, and trained in the domain of interest. Self-assessment can be trained utilizing objective rating scales with behavioral measures through repeated trials of self-grading (number of trials needed depends on trainee factors such as skill level and individual differences). Upon completion of training, self-assessment will be ingrained through practice of utilizing the individual criteria of the self-grading process. The self-grading components can then be aggregated to fit an instinctual, instantaneous self-impression of performance (see Figure 3). In this way, one can now picture self-grading of a particular skill transformed through training into self-impression that can be readily continually during missions or future training.

Consider Methodology and Individual Differences

According to methodological and individual difference problems inherent with self-assessment in general, it is recommended to follow procedures that minimize their effects. These procedures include frame of reference training that gives trainees the correct view of good performance, precise anchors with behavioral comparisons that reduce ambiguity, and diligence in choosing the criterion for establishing criterion-related validity. It is recommended that, when choosing participants for any kind of self-assessment accuracy or validity study, individual differences are controlled for by counterbalancing, random selection, random assignment, and/or statistical controls. It should be recognized that assessment in any form or fashion is fraught with confounds such as these, but through diligent research and understanding possible effects of these confounds, accurate conclusions can be attained. Through the training of self-grading as a

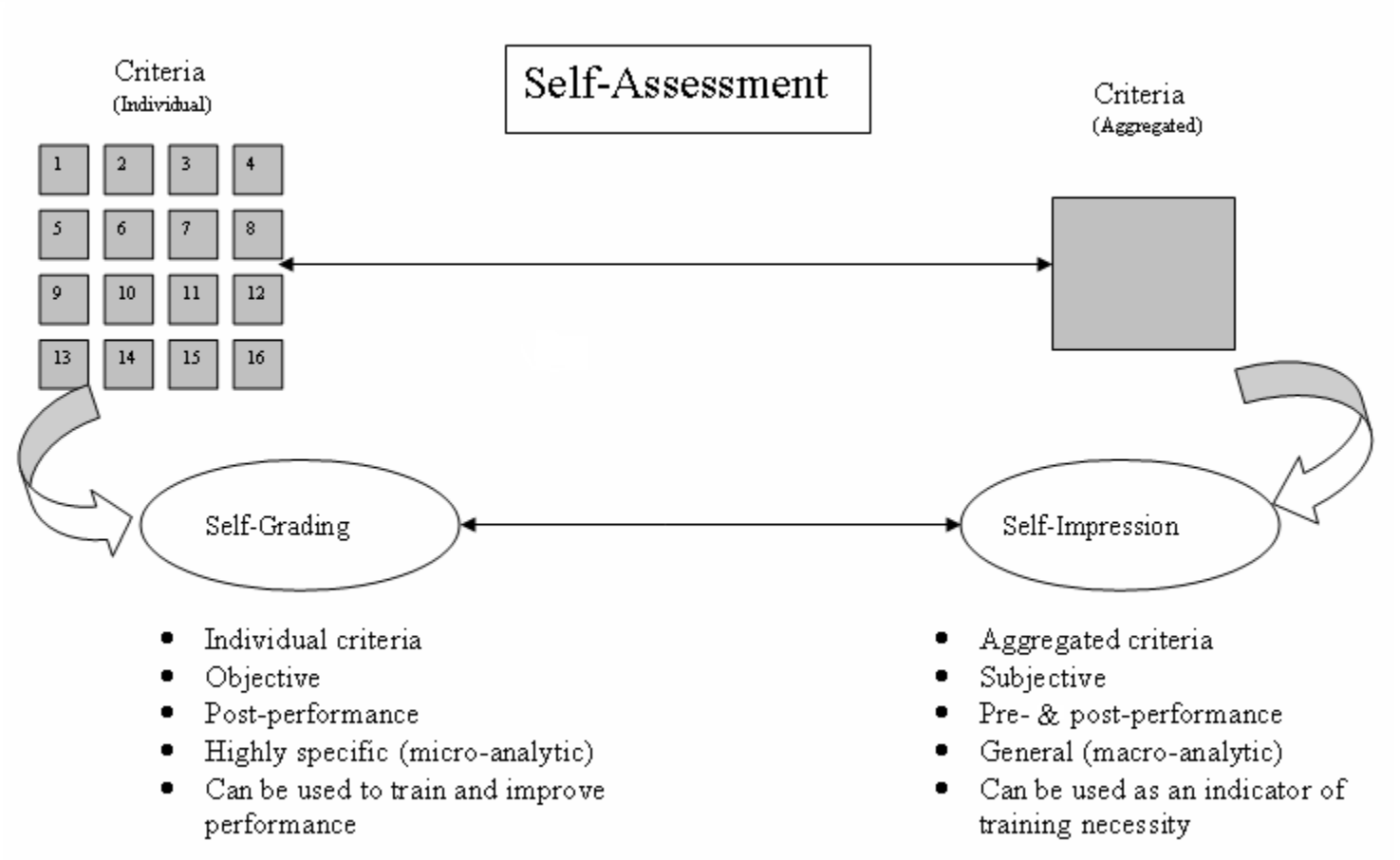

Figure 3. Self-assessment continuum contrasts the individual nature of self-grading criteria with the aggregated nature of self-impression criteria.

skill, individual differences and methodological problems will be minimized through objectivity and increased skill for all participants.

Concluding Remarks

In today's Army, the ability to self-assess is more important than ever due to increased deployment times and less time at home for schoolhouse training. This leads to an increased need for self-directed recognition of training needs. Up to now, the accuracy of self-assessment has been underwhelming at best. The reason for this may be due to the misuse of self-impression when self-grading would be better suited. The terminology we propose in this paper is an attempt to clear the self-assessment waters that have been muddied by similar terms measuring self-assessments with varying degrees of specificity and objectivity. In the future, we intend to further this effort by evaluating the strength of association between the variables discussed (i.e., ambiguity, skill level, training, and individual differences) and the accuracy of self-assessment. We also intend to evaluate the effectiveness of self-assessment training as it generalizes across domains. Specifically, we want to know whether Soldiers – from junior enlisted Soldiers to senior officers – can be trained to accurately assess their own performance and recognize training needs.

The Army should implement training which involves and utilizes the continuum of self-assessment, including self-grading and self-impression. Consideration for the situation and skill level should dictate the point within the continuum that would be most appropriate for a given training assessment. Self-grading could be useful for the introduction and training of skills that have not yet been mastered. Self-impression may also be useful for assessing a Soldier's confidence or self-perception of personality or traits, as well as assessment of performance once self-assessment of a particular domain has been mastered. Overall, we feel the implementation of the continuum has potential to improve the quality of training and skill retention throughout the Army hierarchy.

References

Ackerman, P. L., Beier, M. E., & Bowen, K. R. (2002). What we really know about our abilities and our knowledge. *Personality and Individual Differences, 33*, 587-605.

Andrade, H. G. (2003). Role of rubric-referenced self-assessment in learning to write. *The Journal of Educational Research, 97*, 21-34.

Arnold, H. J. (1976). Effects of performance feedback and extrinsic reward upon high intrinsic motivation. *Organizational Behavior and Human Performance, 17*, 275-288.

Barnsley, L., Lyon, P. M., Ralston, S. J., Hibbert, E. J., Cunningham, I., Gordon, F. C. (2004). Clinical skills in junior medical officers: A comparison of self-reported confidence and observed competence. *Medical Education, 38*, 358-367.

Brewster, L. P., Risucci, D. A., Joehl, R. J., Littooy, F. N., Temeck, B. K., Blair, P. G. (2008). Comparison of resident self-assessments with trained faculty and standardized patient assessments of clinical and technical skills in a structured educational module. *The American Journal of Surgery, 195*, 1-4.

Brutus, S., Fleenor, J. W., & McCauley, C. D. (1999). Demographic and personality predictors of congruence in multi-source ratings. *Journal of Management Development, 18*, 417-435.

Carless, S. A., & Roberts-Thompson, G. P. (2001). Self-ratings in training programs: An examination of level of performance and the effects of feedback. *International Journal of Selection and Assessment, 9*, 217-225.

Castle, N., Garton, H., & Kenward, G. (2007). Confidence vs. competence: Basic life support skills of health professionals. *British Journal of Nursing, 16*, 664-666.

Chur-Hansen, A. (2000). Medical students' essay-writing skills: Criteria-based self- and tutor-evaluation and the role of language background. *Medical Education, 34*, 194-198.

Cutler, B. L., & Wolfe, R. N. (1989). Self-monitoring and the association between confidence and accuracy. *Journal of Research in Personality, 23*, 410-420.

Das, M., Mpofu, D., Dunn, E., & Lanphear, F. H. (1998). Self and tutor evaluations in problem based learning tutorials: Is there a relationship? *Medical Education, 32*, 411-418.

Davis, D. A., Mazmanian, P. E., Fordis, M., Harrison, R. V., Thorpe, K. E., & Perrier, L. (2006). Accuracy of physician self-assessment compared with observed measures of competence. *Journal of the American Medical Association, 296*, 1094-1102.

Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology, 63*, 568-584.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest, 5*, 69-106.

Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of Personality and Social Psychology, 57*, 1082-1090.

Edwards, N. M. (2007). Student self grading in social statistics. *College Teaching, 55*, 72-76.

Ehrlinger, J., & Dunning, D. (2003). How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of Personality and Social Psychology, 84*, 5-17.

Eva, K. W., Cunningham, J. P. W., Reiter, H. I., Keane, D. R., & Norman, G. R. (2004). How can I know what I don't know? Poor self assessment in a well-defined domain. *Advances in Health Sciences Education, 9*, 211-224.

Everson, H. T., & Tobias, S. (1998). The ability to estimate knowledge and performance in college: A metacognitive analysis. *Instructional Science, 26*, 65-79.

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research, 59*, 395-430.

Fitzgerald, J. T., Gruppen, L. D., & White, C. B. (2000). The influence of task formats on the accuracy of medical students' self-assessments. *Academic Medicine, 75*, 737-741.

Fox, S., & Dinur, Y. (1988). Validity of self-assessment: A field evaluation. *Personnel Psychology, 41*, 581-592.

Gordon, M. J. (1991). A review of the validity and accuracy of self-assessments in health professions training. *Academic Medicine, 66*, 762-769.

Hassmén, P., & Hunt, D. P. (1994). Human self-assessment in multiple-choice testing. *Journal of Educational Measurement, 31*, 149-160.

Hayes, A. F., & Dunning, D. (1997). Construal processes and trait ambiguity: Implications for self-peer agreement in personality judgment. *Journal of Personality and Social Psychology, 72*, 664-677.

Hodges, B., Regehr, G., & Martin, D. (2001). Difficulties in recognizing one's own incompetence: Novice physicians who are unskilled and unaware of it. *Academic Medicine, 76*, 87-89.

Koka, A., & Hein, V. (2003). Perceptions of teacher's feedback and learning environment as predictors of intrinsic motivation in physical education. *Psychology of Sport and Exercise, 4,* 333-346.

Kruger, J. (1999). Lake Wobegon be gone! The "below-average effect" and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology, 77,* 221-232.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77,* 1121-1134.

Kruger, J., & Dunning, D. (2002). Unskilled and unaware–but why? A reply to Krueger and Mueller (2002). *Journal of Personality and Social Psychology, 82,* 189-192.

Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology, 82,* 180-188.

Leopold, S. S., Morgan, H. D., Kadel, N. J., Gardner, G. C., Shaad, D. C., & Wolf, F. M. (2005). Impact of educational intervention on confidence and competence in the performance of a simple surgical task. *The Journal of Bone and Joint Surgery, 87,* 1031-1037.

Locke, E. A., Latham, G. P. (1990). Goals and feedback (knowledge of results). In *A Theory of Goal Setting & Task Performance* (pp. 173-205). Englewood Cliffs, NJ: Prentice-Hall.

MacDonald, J., Williams, R. G., & Rogers, D. A. (2003). Self-assessment in simulation-based surgical skills training. *The American Journal of Surgery, 185,* 319-322.

Martin, D., Regehr, G., Hodges, B., & McNaughton, N. (1998). Using videotaped benchmarks to improve the self-assessment ability of family practice residents. *Academic Medicine, 73,* 1201-1206.

Mattheos, N., Nattestad, A., Falk-Nilsson, E., & Attstrom, R. (2004). The interactive examination: Assessing students' self-assessment ability. *Medical Education, 38,* 378-389.

Matthews, M. D., & Beal, S. A. (2002). *Assessing situation awareness in field training exercises*. (Research Report 1795). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

McKinstry, B., Peacock, H., & Blaney, D. (2003). Can trainers accurately assess their training skills using a detailed checklist? A questionnaire-based comparison of trainer self-assessment and registrar assessment of trainers' learning needs. *Education for Primary Care, 14,* 426-430.

Metcalfe, J. (1998). Cognitive optimism: Self-deception or memory-based processing heuristics? *Personality and Social Psychology Review, 2*, 100-110.

Milgrom, P., Weinstein, P., Ratener, P., Read, W. A., & Morrison, K. (1978). Dental examinations for quality control: Peer review versus self-assessment. *American Journal of Public Health, 68*, 394-401.

Mirabella, A., & Love, J. F. (1998). *Self-assessment based mini-after action review (SAMAAR) methodology: Developmental application to division artillery staff training.* (Technical Report 1086). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Moreland, R., Miller, J., & Laucka, F. (1981). Academic achievement and self-evaluations of academic performance. *Journal of Educational Psychology, 73*, 335-344.

Muchinsky, P. M. (1996). The correction for attenuation. *Educational & Psychological Measurement, 56*, 63-75.

Murphy, K. R. (2008). Explaining the weak relationship between job performance and ratings of job performance. *Industrial and Organizational Psychology, 1*, 148-160.

Oskamp, S. (1965). Overconfidence in case-study judgments. *The Journal of Consulting Psychology, 29*, 261-265.

Osterman, P. (2007). 'Implications of methodological advances for the practice of personnel selection: How practitioners benefit from meta-analysis': Comment of Le, Oh, Shaffer, and Schmidt. *Academy of Management Perspectives, 21*, 16-18.

Pallier, G. (2003). Gender differences in the self-assessment of accuracy on cognitive tasks. *Sex Roles, 48*, 265-276.

Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., et al. (2002). The role of individual differences in the accuracy of confidence judgments. *The Journal of General Psychology, 129*, 257-299.

Parker, R. W., Alford, C., & Passmore, C. (2004). Can family medicine residents predict their performance on the in-training examination? *Family Medicine, 36*, 705-709.

Randal, R., Ferguson, E., & Patterson, F. (2000). Self-assessment accuracy and assessment centre decisions. *Journal of Occupational and Organizational Psychology, 73*, 443-459.

Regehr, G., Hodges, B., Tiberius, R., & Lofchy, J. (1996). Measuring self-assessment skills: An innovative relative ranking model. *Academic Medicine, 71*, S52-S4.

Reider, B. J. (2008). *Army Self-Development Handbook*. Washington, DC: Office of the Secretary of the Army.

Ross, J. A. (2006). The reliability, validity, and utility of self-assessment. *Practical Assessment, Research & Evaluation, 11*, 1-13.

Shadrick, S. B., & Schaefer, P. S. (2007). *Development and content validation of crisis response training package Red Cape: Crisis action planning and execution.* (Research Report 1875). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Schraw, G. (1997). The effect of generalized metacognitive knowledge on test performance and confidence judgments. *Journal of Experimental Education, 65,* 135-146.

Shoenfelt, E. L. (1996). Goal setting and feedback as a posttraining strategy to increase the transfer of training. *Perceptual and Motor Skills, 83,* 176-178.

Sidhu, R. S., Vikis, E., Cheifetz, R., & Phang, T. (2006). Self-assessment during a 2-day laparoscopic colectomy course: Can surgeons judge how well they are learning new skills? *The American Journal of Surgery, 191*, 677-681.

Soll, J. B. (1996). Determinants of overconfidence and miscalibration: The roles of random error and ecological structure. *Organizational Behavior and Human Decision Processes, 65*, 117-137.

Stankov, L., & Crawford, J. (1996). Confidence judgments in studies of individual differences. *Personality and Individual Differences, 21*, 971-986.

Story, A. L. (2003). Similarity of trait and construal and consensus in interpersonal perception. *Journal of Experimental Social Psychology, 39,* 364-370.

Strong, B., Davis, M., & Hawks, V. (2004). Self-grading in large general education classes: A case study. *College Teaching, 52,* 52-57.

Sullivan, K., & Hall, C. (1997). Introducing students to self-assessment. *Assessment & Evaluation in Higher Education, 22,* 289-305.

Taras, M. (2001). The use of tutor feedback and student self-assessment in summative assessment tasks: Towards transparency for students and for tutors. *Assessment & Evaluation in Higher Education, 26,* 605-614.

Ulmer, M. B. (2000, June). *Self-grading: A simple strategy for formative assessment in activity-based instruction.* Paper presented at the Conference of the American Association for Higher Education, Charlotte, NC.

Vogt, D. S., & Colvin, C. R. (2005). Assessment of accurate self-knowledge. *Journal of Personality Assessment, 84,* 239-251.

Ward, M., Gruppen, L., & Regehr, G. (2002). Measuring self-assessment: Current state of the art. *Advances in Health Sciences Education, 7,* 63-80.

Ward, M., MacRae, H., Schlachta, C., Mamazza, J., Poulin, E., Reznick, R., et al. (2003). Resident self-assessment of operative performance. *The American Journal of Surgery, 185,* 521-524.

Wells, L. E., & Sweeney, P. D. (1986). A test of three models of bias in self-assessment. *Social Psychology Quarterly, 49,* 1-10.

Young, J. M., Glasziou, P., & Ward, J. E. (2002). General practitioners' self ratings of skills in evidence based medicine: Validation study. *British Medical Journal, 324,* 950-951.

Zakay, D., & Glicksohn, J. (1992). Overconfidence in a multiple-choice test and its relationship to achievement. *Psychological Record, 42,* 519-524.