The Association of Social Work Boards' Licensure Examinations: A Review of Reliability and Validity Processes

Stephen M. Marson, Donna DeAngelis and Nisha Mittal

Research on Social Work Practice 2010; 20; 87
DOI: 10.1177/1049731509347858

Research Article

Research on Social Work Practice
Volume 20 Number 1
January 2010 87-99
© 2010 The Author(s)
10.1177/1049731509347858
http://rswp.sagepub.com

The *Association of Social Work Boards'* Licensure Examinations

A Review of Reliability and Validity Processes

Stephen M. Marson
University of North Carolina at Pembroke
Donna DeAngelis
Association of Social Work Boards
Nisha Mittal
ACT, Inc.

Objectives: The purpose of this article is to create transparency for the psychometric methods employed for the development of the Association of Social Work Boards' (ASWB) exams. **Results:** The article includes an assessment of the macro (political) and micro (statistical) environments of testing social work competence. The seven-step process used to ensure content validity is discussed. The types of reliability and validity methods employed to assess ASWB exams are discussed. Examples are offered to illustrate the statistical methods employed for selecting good items (questions, some of which are not in question form) and eliminating poor ones. A discussion of the finances of reliability and validity research is included. **Conclusion:** Readers will have a deeper understanding of how exams are constructed. The information housed within this article may facilitate preparing students for the exams.

Keywords: *statistics; reliability; validity; measurement; social work exams*

The requirements for social work licensure exist to protect the public. These requirements set minimum standards for education, experience, supervision, and demonstrated knowledge. The social work regulatory board must be assured that a candidate for licensure meets the requirements in all the areas specified by the jurisdiction's regulatory law. The purpose of the social work licensing examinations is to determine whether social workers have the minimum knowledge necessary to practice in a competent and safe manner with little risk to the public they serve (Thyer & Biggerstaff, 1989). This article includes macro and micro aspects of test construction and testing under the Association of Social Board Work Boards (ASWB) examination program. It will discuss the practice analysis, the process of developing items, passing rates (ASWB, 2004a), including score interpretation, and the results of evaluations by independent psychometricians.

Overview of Macro and Micro Aspects of Testing Reliability and Validity

The systematic methods and standards of constructing an exam have radically changed within the past 20 years (Downing & Haladyna, 2006). In recent years, examinations have been based on item response theory (IRT), a more advanced testing model than the

Authors' Note: Earlier versions of this manuscript were presented at the 25th Annual Baccalaureate Program Directors Conference, March 8, 2008 (Destin, FL) and at the 54th Annual Program Meeting of the Council on Social Work Education, November 1, 2008 (Philadelphia). The authors thank Kathleen Hoffman for her editorial work. The authors declared a potential conflict of interest (e.g. a financial relationship with the commercial organizations or products discussed in this article) as follows: Ms. DeAngelis is the Executive Director of the Association of Social Work Boards (ASWB). Dr. Marson is a long-time unpaid volunteer who has served on the ASWB Examination committee; preceding that, he worked for several years as a contracted item writer for ASWB. Dr. Mittal is a staff psychometrician for ACT, Inc., ASWB's testing contractor. The authors received no financial support for the research and/or authorship of this article. Correspondence may be addressed to Stephen M. Marson, Sociology & Criminal Justice Department; University of North Carolina at Pembroke; Pembroke, NC 28373; e-mail: steve.marson@nc.rr.com.

traditional classical theory. ASWB exams are based on the IRT model, which helps increase the precision of estimating passing scores based on the standards initially set by the passing score committee and, thereby, increases the consistency of pass/fail results.

In this section, two levels of exam construction are examined: macro and micro. On a macro level, the standards for agencies and organizations that develop high-stakes and gatekeeping (licensure and certification) exams are presented. These examinations affect whether an individual will be able to practice his or her chosen profession. This encompasses sociopolitical and ethical foundations of the current environment in which exams are developed. On the micro level, concepts of reliability and validity are reviewed.

Macro Environments and Testing

What standards exist for authority and legitimacy in high-stakes examinations? Who or what organization is allowed to construct gatekeeping exams? There is no governmental accreditation system that exists for organizations that maintain and supply professional gatekeeping exams. Historically, Marson (1981) reports on the development of the National Commission for Health Certifying Agencies (NCHCA). In the 1970s, the federal government provided seed money to establish a nongovernmental, private, nonprofit agency to develop objective standards for organizations that offer credentials to health professionals. Since 1981, NCHCA opened its membership to other organizations and changed its name to the National Commission for Certifying Agencies (NCCA). The original intent was to provide a "Good Housekeeping Seal of Approval" for organizations that construct high-stakes professional gatekeeping exams. It was to parallel the accreditation system that we currently have for colleges and universities. In its current form, the National Organization of Competency Assurance does not meet the expectation articulated in the early 1970s (see http://www.noca.org/). As a result, organizations such as ASWB establish their authority and legitimacy via an alternative avenue.

The alternative to having an accreditation agency is producing a national standard of quality assurance and systems that ensure quality. In this respect, a 3-point system using the following elements is employed by most high-stakes testing organizations, including ASWB:

1. A universally accepted standard for exam construction;
2. An organization that includes psychometricians who monitor and statistically screen exam items; and
3. A periodic independent audit in the case of ASWB by a psychometrician having no vested interest in the outcome.

A description of each point follows. First, the universal standard for high-stakes test construction is found in Standards for Educational and Psychological Testing. It was written by three professional organizations: the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (AERA et al., 1999). This monograph addresses professional and technical issues of test development. It includes changes in federal law and measurement trends affecting validity and also affecting the testing of individuals with disabilities or different linguistic backgrounds. The work is revised a minimum of every 10 years to address changes in testing standards, changes in federal laws, changes in statistical processes, and technological advances. At the writing of this document, Standards for Educational and Psychological Testing is undergoing a revision that will be completed in 2009. Because of the rapid technological changes that have a direct impact on psychometrics, an interim monograph is also employed as a standard to assess gatekeeping exams (Downing & Haladyna, 2006).

Second, the professional staff members within these organizations who are responsible for high-stakes examinations are usually not psychometricians. Rather, they are professionals with degrees in various fields such as social work, computer science, or English. These staff members coordinate test construction activities with assistance from testing experts. The ASWB examinations are contracted to ACT, Inc., for test development, administration, security, and monitoring of the psychometric properties of the test. ACT, Inc., is a nonprofit testing organization that develops and administers qualifying examinations for postsecondary education and provides exam development and administration assistance to organizations that use certifying and licensure examinations. ACT employs a large staff of psychometricians who are experts in testing and measurement. This organization develops and administers a number of high-stakes examinations in coordination with content experts in various fields. With ASWB, it provides psychometric expertise and abilities to combine with the subject matter knowledge of social workers.

Third, the standard practice for organizations that construct and administer gatekeeping exams is to periodically complete a summative and formative assessment of the product. The organizations' boards of directors have a fiduciary responsibility to assure that such an evaluation complies with the highest possible

standards. Failure to perform such an assessment may subject the organization to serious civil litigation. Boards of directors are acutely aware of these consequences. To comply with this standard, ASWB's Board of Directors periodically mandates an external and independent audit that includes both summative and formative dimensions of the exam (Haladyna, 2000; Cizek, 2008). The results of both of these evaluations were that the ASWB examinations are valid, reliable, and defensible.

Micro Aspects of Testing

Two standard psychometric concepts are employed in testing: reliability and validity. These are discussed in brief.

Figure 1
Decision Accuracy

		True ability	
Outcome of exam		**Social workers at or above minimal competency for public protection**	**Social workers below minimal competency for public protection**
	Pass	Correct decision A	Incorrect decision B
	Fail	Incorrect decision C	Correct decision D

Reliability. Reliability is an estimate of consistency of test scores, the degree to which examinees score the same thing over replications of a measurement procedure (Brennan, 2001). That is, do test takers receive the same score (within some measurement error) in repeated attempts? Do the candidates who fail continue to fail on repeated attempts if there is no change in knowledge, skills, and abilities (elements of practice combined and abbreviated as KSAs)? The psychometric concept of reliability is complex and often misunderstood. For the licensure examinations, Figure 1 provides a simple guide for assisting readers in understanding measurement of true ability using reliable tests and the function it serves in protecting the public.

In a perfect world with a perfect exam, all test takers will fall into categories A or D. Competent social workers will pass the test; incompetent ones will not. The reality of any exam is not a clear dichotomy. Insurance companies can testify that a number of practitioners fall into category B (e.g., a licensed social worker sued for incompetent practice). We believe we have met social workers who fall into category C; social workers who seem to be competent but cannot pass or have difficulty passing the exam.

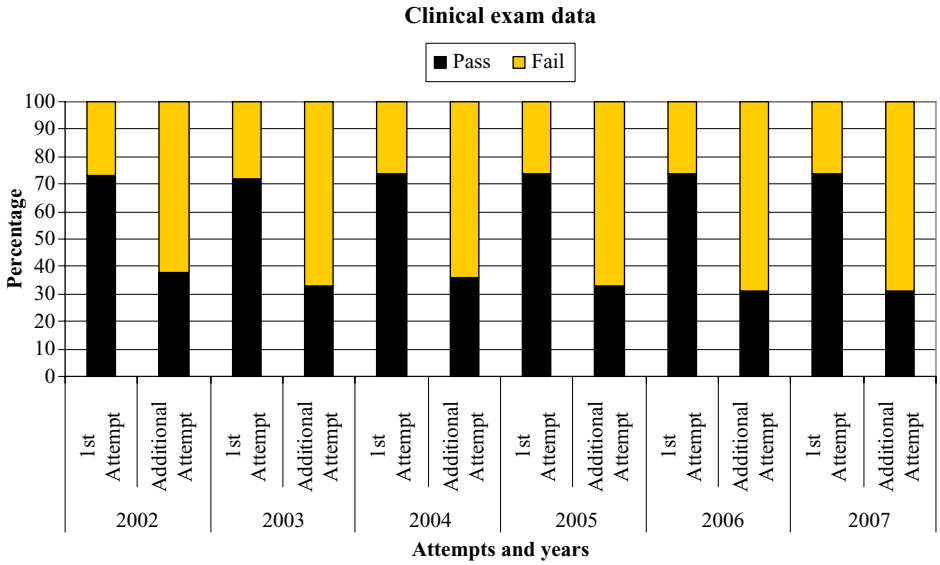
A psychometrician's primary focus is to maximize the test information function at the cut score so that competent candidates fall into category A, incompetent candidates fall into category D, and a negligible number fall into categories B and C. Even in an ideal exam, a few candidates are likely to fall into categories B and C due to measurement error. For licensing examinations like ASWB's (where all candidates who pass are considered competent and candidates who fail are considered incompetent), the decision consistency in pass/fail decisions is considered a more appropriate form of reliability than the traditional classical concept of reliability, the Kuder-Richardson Formula 20 (KR-20; Haertel, 2006). Reducing the number of social workers who fall into categories B and C is a function of this decision consistency. As the decision consistency increases, the number of people who fall within categories B and C decreases.

For licensure examinations, which serve to protect the public by licensing candidates who pass the examination, psychometricians like to see high decision consistency (in the nineties) in the nineties, The ASWB examinations have shown high reliability estimates, in the nineties, both by the preferred advanced IRT model (decision consistency in pass/fail decisions) and the less relevant classical standards (KR-20, test reliability measure as shown by its internal consistency). A classical test score reliability that is close to 0.8 is considered acceptable; one that approximates 0.9 is considered excellent (DePoy & Gitlin, 1998; Marlow, 2005; Royce, 2008; Rubin & Babbie, 2008). This means that the testing data support the ASWB exams in making the best possible pass/fail decisions for competent/incompetent candidates, thereby protecting the public by licensing the deserving candidates.

There is a consistently high failure rate of the repeat test takers (candidates who were unsuccessful in their initial attempt—see Figure 2) of the ASWB exams. Since these exams are considered good screening instruments to filter the minimally competent candidates who demonstrate adequate knowledge, skills, and abilities for safe practice, it is recommended that before the next attempt, the repeat candidates should enhance their knowledge to increase their probability of success.

Validity. "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of tests" (AERA et al., 1999, p. 9). Validity is an argument for assessment of the evidence that the test measures what it is supposed to measure. This argument also pertains to how well the test achieves its intended purpose as well as evidence

Figure 2
Clinical Exam Data



that the test does not measure something other than the target attribute. An assessment may be valid for one purpose but not for the other. For example, there should be evidence that a test for mathematical knowledge is appropriate for the intended purpose, but this same test should not necessarily be a test of reading ability. Support for valid use of the social work examinations would mean that the exams are constructed using the standard guidelines as specified in the AERA et al. (1999), that they are based on a definition of what a level of social work practice signifies, and that there is evidence that scores from each category of the examinations can be used to identify who has adequate knowledge for safe practice at each level.

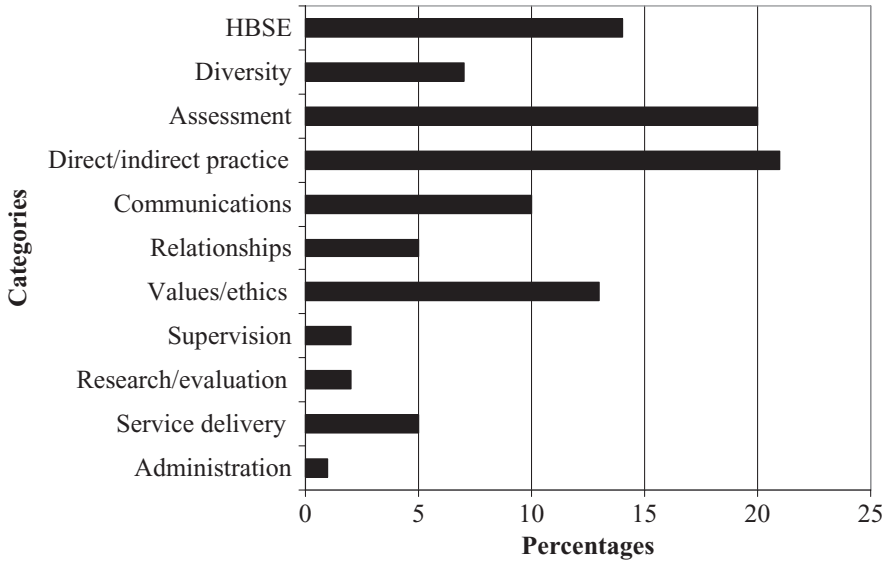
The test items are written only by trained social work professionals and are intended to reflect the testing of only the critical knowledge areas identified by the test specifications. After the initial entry, each item undergoes several tiers of critical review by the ASWB Examination Committee, consisting of a group of social work experts, to ensure that it represents at least one of the KSAs identified as critical and important to social work practice. The committee also ensures that the response to the test items is not unfairly influenced by geographical region, race or ethnicity, gender, or level of proficiency in English (ACT, 1998; Mittal, 2004). Additionally, they ensure that each test item has correct and verifiable information with a single best answer. Once the item is approved, it is edited for appropriate style, consistency, and reading level. As a final step, the test items are pretested. Only items that meet the highest psychometric standards for appropriate difficulty (percentage of candidates answering the items correctly) and discrimination (the ability to distinguish between high and low scoring candidates) are used for the scored exam.

The ASWB organization and the testing company together ensure that all the appropriate steps for test construction are taken: analysis of professional practice, development of test specifications from the practice analysis, item writing based on test specifications, item review for appropriate wording and relevancy to practice, pretesting of items by using them on tests on an unscored basis to screen for unexpected performance, and assembly of a final test form based on psychometrically sound test items. The ASWB test scores in each of the three categories (Bachelors, Masters, and Clinical—the Advanced Generalist exam was not included in this discussion because of the small numbers of candidates who have taken it in the past few years) reflect the knowledge of the candidates in the specific category tested and based on performance, the evidence of adequate competency of safe practice at that level.

Task Analysis

The foundation for content validity is the task analysis and is mandated by AERA et al. (1999), Downing and Haladyna (2006), Haladyna (2000), and Cizek (2008). First, the examination items are based on knowledge statements developed through a North American practice analysis survey in which social workers in the United States and Canada were asked to identify and

Figure 3
Content Areas in Bachelors Exam



rank the tasks they must know how to perform on the first day of their job. The data from this survey are compiled by the test company and are analyzed by social work subject matter experts, who then construct the content outline. The respondents to the survey statistically reflect the makeup of the profession, as does the composition of the subject matter experts who analyze the data. The last survey was conducted in 2001–2003, with the latest one underway in 2008. The examinations that began being administered on May 17, 2004, test content determined by the results from the survey information.

This analysis includes a random survey of thousands of practicing social workers. The results outline the professional tasks in which social workers are involved. The analysis ranks each of the tasks by assessing the frequency and importance that survey respondents indicated during their task-by-task responses, as well as whether or not the survey respondents indicated that the knowledge to perform the task was necessary at entry to the profession or could be learned "on the job" (Questionnaire for the task analysis, retrieved 2007 at www.aswb.org). Tasks that are frequent and important activities among practicing social workers have greater weight as reflected by the number of items on the examination when compared to tasks that show less frequent activity. For example, the task analysis demonstrates that knowledge of social work history is not employed in the daily practice of social work. Thus, one will not find "history" items on the exam. The bar charts in Figures 3–5 present the results from the Bachelors, Masters, and Clinical analyses (respectively). The percentages indicate the corresponding proportion of items that are included on each exam (ASWB, 2004b).

Steps for Item Development

Test item construction lays the foundation for a valid test. Downing and Haladyna (2006) outline the most recent standards for item writer training. Each of the eight steps used by ASWB in item development, including training writers, is outlined in Figure 6 and is discussed in detail.

Stage 1: Recruitment of Item Writers

Several common sense eligibility criteria are employed for recruitment of item writers:

- A degree in social work from an accredited institution.
- A license or certification in one's jurisdiction (state or province).

Announcements of the need for item writers are made in several venues. These include but are not limited to ASWB's Web site, The Council on Social Work Education's Annual Program Meeting, the National Association of Black Social Workers' annual conference, the Baccalaureate Program Directors' annual meeting, the annual meeting of the Rural Social Work

Figure 4
Content Areas in Masters Exam

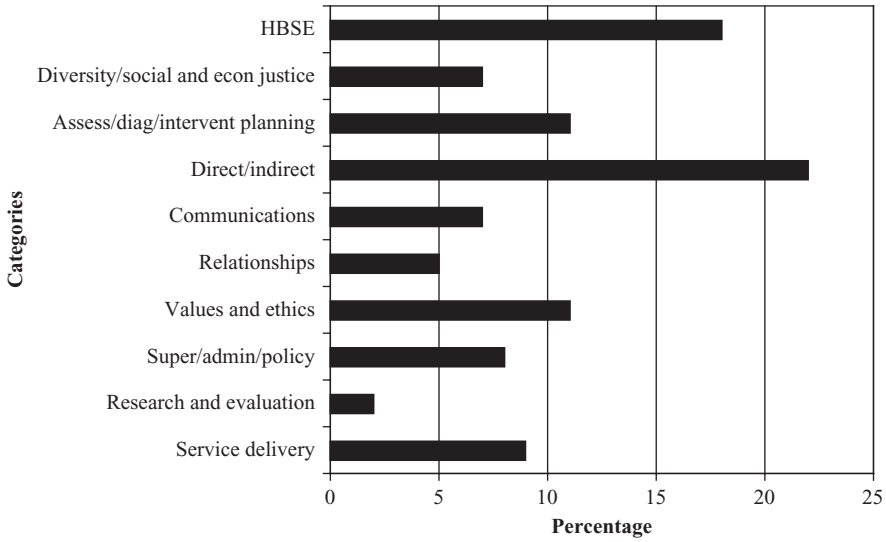
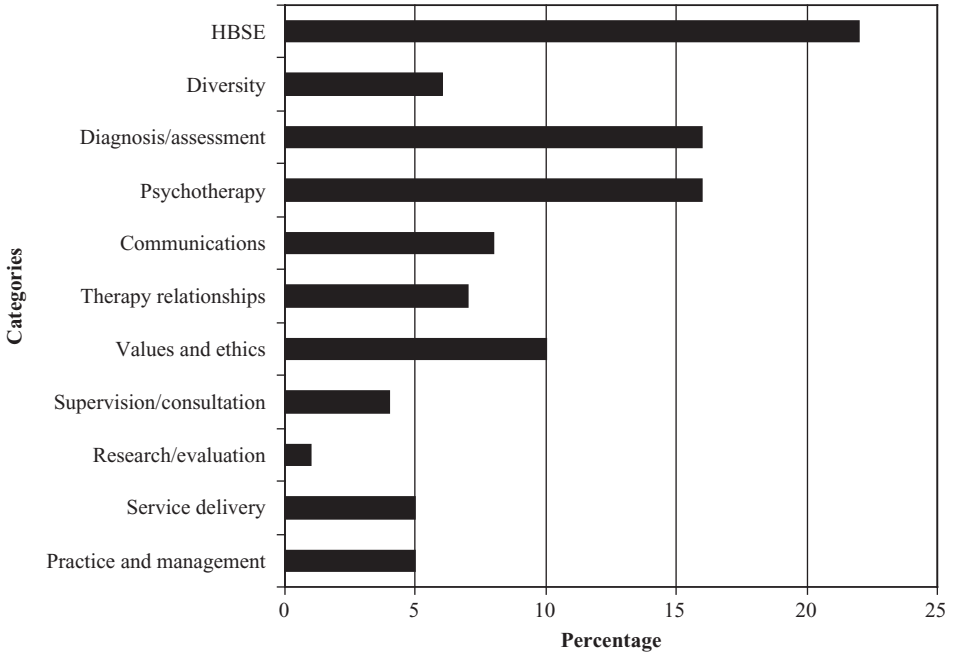
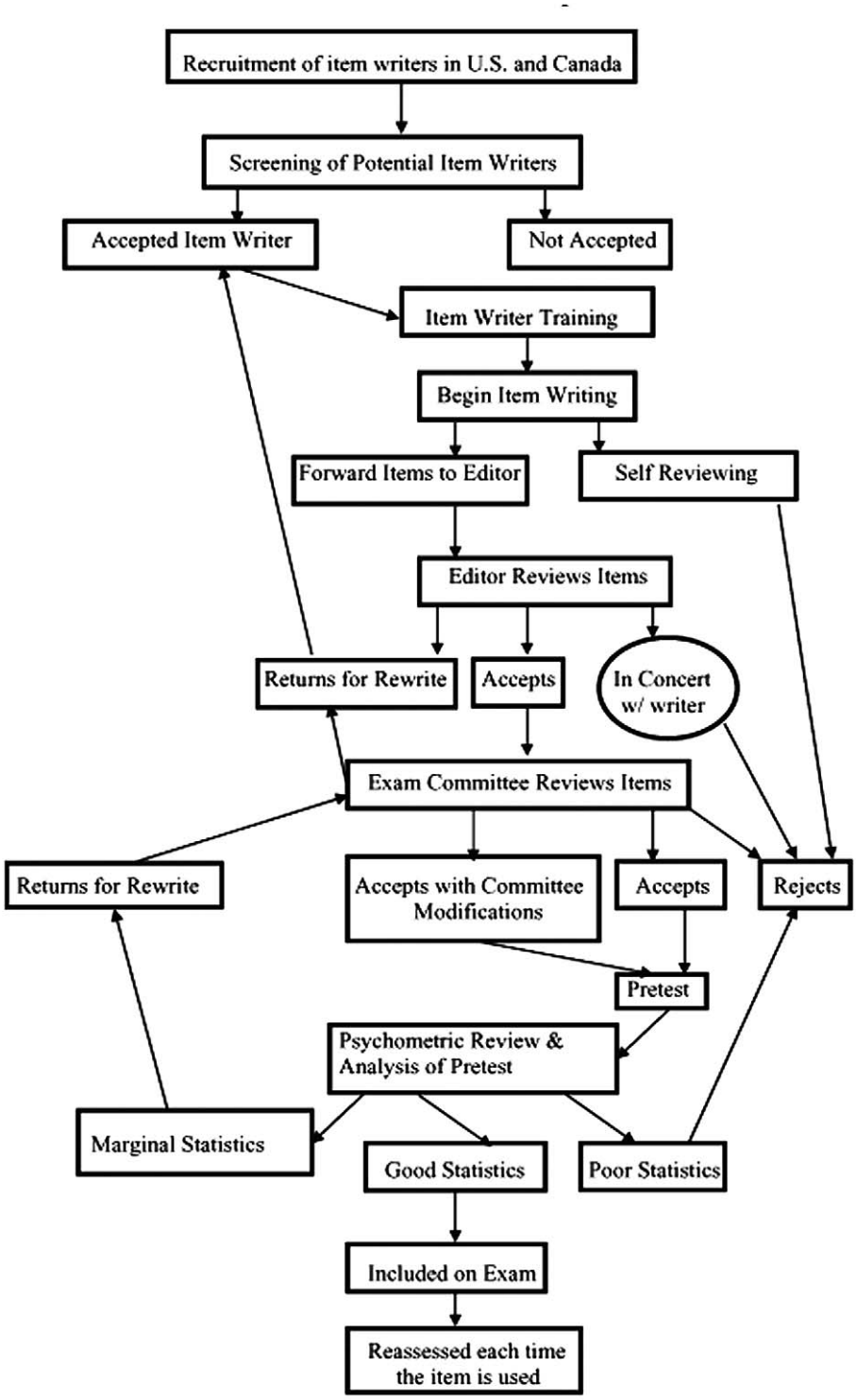


Figure 5
Content Areas in the Clinical Exam



Caucus, other annual conferences, and newsletters of NASW chapters and others. ASWB targets announcements of its annual search for writers and makes choices among applicants in ways that help recruit and retain item writers that reflect the diversity in race and ethnicity, gender, area of practice, and geographic location within the profession. Social workers who want to be considered for training as item writers are asked to send a letter of interest and a copy of their curriculum vitae to ASWB.

Figure 6
Flowchart for Item Development



Stage 2: Screening Potential Item Writers

Social workers applying to become item writers are sent a screening document that gives the rules for item writing standards for content validity. These social workers are tested on their ability to apply these rules in writing three sample items and editing/correcting three other flawed items. Many potential writers do not

complete the task and are not considered further for item writer training (ASWB, 2005). For the 2009 class of writers, for example, there were about 160 screening tests sent out. Fewer than 50 were completed. For those who complete the task, the ASWB staff assesses the quality of their work. Essentially, this is a pass/fail test. Those who perform poorly are not invited to item writer training. Those who complete the task in an acceptable manner are further reviewed in light of the needs of the program for demographic and subject matter diversity. The licensed social workers who are the best choices in both areas are invited to attend the training.

Stage 3: Item Writer Training

Item Writer Training is an annual 3-day event that takes place in Northern Virginia. On the average, 20–25 candidates are accepted to item writer training (ASWB, 2003, 2004c). During this training event, the participants are introduced to the basic concept of content validity. The criteria for content validity are contained in a number of rules and directions, both as included in the screening document and in an Item Writing Guide (ASWB, 2007a).

Stage 4: Item Writing

After completing item writer training, the participants return home and begin to compose items. For each writer, the task is to compose 30 items within a 6- to 8-month period of time (ASWB, 2007b). The assignments are made in 10-item increments and depend on the writers' areas of expertise and the need for items in specific content areas on the examinations. Two dimensions of this task are critical. First, the final product must comply with the standards of content validity. Second, item writers are expected to employ self-editing. Thus, as illustrated in Figure 6, Stage 4 is the first point where items are rejected.

Stage 5: Editor Reviews Items

Upon composing 10 items, the item writer submits his or her final product to one of five social work editors. The social work editors review each item in accordance with content validity standards (AERA et al., 1999; Downing & Haladyna, 2006). After the review, the items are placed into one of three categories:

- Item is accepted. In this case, the item is forwarded to ASWB's exam committee.
- Item is rejected. The editor determines that an item does not comply with minimum standards of content validity. In concert with the item writer, a decision is made to reject the item and return it to the writer with the reasons for rejection. At this stage, we find the second opportunity to reject an item.
- Item is returned to item writer for revision. In concert with the item writer, the editor identifies weaknesses of an item and makes suggestions to improve the item to comply with content validity standards. The work is returned to the item writer for revision or sometimes complete rewriting. After the item is reworked, the item is returned to the editor. At this point, the item is processed as if it were a new item.

When the editor accepts 30 items to be forwarded to the ASWB Examination Committee (within the specified period of time, although extensions are granted if there are acceptable reasons for the delay), the item writer is compensated with a check for US$1,000. However, this payment does not mean that the items will appear on the examinations. Within the process, there are two additional points at which items can be rejected.

Stage 6: Exam Committee Review

The ASWB Examination Committee has 18 members from social work practice and education who are also diverse by race, ethnicity, culture, gender, practice areas, and geography. Committee members are usually appointed from among successful item writers. The exam committee is divided into three parts, each working on a different category of the examinations. They are

- Clinical
- Masters
- Bachelors

These committees review every new item and must reach consensus on each item before it is pretested on the social work examinations. Each committee specifically looks for only one correct answer for each item. If the committee cannot come to consensus, the item is either discarded or changed. There are several versions, or forms, of each ASWB examination category available to be given to test takers at the same time. Members of the Examination Committee or members emeritus who come to committee meetings for that purpose review all the items again on each form of the examination before it goes on-line to the test centers to be administered. The items on each of these forms are different, but the content that is being tested is the same. Candidates are given a different form of the examination if they must retest.

As in Stages 2 through 5, the committee's central focus is the assessment of content validity. ACT and ASWB staff begin each meeting by providing the "Item Review Guidelines" handout (Downing & Haladyna, 2006) and reviewing it orally so that the approved items

Figure 7
An Example of a Statistically Poor Item

A social worker approaches a client in the waiting room of a mental health clinic. The client stands, then has a seizure. The social worker observes that the seizure is mild and that unconsciousness is brief. The client is breathing, and no injury or major consequence is evident. The social worker should NEXT:

A. inquire whether the client has had previous seizures
B. tell other clients to leave the area
C. ask the client if medical care is necessary
D. call for emergency medical services

Correct Response = D

SPLIT	N	A%	B%	C%	D%
Upper (27%)	119	18	2	29	51
Middle (46%)	204	18	3	29	50
Lower (27%)	119	1	15	3	67
Total	442	17	2	26	55
Item Difficulty					55
Pbis		.03	.01	.14	−.15

meet the highest standards. Employing these criteria, members of the committee systematically assess every word in every item for content validity. All of these items have been approved by the social work editor during Stage 5. The committee has three options.

- The most uncommon option is to accept the item as written and forward it to Stage 7 where "item analysis" is assessed.
- The most common option is to accept the item with modifications. Here, fine-tuning takes place. For example, English words that have a history of confusion for Spanish-speaking test takers are replaced with words that are clear to the reader whose primary language is not English.
- The final option is rejection. As outlined in Figure 6, this is the third point in the process in which an item can be rejected.

Reviewing items is a time-consuming process. Generally, one committee member is assigned the task of being a timekeeper. Completion of the modifications is limited to anywhere from 5 to 8 min. If the item cannot be saved within the time frame, the item is often submitted to an individual committee member who will produce modifications based on the committee's deliberations. These rewrites take place outside of the committee's meeting. On rare occasions, which are not included in the outline in Figure 6, rejected items can receive an additional review. If this additional review produces changes that appear to improve the item, the item is sent back to the Examination Committee as an edited item. Every effort is made to produce modifications to save the item. However, some very easy or very difficult items are never forwarded to Stage 7 (ACT, 2008).

Stage 7: Psychometric Review

In Stages 2 through 6, emphasis is placed on content validity. Within Stage 7, we move from the assessment of validity to the more complex item analysis. Within this stage, inferential statistics are employed to differentiate between psychometrically acceptable and unacceptable items. Items are pretested before they can be used as scored items. When an item is being pretested, it means that the item appears on the examination but does not count toward the candidate's score. An item is approved for use as a scored item only if its statistical performance is acceptable. The system of pretesting items protects examination candidates by using only items that have been proven effective in testing relevant knowledge.

Each exam includes a total of 170 items: 150 are scored items that have been approved for validity, reliability, and acceptable item statistics; 20 are new items (unscored) that are included to assess their statistical performance before including them as scored items.

- Approximately 35–65% of the items demonstrate psychometrically acceptable performance (ACT, 2008). These items are loaded into a secure database and are available for use as scored items. However, only items that show the highest statistical performance (a maximum of 25% of the total items pretested) are selected for the scored portion of the exam.
- The items showing poor statistics are either deleted permanently from the ASWB database or salvaged after modification by the Examination Committee. After such modification, the item is again pretested as part of the 20 new items included in a test form.

In Figures 7 and 8, two examples illustrate how item analysis is used. In Figure 7, we see the statistics for a rejected item, and in Figure 8 the statistics for an accepted item.

This frequency table shows how a group of 442 candidates responded to one item on an ASWB test. Based on the total scores achieved on the exam, the entire group was divided (split) into three subgroups. One group, those whose abilities as indicated by overall exam performance fell into the highest category, is identified as upper; second, a comparatively larger group, again going by overall performance, are classified as middle; and the third group as lower, under the column head "SPLIT." Each of the three rows shows the total

Figure 8
An Example of a Statistically Good Item

Keyed Response = A

SPLIT	N	A%	B%	C%	D%	E
Upper	350	89	3	1	7	0
Middle	598	67	9	1	23	0
Lower	350	36	11	3	49	0
Total	1298	65	8	2	26	0
Item Difficulty		65				
Pbis		.44	−.11	−.09	−.39	

Note: Because the data is from a test item currently being used, the entire item must be concealed.

number (*N*) in each group and the percentage of candidates who selected each of the four possible responses (columns A–D) in that group. In this example, the line for the upper group presents the responses of those candidates who scored (their total score) in the top 27%, while the line for the lower group reports the responses of those who scored in the bottom 27% overall. The middle row (46%) presents the choices (A–D) of the rest of the candidates (46%). Finally, the "total" row presents the total number of candidates who responded to this item and the percentage who selected each of the four possible responses. "Item Difficulty" shows the percentage of candidates who selected the correct response. "Pbis" (Point biserial, a measure of the association of a continuous variable and a binary variable) shows the association between the total test score and the candidate's performance on the item (correct or incorrect). The overall score is the continuous variable; the performance on a single item is the binary variable. An item will have higher Pbis if the high scoring candidates get the item correct more frequently than do the low scoring candidates.

This item does not perform well statistically. The item difficulty is 55 (a total of 55% of the 442 total candidates selected the correct key) and a negative discrimination (Pbis = −.15) means it does not discriminate well between the high and low scoring candidates. It is evident from the data that a higher percentage of lower group candidates (67%) selected the correct response compared to only 51% of the upper group candidates. Although the key is correct, this item fails to meet the psychometric criteria of acceptable performance (item difficulty between 40 and 90, and Pbis > .15). It was a pretest item (one of 20 items) and was deleted from the ASWB database.

A group of 1,298 candidates (Figure 8) responded to this statistically well-performing item. The item difficulty is 65 (a total of 65% of the total 1,298 candidates selected the correct response) and good discrimination (Pbis = .44), which means it discriminated well between the competent and the incompetent candidates. It is also evident from the data that a higher percentage of upper group candidates (89%) selected the correct response compared to only 36% of the lower group candidates. This item was a pretest item and was accepted for inclusion into the ASWB database. Since the item is secure and currently in use, its text is not provided in this article.

Along with the statistical performance of items, ASWB exams are analyzed for ethnic and gender bias. The results of empirical analyses provide information as to whether any particular subgroup performs better than another subgroup of comparable ability on the test item. Very few items show statistical bias in the ASWB exams. The existence of statistical bias does not confirm that the item is biased, but it shows a possibility of bias. Such items undergo a sensitivity analysis by experts for the existence of any content bias. If the committee finds the existence of content bias, it is deleted from the database. In the ASWB exams bias is less than 1%.

Stage 8: Continued Statistical Assessment

Once an item is accepted for inclusion in the ASWB database, it is available for selection in the scored portion of the exam. However, its statistical performance is still closely monitored. If an item shows a dramatic change in performance (item difficulty), it is flagged for review. Possible reasons for such a dramatic change include a revision in currently accepted social work practices, a revision in the laws or regulations, or a security breach. These flagged items are reviewed by experts for modification or deleted permanently from the database. If the item is modified, it must go through the pretest cycle once again to regain its status as an acceptable item.

Items are also reviewed for changes in accepted terminology. For example, after the welfare reform, all items that included or referred to the term "AFDC" were pulled from the exam and returned to the Examination Committee for modifications. If modifications were possible to an item, it was recycled and became a new item to be pretested in the future. To repeat an important point again, even though the item was formerly accepted, additional modifications do not

mean that the item is included on the exam without pretesting.

Readability Study

To ensure that the knowledge of English language does not adversely affect ESL candidates (those for whom English is a Second Language), ASWB had a readability study conducted on its examinations. The results of the study showed that the examinations read at the same level as 10th grade textbooks, except, of course, for terms of art related to the social work profession. Since most of the social work text books and journal articles are written at the post-secondary or college level, it was concluded that the knowledge of the English language does not affect the performance of ESL candidates (ACT, 1997; ACT, 1998).

Passing Score Study

Who sets the passing score? Can that score be adjusted? A total of 32 content experts from the Practice Analysis Task Force and the 2003 Examination Committee met together under the guidance of a psychometrician to conduct the passing score study using the Modified Angoff Method, a widely recognized psychometric process described as having judges think of "a number of minimally acceptable persons, instead of only one such person, and ... estimate the proportion of minimally acceptable persons who would answer each item correctly" (Thorndike, 1971). This was done for each of the four ASWB examinations. This study involved a group exercise to conceptualize a "minimally competent" social worker, group members taking the examination, and then in a separate review, estimating "how many of the hundred minimally competent" social workers previously conceived would get each item correct. These ratings were averaged and the results shared with the whole group, outliers were discussed, and then revised in light of the discussion for a total of three iterations. This process was conducted for each of the four ASWB examinations (Mittal, Cartmill, & Vincent, 2005).

The final ratings were computed and averaged into the recommended number of items that must be answered correctly to pass—what is referred to as the raw passing score—for each of the four examinations. Then, the psychometrician conducted two other group exercises to validate the passing score for each exam. As a final step, the ASWB Board of Directors, themselves a diverse group of volunteers, was asked to approve the raw score that a candidate should obtain to pass each test. The Board of Directors made an informed decision only after the critical review of the historical pass/fail data and the projected pass/fail numbers that would be obtained if the results of the passing score study were accepted (Mittal et al., 2005).

The actual number of items required to be correct to pass that was calculated in the passing score study is referred to as the raw score. As mentioned earlier, there are a number of forms for each examination reflecting the same required content. While raw scores can reflect variations in the difficulty of individual items on a particular examination, these variations are accounted for through an equating process. Equating essentially moves the required raw passing score up or down depending on the difficulty levels of individual items. Through equating, the passing raw score is adjusted for each form of the examination so that fewer correct items are needed to pass a more difficult form of the test. Thus, the ability that needs to be demonstrated remains the same from test to test. The passing raw score for each form is the same for all jurisdictions. Every candidate must meet the same passing standard regardless of jurisdiction.

Passing Rates

The passing rates for "first-time" test takers are considered the most accurate picture of test performance. As a group, theoretically each person taking one of the ASWB examinations for the first time has the same chance of passing, so the sample of first-time examinees acts as the control group. When a candidate fails an examination, that candidate's chances of passing are less than they were the first time he or she took it. When the candidates who fail are included in the total group passing average, that average would be lower because those candidates are more likely to fail again, bringing down the passing average for the total group.

Passing rates for first-time examinees on the ASWB exams are relatively stable. The passing rates for first-time examination candidates in 2008 are as follows (ASWB, NDb).

- Bachelors—77.3%
- Masters—74.0%
- Clinical—75.9%

What does it mean when a person fails to obtain a passing score? The ASWB exams are intended to protect the public by screening the competent candidates from the incompetent candidates. Successful performance on the exam suggests that the candidate meets the minimum

competency requirements for safe practice. Unsuccessful performance suggests that the candidate needs to enhance his knowledge, skills, and abilities.

What does it mean when a school does not have an adequate passing rate? The majority of candidates who do not pass have failed to acquire adequate knowledge, skills, and abilities for safe social work practice. It is important to realize that no conclusions can be drawn from only the passing rates of graduates. When employing the test as part of a summative program evaluation, it is necessary to demonstrate a time series pattern. This means that a university researcher needs to collect data over time with a minimum of three and optimum of seven cohorts. Only patterns over time can suggest changes.

Balancing Cost and Validity

Although certification and licensure examinations usually rely exclusively on job analyses and blueprint reviews, more evidence for validity would certainly be desirable. However, in order to conduct such a research study, an organization that does high-stakes licensing examinations would need to consider several factors: cost, feasibility, practicality, and benefits. Examination items are very expensive. Four years ago, the cost of developing items to achieve an acceptable level of psychometric standard and content validity was estimated at US$900 per item (Marson, 2005). Psychometric standards have been slightly increased since then, and the cost as gone up accordingly. This cost is transferred to test takers in the application fee. ASWB has tried to keep the licensing examination fee reasonable. This fee was set at US$175.00 in 2000 and became effective in 2001. The fee has not been increased since that time.

Discussion and Application to Practice

The primary purpose of this review is to provide a comprehensive description of how the social work licensing examinations are developed. The ASWB social work licensing examination program has had regular independent psychometric evaluations, the last one concluded in July 2008. These evaluations have consistently found that the ASWB examination program meets all of the Standards for Educational and Psychological Testing. It was also found by two independent evaluators (Haladyna, 2000 and Cizek, 2008) that the social work licensing examinations are valid, reliable, and defensible. In Cizek's evaluation he concluded, "The examinations produced by the ASWB have been guided by identified best practices in the field of psychometrics and all phases of the ASWB examinations appear to have been guided by the Standards for Educational and Psychological Testing (AERA et al., 1999) and other accepted, defensible professional practices for developing, administering, scoring, and evaluating the ASWB assessments. All of the evidence and documentation reviewed in the preparation of this report suggest that the ASWB examinations produce highly reliable and valid scores for examinees and the jurisdictions that are the primary users of those scores." He also noted in his report, "The ASWB examinations are developed to make pass/fail distinctions based on occupational requirements, not number of individuals need for the workforce."

This article is unique because it presents two types of information. First, most of this material is published elsewhere in a variety of different venues. Thus, the authors were able to bring together research findings from sources that are scattered and that have emphasized only portions of the licensing examination picture. To provide a presentation that is informative while safeguarding examination security, charts were made rather than presenting detailed data. Second, new information never presented in any venue is included.

Since the publication of Pincus and Minahan (1973), empirically based accountability in social work practice has slowly continued to grow and progress. The painstaking work done by ASWB and ACT to ensure that the social work examinations meet all standards of accountability is described in this article in a way that makes the process as transparent as possible without breaching the required security of the exams.

The test development process involves the reliance on psychometric support combined with subject matter expertise. ASWB has demonstrated a commitment to ensuring that both an excellent psychometric foundation and a diverse, representative subject matter expertise are combined to make the exams an appropriate gate keeping tool.

The association has not only obtained excellent psychometric services but has had periodic independent reviews of the examinations by independent psychometricians to ensure that all standards are met. It has also involved subject matter experts diverse in gender, race, ethnicity, geographic location, practice area, and practice setting over the 30 years of its existence.

The central focus of this article has been to illustrate reliability and validity processes, absolutely essential when the successful completion of such an evaluation instrument determines whether or not someone can work in a profession. These very high stakes require careful, responsible, and ethical construction of such

instruments. ASWB has consistently shown over the years that it executes its responsibility regarding test construction and administration according to established psychometric theory and practice, and with adherence to social work ethics.

With advances in technology and in the science of measurement, gate keeping examinations afford increasingly greater precision. ASWB's examinations are carefully kept in the vanguard of these advances, with equal attention to the excellence of subject matter expert participation.

References

ACT. (1997, November). *Discussion paper on procedures for testing ESL candidates for social work licensure, prepared for the American Association of State Social Work Boards (AASSWB)*. Unpublished manuscript.

ACT. (1998, November). *Readability analysis of licensure exams for the American Association of State Social Work Boards (AASSWB)*. Unpublished manuscript.

ACT. (2008, August). Annual technical report, prepared for the Association of Social Work Boards.

AERA (American Educational Research Association), APA (American Psychological Association), & NCME (National Council on Measurement in Education). (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

Association of Social Work Boards. (2003, Summer). *The Itemizer: News for item-writers*.

Association of Social Work Boards. (2004a). *The exam "blue book."* Culpeper, VA: Author.

Association of Social Work Boards. (2004b). *Analysis of the practice of social work 2003: Final report.* Culpeper, VA: Author.

Association of Social Work Boards. (2004c, Summer). *The Itemizer: News for item-writers*.

Association of Social Work Boards. (2005, Winter). *The Itemizer: News for item-writers*.

Association of Social Work Boards. (2007a). *The item writing guide*.

Association of Social Work Boards. (2007b). *Item writer contract*.

Association of Social Work Boards. (ND). (2007). *The questionnaire for the task analysis*. Retrieved June 10, 2007, from http://www.aswb.org/Practice_analysis_files/Web_form_b.pdf

Association of Social Work Boards. (NDb). 2007 pass rates. Retrieved March 4, 2009, from http://www.aswb.org/SWLE/2007passrates.asp

Association of Social Work Boards. (2008, December). The year in review. *Association News*, p. 2.

Brennan, R. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, *38*, 295-317.

Cizek, G. J. (2008). *Confidential report: Review of the ASWB licensure examinations.* Unpublished paper. Chapel Hill: University of North Carolina.

DePoy, E., & Gitlin, L. N. (1998). *Introduction to research: Understanding and applying multiple strategies.* St. Louis: Mosby.

Downing, S. M., & Haladyna, T. M. (Eds.). (2006). *Handbook of test development.* Mahwath, NJ: Lawrence Erlbaum Associates.

Haladyna, T. M. (2000). *Confidential report: Evaluation of the ASWB licensure examination.* Unpublished paper. Phoenix: Arizona State University.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 99-101). Westport, CT: American Council on Education/Praeger.

Marlow, C. R. (2005). *Research methods for generalist social work.* Belmont, CA: Brooks/Cole.

Marson, S. (2005). Social work. In K. Kemp-Leonard (Ed.), *The encyclopedia of social measurement* (Vol 3, pp. 539-546). NY: Elsevier Academic Press.

Marson, S. (1981). *The validity of the ACSW examination.* Unpublished paper. Raleigh: North Carolina State University.

Mittal, N. (2004). *Diversity issues in the examination process: A psychometric perspective.* Paper presented to the ASWB Spring Education Meeting, Calgary, Canada.

Mittal, N., Cartmill, S., & Vincent, E. (2005). *Increasing the validity of the standard-setting process for licensure examinations*. Paper presented at American Educational Research Association, Montreal, Quebec, Canada.

Pincus, A., & Minahan, A. (1973). *Social work practice: Model and method.* Itasca, IL: F.E. Peacock.

Royce, D. (2008). *Research methods in social work.* Belmont, CA: Brooks/Cole.

Rubin, A., & Babbie, E. (2008). *Research methods for social work.* Belmont, CA: Brooks/Cole.

Thorndike, R. L. (1971). *Educational measurement* (2nd ed., p. 515). Washington, DC: American Council on Education.

Thyer, B. A., & Biggerstaff, M. A. (1989). *Professional social work credentialing and legal regulation: A review of critical issues and an annotated bibliography.* Springfield, IL: Charles C. Thomas.