

Texas Psychology Exam Project

Technical Advisory Group Feedback

1. Overall Assessment of Competency

To define competency, can that be fully accomplished as part of the job task analysis?

Technical Advisory Group Discussion

Members of the Technical Advisory Group agreed that a Job Task Analysis can play a central role in defining competency, particularly for purposes of defensibility and legal review. Conducting a practice or job analysis was noted to be especially important in addressing potential challenges related to fairness or bias in the exam.

The group discussed the types of competencies that the JTA would target, noting that licensure examinations most commonly assess knowledge and the application of knowledge. Members emphasized that the examination should focus on a candidate's ability to recognize relevant issues and determine appropriate courses of action in professional situations.

The committee also discussed whether the exam could meaningfully assess skills in addition to knowledge. Examples raised included:

- Technical skills, such as the use of specific therapeutic approaches (e.g., Socratic dialogue)
- Relational or interpersonal skills, such as counseling ability or cultural responsiveness

Members noted that certain skills may be difficult to measure directly through a written examination and may instead be reflected indirectly through scenario-based questions that evaluate the application of professional judgment.

The Technical Advisory Group also discussed who should be included in the JTA sample. Members emphasized that the job analysis should primarily reflect the experiences of current practitioners, as they are best positioned to identify the knowledge and tasks that are most important for competent practice. While training expectations may overlap with practice demands, the group emphasized that practitioners should drive the identification of critical knowledge and tasks, rather than relying solely on training standards.

The group also reiterated earlier concerns that the framing of the proposed exam appeared largely oriented toward clinical or health service psychology, and that other licensed practice areas, such as industrial-organizational psychology, may have substantially different practice demands. As a result, licensing boards may ultimately need to determine which competencies are most critical for their regulatory purposes and whether different practice areas should be addressed differently in a licensure examination framework.

2. Assessment of demonstrable competency

What are the pros and cons of separating portions of the exam into a traditional knowledge assessment and a practical/skills-based assessment?

Is there merit and validity arguments to defend fully integrating an assessment of knowledge and skills through demonstration of application of knowledge?

What assessment methods are objective and scalable that could be used to assess knowledge and skills?

Technical Advisory Group Discussion

Members of the Technical Advisory Group discussed potential advantages and disadvantages of separating knowledge-based and skills-based components of the exam. Several members emphasized that the application of knowledge is particularly important in the context of licensure examinations because of the exam's role in protecting the public. The committee noted that the ability to interpret information, recognize important issues, and determine appropriate next steps in professional situations is central to competent practice.

At the same time, members also noted that certain foundational knowledge remains essential. Examples discussed included core measurement and assessment concepts such as standard error of measurement and interpretation of standard scores, which are necessary for psychologists to interpret test results appropriately and make sound professional judgments.

Several members noted that separating knowledge and applied components may allow each to be assessed using formats best suited to the construct being measured. Knowledge can be assessed efficiently using objective-response formats such as multiple-choice questions. Skills, by contrast, may benefit from more complex item formats designed to capture decision making, reasoning, and professional judgment.

Separating knowledge and applied components may also provide advantages from a measurement and operational perspective, including clearer score interpretation and greater control over psychometric procedures such as dimensionality, calibration, and score reporting. In addition, separate sections may simplify item development and test construction. However, some members noted that separating the constructs could introduce reliability considerations, as each section would require a sufficient number of items to produce stable scores. As a result, dividing the exam into separate knowledge and skills sections may require a larger total number of items to maintain adequate reliability.

That said, some members emphasized that professional competence in practice is inherently integrated, and separating knowledge from application may create an artificial distinction. Integrating knowledge and applied reasoning within case-based scenarios may better reflect how psychologists apply knowledge in professional contexts.

One member noted that an application-focused approach could involve structured, stepwise scenarios in which candidates demonstrate increasingly complex decision-making processes. For

example, tasks might progress from demonstrating skills in straightforward contexts to more complex or culturally nuanced situations.

The committee also discussed the potential use of situational judgment items or similar scenario-based formats. Members cautioned that these item types may present test security challenges, as the scenarios can be distinctive and easier for candidates to remember and share. As a result, these items may need to be replaced more frequently to maintain exam security. For examinations with lower candidate volumes, the need for frequent replacement of such items may make them less practical.

Overall, the discussion reflected that both approaches could be valid. Some members suggested that a hybrid design might provide the most balanced solution, combining a foundational knowledge component addressing core public-safety content with case-based or scenario-based items designed to measure judgment, integration of information, prioritization, and next-step decision making. Some members also noted the potential for process-based or sequential decision tasks that capture how candidates work through professional problems over multiple steps.

The committee generally emphasized that exam design decisions should balance valid measurement of applied competency with practical considerations related to development, scoring, cost, and administration.

Use of Artificial Intelligence / Open ended responses

The committee expressed cautious opinions that AI may someday be used in a licensure exam to score more open-ended, response driven assessment methods. Currently, the technology has moved beyond some of the early limitations in its ability to qualitatively score written responses, such as having to rely on length of answer as a factor in scoring. Some assessments are now using AI scoring that looks for key words and concepts, which gets closer to a qualitative assessment. However, this still falls short of the ability to fully analyze and score a written response for knowledge and application of those concepts.

The advancement of technology generally, though, may present opportunities for open-ended short responses that can be more objectively scored. For example, exam takers could be presented a scenario and then asked to identify a diagnosis. The exam can have a pre-established list of all DSM diagnoses. The exam taker can be required to type in a diagnosis fully, or begin typing and choose from available answers that are displayed. In this way, the exam taker must initiate the answer as an open-ended response, but still ultimately selects from answers that can be objectively scored.

With the growth of AI over the next few years, while the exam would be developing, it is also possible that true short answer responses about symptoms, diagnoses, or treatments could be effectively assessed for required key words/concepts using AI.

3. Exam Validation

What steps, tools, benchmarks, etc., should we make sure to require of the test development vendor?

What is the proper use of predictive validity in licensure examination?

Technical Advisory Group Discussion

Members of the Technical Advisory Group emphasized that strong psychometric validation and defensibility should be central requirements for the examination. The committee noted that any vendor selected to develop the exam should demonstrate expertise in modern test development practices, psychometric modeling, and licensure examination design.

The committee indicated that vendors should follow established professional testing standards, including widely recognized guidelines for educational and psychological testing. Vendors should demonstrate the ability to support defensible validation processes, including job or practice analyses, blueprint development, item development procedures, and statistical evaluation of test performance.

Members noted that the vendor should have demonstrated experience with objective scoring methods, particularly for examinations incorporating case-based or scenario-based items. Where more complex item formats are considered, the vendor should provide clear evidence that scoring methods are reliable, standardized, and defensible.

The committee emphasized the importance of ensuring that exam content is closely aligned with professional practice, typically through a rigorous practice or job task analysis. Vendors should be able to demonstrate how exam content is derived from and supported by such analyses.

Members also discussed the potential use of multiple item formats, including case-based questions and other formats designed to assess applied judgment. If interactive elements such as video or audio scenarios are considered, the vendor should demonstrate experience implementing such formats while ensuring accessibility and appropriate accommodations.

One suggestion raised was that the vendor should be able to develop a sample or pilot form of the examination that could be independently pilot tested or reviewed by an external body prior to full implementation.

The committee also emphasized the importance of strong security procedures, data management practices, and ongoing psychometric monitoring, including evaluation of item performance, exam reliability, and fairness across candidate groups.

Predictive Validity

From a validation perspective, there was strong agreement that content validation remains the primary foundation for defensibility, with a clear linkage from job analysis to test blueprint to item development. Members noted the lack of a universally accepted external criterion for professional competence reinforcing the importance of a content-based approach. Even among

leading organizations such as Educational Testing Service have described the difficulty with external criterion. While group differences in outcomes such as pass rates can be examined when data permit, these must be interpreted cautiously and in the context of appropriate standard setting.

The discussion acknowledged the broader measurement literature showing relationships between ability measures and outcomes such as academic performance and income while also recognizing ongoing debate about interpretation and fairness. Exams like the SAT illustrate the complexity as they measure developed knowledge and skills but remain correlated with general ability.

Multiple members discussed the challenges associated with establishing predictive validity for licensure examinations. Members noted that there is currently a lack of widely accepted criterion measures capable of meaningfully evaluating professional competency over time

One member suggested that predictive validity could be examined by comparing licensure exam performance to a more comprehensive clinical assessment. The committee noted that no such standardized assessment currently exists and that developing one would introduce significant complexity. As a result, this approach was not considered practical or relevant for licensure purposes.

Members also discussed the distinction between measuring minimal competency at the point of licensure and attempting to predict professional performance at a later point in time. Professional performance may be influenced by many factors unrelated to the licensure examination, including supervision, experience, workplace context, organizational factors, motivation, professional development opportunities, and practice setting.

Because of these limitations, members expressed significant reservations about the usefulness of predictive validity studies for licensure examinations. The committee generally viewed predictive validity as impractical and of limited relevance for licensure decisions and noted that it would be unreasonable to require predictive validity as a standard for licensure examinations.

Members also noted that professional licensure examinations historically have not been designed or validated using predictive validity evidence, and that such evidence has not been required to establish the defensibility of high-stakes licensure examinations. Further, no licensing exam meets a standard of demonstrating predictive validity.

While the concept of predictive validity continues to be of interest to some stakeholders within the broader community, the committee emphasized that validation of professional entry examinations has traditionally relied on content validity supported by systematic practice analyses and ongoing psychometric evaluation, rather than attempts to predict future professional performance.

Members also discussed that validation should be understood as an ongoing process rather than a binary determination. Rather than viewing an exam as simply “valid” or “not valid,” validation involves accumulating evidence from multiple sources, including practice analyses, content development procedures, statistical analyses of exam performance, and ongoing monitoring of exam results.

Measuring Competency vs Predicting Good Practice

Building on the earlier discussion of predictive validity, members discussed the distinction between measuring competency at the moment of testing and attempting to predict future professional performance.

The committee generally agreed that the primary goal of a licensure examination should be to measure competency at the time of testing, ensuring that candidates possess the knowledge and judgment necessary for safe entry-level practice.

Members noted that while one might reasonably expect that a candidate's current competency would relate to their later professional performance, demonstrating this relationship empirically is difficult for the reasons previously discussed.

Several members also emphasized the importance of recognizing the limits of what an examination can measure. For example, knowledge of ethical guidelines can be assessed through examination, but knowledge alone does not necessarily predict ethical behavior or professional conduct over time.

Members suggested that exam documentation should acknowledge these limits and clarify that the exam measures knowledge and decision-making relevant to entry-level competency, rather than attempting to directly predict long-term professional behavior.

The committee also noted that assessment methods should strive for ecological validity, meaning that testing tasks should approximate real-world professional situations as closely as possible. Scenario-based or case-based approaches may help support this goal by requiring candidates to apply knowledge and judgment in contexts resembling professional practice.

Predictive Validity as a Policy Consideration

In light of these considerations, members discussed whether predictive validity should be considered a reasonable goal for an examination intended to measure entry-level competency.

Several members suggested that predictive validity may represent an unreasonable or impractical goal for licensure examinations, given the methodological challenges and purpose of the examination previously discussed. Some members noted that while predictive evidence may provide supplemental evidence, establishing such evidence could require longitudinal studies conducted over many years, which would likely be outside the scope of the core exam validation process.

The committee generally indicated that content-based validation linked to job or practice requirements should serve as the primary form of validity evidence for licensure examinations. Predictive evidence, if pursued, could potentially be explored through longer-term research efforts separate from the development and validation of the exam itself.

Ultimately, the committee's consensus was that dissatisfaction with the current exam may be better addressed by strengthening both the definition and assessment of competence and more effectively determining whether candidates are prepared to deliver services rather than relying on

predictive validity evidence. The committee emphasized that licensure examinations are best designed to assess the competency required for safe entry into practice instead of attempting to predict long-term professional performance. However, it was recognized that some stakeholders strongly advocate for assessing predictive validity. As such, the committee suggested that consideration be given to including a predictive validity study in the RFP process, with the understanding that such work would be exploratory and supplemental to core validation efforts. Based on proposed costs and staff resources, the board could determine the appropriate level of emphasis to place on this work.

4. Adaptive Assessments

What are the pros and cons of using adaptive testing in this examination?

Technical Advisory Group Discussion

Members of the group discussed the potential advantages and disadvantages of using adaptive testing methods for the proposed examination.

Several members noted that adaptive testing can provide efficiency advantages, particularly when assessing knowledge-based content. Adaptive exams can adjust item difficulty based on a candidate's responses and may allow testing to terminate earlier once it has been statistically established that a candidate has passed or failed. In addition, adaptive testing can allow candidates who are near the passing threshold to answer additional items that help confirm a pass or fail decision.

Members also noted that adaptive testing can support more advanced routing and termination rules, potentially allowing the exam to focus on items that are most informative for estimating a candidate's ability.

However, several potential limitations were also discussed. Effective adaptive testing typically requires large, well-calibrated item pools, which may be difficult to establish for a new testing program. Members noted that adaptive testing significantly increases item needs (5x to 10x the item requirements), both to support calibration and to maintain secure item rotation over time. For this reason, several members suggested that adaptive testing may be more appropriate for a mature testing program with a large and well-developed item bank, rather than as an initial design approach.

Members also noted that adaptive testing may place greater demand on highly discriminating items, which can lead to those items being administered more frequently. Over time, this may increase pressure on item development efforts and create additional challenges related to maintaining item security and replenishing the item pool.

Another issue discussed was candidate perception and acceptance. Some candidates report concerns that missing early questions may negatively affect their exam outcomes, even though this perception is generally inconsistent with how adaptive scoring functions.

The committee also discussed the implications of adaptive testing for candidate score reporting. When all candidates receive the same exam form or sections, it may be easier to provide detailed feedback by content area, which can help candidates understand their performance and prepare for future attempts. Adaptive testing can make this type of feedback more complex, depending on the exam design.

Members also noted that while adaptive approaches may be feasible for knowledge-based assessments, implementing adaptive methods for skills-based or complex scenario-based items could be more difficult depending on the format used.

Overall, the committee indicated that adaptive testing offers potential efficiency advantages, but it also introduces operational, psychometric, and development challenges. Members suggested that adaptive testing may be best considered after a testing program has matured and developed a sufficiently large and stable item bank, rather than as a starting design for a new licensure examination program.

5. Identifying and Preventing Bias

Other than a traditional committee that reviews performance on individual questions to identify potential bias, item analysis, and DIF, are there other processes or tools you recommend to prevent bias?

Technical Advisory Group Discussion

The Technical Advisory Group discussed methods for identifying and mitigating bias beyond traditional approaches such as item analysis, DIF, and committee-based bias review. While bias sensitivity panels have limited empirical support, members noted they can be useful for identifying issues like stereotypes or problematic wording and may enhance stakeholder confidence when used as a supplementary process. DIF was affirmed as a core analytic tool, though it is limited by large sample requirements and reliance on simplified (sometimes arbitrary) group comparisons. As an enhancement, the group discussed latent regression and related modeling approaches, which allow for the inclusion of multiple variables simultaneously, estimation of underlying ability, and separation of true group differences from item-level bias. Latent regression can supplement DIF by modeling overall proficiency and examining whether observed racial differences persist after accounting for other measured candidate characteristics. Its usefulness depends on the quality and scope of available background data, and it cannot control for important influences that are not collected. Examples of potentially useful variables to collect upfront included training pathway, program, and experience level. These approaches are more flexible and can control for covariates, but should be used in addition to, not in place of, traditional DIF.

6. Exam Administration

Other than for practical reasons, is there any reason to limit the time allowed to take the exam?

What should we understand or consider about the process of translating an examination and validity? Discuss validation by language in development versus validation via translation? With a small volume of a specific language, what do you recommend? How much does dialect impact translation (Latin American Spanish vs Castilian Spanish)?

Technical Advisory Group Discussion

Time Limits

The discussion reflected broad agreement that there is no clear theoretical or psychometric rationale for imposing strict time limits on the examination. Participants emphasized that the primary purpose of the assessment is to evaluate competence, not speed, and that time constraints should not interfere with a candidate's ability to demonstrate their knowledge and skills. In this sense, the group generally supported an approach in which time limits, if used, are designed to be sufficiently generous so that the exam is not speeded for the vast majority of examinees.

From a measurement perspective, there was recognition that overly restrictive time limits can introduce construct-irrelevant variance, particularly if performance becomes dependent on reading speed, test-taking pace, or language proficiency rather than the intended domain of competence. Language differences were specifically noted as a factor that may necessitate additional time for some candidates. As such, providing adequate time was viewed as an important consideration for fairness and accessibility. It was also recommended that response time data be monitored to evaluate whether speediness is influencing outcomes, with any such effects raising potential validity concerns.

The group acknowledged that practical considerations may justify some limits on testing time. These include concerns related to test security, such as the potential for item exposure or cheating, as well as logistical constraints associated with exam administration.

One participant observed that extending time for live or proctored verbal responses may be more challenging due to the nature of real-time administration. This raised the possibility that different timing approaches may be needed depending on the format of the assessment.

Language / Translation

The discussion highlighted that language and dialect present significant challenges in test development, particularly for widely spoken languages such as Spanish and French, where meaningful differences exist across regions. There was clear agreement that one cannot assume a single translation will function equivalently across dialects, and that linguistic differences may affect both clarity and validity. Examples were offered to illustrate this point, including phrases such as "I think things will get better" or "feeling blue," which do not translate cleanly or meaningfully into some languages, underscoring the risk of construct distortion through direct translation.

Participants also emphasized that translation addresses linguistic differences but does not fully capture cultural context, which may further impact how items are interpreted. Some approaches, such as parallel development using multiple language panels, were noted as potential ways to improve comparability. However, these approaches introduce additional complexity and resource demands. An example was provided of a program that developed multiple Spanish versions to account for dialectical and cultural variation.

Several members of the group expressed caution about pursuing translation given these challenges. Concerns were raised about the difficulty of demonstrating equivalence across language versions, the limitations of methods such as differential item functioning analyses in cross-language contexts, and the broader challenge of ensuring both linguistic and cultural validity. As an alternative, some members suggested that providing extended time for candidates may be a more practical and defensible approach to addressing language-related barriers.

Overall, there was strong agreement that translated versions of an assessment must be approached with attention to both linguistic and cultural factors throughout the development and validation process. Equivalence across languages must be demonstrated empirically rather than assumed. It should also be recognized that translation itself may introduce more challenges than it resolves.

AI translation

The members of the group raised significant concerns about the validity of offering an AI-translated version of the exam alongside the original English as an accommodation. This approach assumes that a fully equivalent translation can be produced in real time, which participants viewed as unlikely given the complexity of language and the need to preserve meaning across contexts. From a standards and validity perspective, it was emphasized that when multiple language versions are used, there must be evidence that scores have the same meaning and support the same decisions, including equivalence in construct measurement, difficulty, and interpretation. The group expressed concern that this process probably not meet these requirements without rigorous SME review.

Practical constraints were also noted, including challenges integrating AI into secure testing environments and the risk of flawed or inconsistent translations. Overall, the group's view was that this approach would likely not be a defensible option.

7. Options for Exam Division / Modularized

What psychometric impacts do you see with options to making the exam modular / customized? Please advise on the test length required to provide a valid pass score by domain.

Technical Advisory Group Discussion

The group explored the distinction between a single, comprehensive exam and a modular approach (e.g., one 200-item exam versus four 50-item modules). Both compensatory models (where strengths in one area offset weaknesses in another) and conjunctive or multiple-hurdle models (where candidates must pass each module independently) were discussed as viable options. The choice between these approaches was seen as dependent on the profession's priorities and the results of the job analysis. For example, in safety-critical professions, such as aviation, a conjunctive model may be more appropriate because no single area of weakness can be offset by strengths elsewhere.

There was discussion about the implications for how competence is conceptualized. A modular approach requiring candidates to pass each section may better ensure minimum competence across domains but it may not capture consistency of performance across the full scope of practice. A single compensatory exam may allow candidates to demonstrate overall competence. If they have relative weaknesses in specific areas this can be offset by strengths in other areas.

Members also noted that modular testing could potentially improve eventual pass rates by breaking the exam into smaller components. This should be balanced against concerns about maintaining standards, and ensuring consistent competence. Certain domains (Ethics) were identified as potentially warranting stricter standards or higher cut scores regardless of the overall model.

Overall, both modular and comprehensive approaches were considered defensible provided they align with how competence is defined and are supported by the practice analysis. The group noted that most licensure exams are treated as measuring a single overall construct of professional competence even though performance may vary across specific content areas.

The group noted that test length by domain depends on achieving adequate reliability for high-stakes decisions. With typical inter-item correlations of a domain composed of about 20 items would yield an estimated reliability of approximately .78, which is generally considered insufficient. As a result, for a modular approach to test individual domains, roughly double the number of items per domain would likely be needed to reach acceptable reliability levels.

Masters / Supervised practice versus Ph.D. / Independent practice

The group concurred that differences in scope of practice between supervised and independent practitioners should determine design. One approach would be to maintain a single exam and establish different performance standards through the standard-setting process. In this model, separate cut scores could be set to reflect the differing expectations for supervised versus independent practice. However, this approach may include content that is less relevant for those practicing under supervision.

Alternative approaches include incorporating both directly into the job task analysis or conducting separate JTAs which would result in different blueprints and distinct exam forms. This would be likely be more relevant in content. However, supervised practice candidates that later apply for an independent license would be required to take a separate exam. Overall, the group noted that different defensible options exist.