ACUITY INSIGHTS

Professional Skills Development

Manual 2026

Table of Contents

| Preface 6

| CHAPTER 1: INTRODUCTION 8

Key Features 10

Uses of the Professional Skills Development tool 11

Intended Purposes and Use 11

Applicability Across Program Types 12

Principles of Use 14

Respondent Rights 14

| CHAPTER 2: INTERPRETATION 16

What the Professional Skills Development tool Scores Tell You 16

Overarching Construct Definition 16

Dimensions 16

PSD Interpretation 19

Step by Step Interpretation Guidelines: Learner Development 20

Program Level Application 22

Case Examples 23

Case 1 - Student: Developing/Moderate Results across Domains 23

Case 2 - Student: All developing competency scores, potentially unengaged with the task 24

Case 3 - Program level application: University of Waterloo, Pharmacy 26

| CHAPTER 3: TEST DEVELOPMENT 28

Overview 28

Background 28

Prototype 29

Identified Improvements 30

Current Version of the Professional Skills Development tool 32

Framework: Dimensions and Competencies 32

Competency Mapping 32

Thematic Analysis 32

Literature Review 33

Test Structure 33

Behavioural Tendency Questions 33

Open-Ended Responses 33

PSD paradigm Components 34

Modules 34

Scenarios....35
Questions....35
Typed Responses....36
Storyline, Scenario and Question Development....36
Storyline development and script writing process for both video and word based scenarios....37
Stimuli production....37
Question writing process....38
Sample storyline, scenario and questions....38
Basic Storyline....38
Scenario....38
Questions....39
Scoring and Feedback....39
Feature Extraction....39
Regression....40
Feedback Generation....41
Feedback Generation Quality Assurance....43
AI Ethical Framework....44
Fair and Equitable....45
Reliable and Rigorous....45
Responsible and Accountable....45
Secure and confidential....46
Transparent and explainable....46
Descriptive Statistics by Program Type....47
| CHAPTER 4: RELIABILITY....50
Reliability Summary....50
Internal Consistency Reliability....51
Standard Error of Measurement / Confidence Intervals....53
Parallel Forms Reliability....53
Alternate Test Form Internal Consistency Comparability....54
Distributional Comparability....54
Inter-Rater Reliability....55
PSD Inter-rater Reliability with Human Expert Judgment....55
Quadratic Weighted Kappa (QWK)....56
Mean Absolute Error....58
Reliability Conclusions....58

| CHAPTER 5: VALIDITY **60**
Validity Summary 60
Evidence that PSD provides quantifiable measurement of the targeted constructs (Construct Validity) 62
Confirmatory Factor Analysis: Evidence Supporting the Validity of the PSD Structure 62
CFA Results 63
Model Fit 63
Factor Loadings 63
Dimension Relationships 65
Evidence Based on Relations with Other Variables: Casper Scores 65
Sensitivity of AI Scores to Scenario–Question Alignment 67
Correspondence to Scores based on Human Expert Judgement 68
PSD Score correlation with Human Expert Judgment 68
Evidence that PSD effectively distinguishes between performance levels (Discriminative Validity) 69
First year v Fourth year students 69
Known-Groups Validity 71
Group Comparisons at Test 1 71
Group Comparisons at Test 2 72
Evidence that PSD is appropriately sensitive to developmental change (Sensitivity to Change) 74
Fairness: Evidence Regarding Score Comparability Across Demographic Groups (Generalizability Across Groups) 77
Accommodation 78
Community Size 79
Gender 80
Language 80
Parental Income 81
Race / Ethnicity 82
Regression Analysis of Group Comparisons 83
Differential Item Functioning (DIF) by Group 86
| References **91**
Appendix A - Crosswalk Between PSD Dimension model and Across Program Competency Frameworks **97**
Medical Education 97
AAMC/AACOM/ACGME Foundational Competencies for Undergraduate

Medical Education....97
Business Education....98
AACSB Accelerators....98
NACE Career Readiness Competencies....98
Engineering....99
Washington Accord Graduate Attributes....99
Law....100
IAALS Foundations....100
Appendix B - Prototype Research Results....**102**
Initial Psychometric Findings: Reliability....102
Prototype: Validity....103
Correlations with Other Metrics....103

Legal Disclaimer

This manual has been developed in accordance with the Standards for Educational and Psychological Testing (2014) and is intended to provide comprehensive documentation of the Professional Skills Development tool as of the date of publication. While reasonable efforts have been made to ensure accuracy and completeness, Acuity Insights Inc. ("Acuity Insights") makes no representations or warranties, express or implied, regarding the accuracy, reliability, or completeness of the information contained herein.

This manual is subject to periodic updates, which may occur without prior notice. Acuity Insights will make reasonable efforts to notify registered programs of material updates; however, it is the responsibility of users to ensure they are referencing the most current version. Updated versions are available through our website.

This manual describes the intended use, purpose, and interpretation of the Professional Skills Development tool. The tool must be used only in accordance with the guidelines and limitations described herein. Acuity Insights disclaims any liability arising from use of the tool outside of its intended purpose or in a manner inconsistent with this documentation.

Users who choose to apply the Professional Skills Development tool in ways not explicitly described in this manual do so at their own risk and are solely responsible for establishing the validity, appropriateness, and interpretation of such use. Results derived from any non-standard use should be interpreted with caution.

This publication, including all content, materials, and intellectual property contained herein, is the exclusive property of Acuity Insights Inc. No part of this publication may be reproduced, distributed, modified, stored in a retrieval system, or transmitted in any form or by any means, including electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of Acuity Insights Inc., except for permitted use as expressly authorized in this manual.

The Professional Skills Development tool, including its content, scoring methodologies, and associated materials, is protected by applicable intellectual property laws. Unauthorized use, reproduction, or adaptation may result in legal action.

Matt Holland
CEO, Acuity Insights

Preface

Acuity Insights is proud to introduce *Professional Skills Development (PSD).* PSD is a tool that has three main components: a formative assessment, results & feedback, and learning exercises & resources. This tool draws on our expertise in assessment design and our experience with open-response SJTs through the long-established Casper test used for admissions purposes.

The *Professional Skills Development* tool offers program leaders a means to assess and foster personal and professional skills including, for example, collaboration, leadership, empathy, ethics, and critical thinking during training and coursework. The tool also provides learners with feedback based on their responses, as well as resources and exercises they can use to reflect and further develop their skills. It enables continuity and an "end-to-end" approach to assessment, giving educators and learners actionable insights at multiple stages of the academic journey. By measuring key skills in a structured and quantifiable way, we can help identify areas for growth that will most benefit learners as they prepare to become well-rounded professionals.

The *Professional Skills Development* tool presents test takers with realistic, complex scenarios and evaluates their responses in an open-response format. This approach encourages individuals to articulate their reasoning and reflects the reality that there are often multiple thoughtful ways to approach a situation. By mirroring the complexity of real-world dilemmas, the tool enhances ecological validity and provides richer insights than closed-response formats alone.

At Acuity, our team is focused on advancing products, services, and insights that foster holistic success. Working alongside academic program partners, subject matter experts, faculty, and learners, we strive to provide decision makers with tools

that deliver higher-fidelity perspectives and support meaningful insights and actionable feedback to guide student development.

This manual outlines the purpose and intended uses of the *Professional Skills Development* tool, while also documenting the scientific evidence underlying its development. The results to date are highly encouraging and we are committed to ongoing refinement. We welcome feedback, alternative interpretations, and opportunities for collaboration, as all of these efforts ultimately strengthen the assessment and its impact.

I would like to thank the many individuals who contributed to this work. Our dedicated internal team worked tirelessly to bring the assessment to life, and our external partners — including learners, educators, and program leaders — provided invaluable insights and feedback that shaped its evolution. The result is a refined tool that reflects the best of collaborative innovation, and we are proud and excited to share it with you.

Matt Holland
CEO, Acuity Insights

CHAPTER 1: INTRODUCTION

Over the last several years, the landscape of professional programs has changed, such that greater emphasis has been placed on assessing not only the technical skills of learners, but also the non-technical skills. Scholars have highlighted the importance of assessing and developing personal and professional skills across higher education including, but not limited to, psychology, law, business, international relations, etc. (Wilson et al., 2013; Dagilyte & Coe, 2014).

A variety of methods are used to assess personal and professional skills in higher education, including self- and peer-assessments, interviews, portfolios, direct observation, and OSCEs. However, these approaches often face limitations related to subjectivity, resource demands, and challenges in capturing complex, context-dependent behaviors, with ongoing concerns regarding reliability, validity, and scalability (Patterson et al., 2016; van der Vleuten & Schuwirth, 2005).

The situational judgment paradigm has a proven history of being a successful format for investigating and measuring skills related to social intelligence and non-technical professional skills (e.g., Patterson et al., 2016; Jackson & Chapman, 2012; Dagilyte & Coe, 2014; Hagen & Bouchard, 2016; McDaniel et al., 2007; Whetzel et al., 2008; Lievens et al., 2019; Lievens & Peeters, 2008).

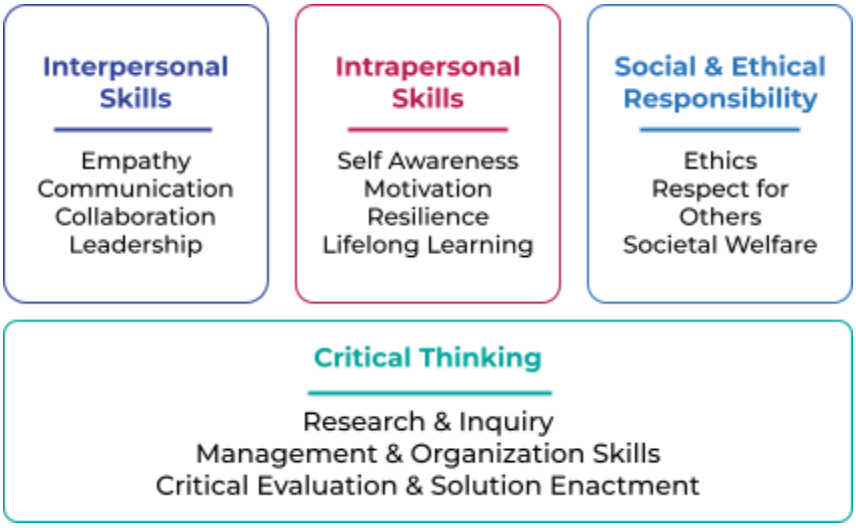
SJTs provide test takers with a series of hypothetical scenarios and assess the individual's response to the situation (Patterson et al., 2016). SJTs are a useful instrument for understanding how an individual would likely react or behave in a future setting, which makes them ideal for assisting in the admissions decision process (Patterson et al., 2016). There are two theoretical underpinnings of SJTs: 1) behavioural consistency theory and 2) implicit trait policies theory. Behavioural consistency theory posits that the best predictor of future behaviour is past behaviour; in this sense, an applicant's response in test situations provides an estimate of an applicant's future behaviour outside of the test context (Patterson et al., 2016). Implicit trait policies theory suggests that individuals' unique traits or characteristics impact their perceptions on what would constitute an effective or appropriate behaviour in various situations (Patterson et al., 2016). Taken together, this theoretical framework supports the notion that a sample of an individual's behaviour (for example, during an SJT) is informative of their behaviour outside of the testing environment, and of their future behaviour.

While SJTs have proven to be valuable tools in summative contexts (Acuity Insights, 2026), there remains an open question as to whether the same paradigm can be adapted for in-program formative use. Accordingly, there is a need for a similar tool that can both measure and support the development of non-technical skills.

Utilizing insights from faculty and learners, we conceptualized, created, and formally developed a situational judgment test (SJT) for *formative* use, which can help assess professional skills, develop them further, and track progress.

The *Professional Skills Development* tool is an assessment that measures and reports on several domains. It provides a detailed report that both quantitatively and qualitatively describes areas of relative strength and weakness. The model used by this *Professional Skills Development* tool is shown in *Figure 1.1*.

Figure 1.1. Professional Skills Development (PSD) Model.



By providing feedback on Interpersonal skills, Intrapersonal skills, Social & Ethical Responsibility, and Critical Thinking, and by drilling deeper into subcompetencies like Empathy, Self-Awareness, Ethics, and Research & Inquiry, an array of dimensions and related competencies, students are provided insights into their skills and can readily target areas for growth. The feedback includes ready-made and easily accessible learning resource materials (e.g., readings, short videos) and exercises that directly connect to the competencies evaluated by the assessment. These exercises provide a very direct and practical way for students to enhance their personal and professional skills.

For Programs, individual level results are summarized to identify learner development areas, and facilitate the early identification and remediation of learners in developing and refining these skills. Aggregate level results are also provided so that curriculum benefits can be tracked, and adjustments to programs can be made when certain aspects of professionalism are identified as being less well developed than others.

Key Features

The Professional Skills Development (PSD) tool by Acuity Insights is unique in higher education as non-technical skills are often assessed based on subjective evaluations and qualitative judgements. PSD offers a scientifically sound way to obtain standardized measurements of personal and professional skills.

For learners,

- They can quickly identify areas in which they are excelling and areas to target for growth
- They receive individualized feedback for each of the four dimensions (e.g., Interpersonal Skills) based on their responses to the scenarios. This feedback includes both comments related to what they did well, as well as tips for improvement.
- They also receive a curated list of resources and exercises for each of the subcompetencies (e.g., Empathy).
- They are provided with visual representation of their progress over time.
- They can take the assessment at several time points using the alternate parallel forms of the PSD.

For programs,

- They receive individual learner results for each dimension, as well as aggregate insights for the student cohort.
- They can use this information to help direct curriculum improvements (useful for Continuous Quality Improvement processes).
- They can easily identify which learners need more support.
- They are better able to document and provide evidence that the program helps to improve these professional skills.
- They can use the results to support the evaluation of accreditation requirements.

Uses of the Professional Skills Development tool

Intended Purposes and Use

Consistent with the principles outlined in the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014), the validity of an assessment is defined in relation to its intended uses. Accordingly, the Professional Skills Development (PSD) assessment was designed to serve several specific purposes, outlined below:

- **Provide quantifiable measurement**
    - for competency areas related to personal and professional skills (specifically interpersonal skills, intrapersonal skills, social & ethical responsibility, and critical thinking).
- **Meaningfully differentiate between performance levels**
    - Individuals with more developed skills in the targeted areas will score higher.
- **Effectively capture and support changes in development**
    - Support skill development by providing feedback that identifies areas to target for improvement and by providing resources that directly connect to the constructs being assessed.
    - Provide a means for these skills to be tracked over time through repeated assessment. Multiple forms of the PSD are available to support this usage.

There are many benefits associated with the use of this tool, including:

- Helping learners graduate job-ready by tracking soft skills and professional competencies.
- Helping learners build the soft skills that define strong professionals: like giving and receiving feedback, managing conflict, and navigating professional relationships.
- Giving learners an opportunity to practice how to handle miscommunication, difficult conversations, and ethical gray areas: the situations where new grads often struggle most.
- Helping learners gain the confidence to lead and collaborate across teams, even when things get messy, fast-paced, or ambiguous.
- Identifying learners who need support early, reducing the risk of costly remediation processes later in the program.

Generally, the purpose of this tool is to foster learner development and it is meant to be used for formative purposes. Although scores are made available, these are included to support the identification of areas of strength and pinpoint areas to target for development. These scores are not intended to be directly integrated into a formal summative grade, and use of the scores in that fashion is discouraged.

Applicability Across Program Types

The PSD tool was designed for broad applicability across diverse program types. Its dimensional framework emerged from an extensive analysis of competency frameworks spanning multiple disciplines and captures the core transferable skills common across them. As a result, PSD is well suited for use in a wide range of educational contexts. Sample applications are outlined below.

MBA Programs

In MBA contexts, PSD can be used to:

- Benchmark durable, non-technical skills early in the program
- Guide coaching, advising, and targeted student support
- Integrate structured reflection into professional development courses
- Strengthen AACSB assurance of learning (AoL) evidence across cohorts
- Measure growth through repeated administration over time

Undergraduate Business Programs

Undergraduate business programs can apply PSD:

- In courses such as HR, ethics, communication, professional development, first-year orientation, capstone strategic management, and senior “life design” courses
- As individual assignments or structured reflective projects
- To support student retention initiatives and targeted remediation
- For continuous Improvement for AACSB assurance of learning outcomes and "closing the loop" activities

Health Sciences Programs

In health sciences education, PSD can be used across courses including first-year labs, Foundations of Clinical Practice, Interprofessional Communication, Patient Care and Professionalism, clinical readiness or pre-clinical skills courses, and senior practicums or clinical rotations. Applications include:

- Pre- and post-placement administration to assess developmental growth
- Required learning activities paired with guided self-reflection
- Longitudinal use across multiple curriculum stages to compare early, pre-clinical, and clinical development
- Integration into interprofessional education sessions to strengthen teamwork and communication across disciplines
- Supporting ethical reasoning, empathy, professionalism, and self-awareness in patient care contexts
- Helping students better understand their responses to complex clinical and interpersonal scenarios
- Providing faculty with actionable insights into cohort strengths and skill gaps in clinical readiness, collaboration, and professional judgment

Engineering Programs

Engineering programs can use PSD in conjunction with courses such as Engineering Ethics and Professional Practice, Introduction to Engineering Design, first-year orientation courses, senior capstone design, and aerospace engineering professional preparation. Typical uses include:

- Required course activities paired with structured reflection assignments
- “Parallel cohort” designs comparing first-year and senior students to evaluate developmental progression
- Enhancing ethical reasoning, self-awareness, communication, and teamwork through reflection and feedback
- Measuring changes in teamwork, communication, and professional readiness from entry to graduation
- Supporting ABET accreditation and continuous program improvement efforts

Linkages between key competency frameworks across disciplines and the PSD framework are presented in Appendix A.

Principles of Use

The *Professional Skills Development* tool was carefully developed and researched to provide the most useful protocols for measuring specific domains of personal and professional skills. Any single test has inherent limitations; however, when used appropriately, the *Professional Skills Development* tool can be extremely useful in the entire process of collecting objective data on these domains, providing targeted resources to support student development, and for tracking improvements over time.

The *Professional Skills Development* tool is not intended to be the sole method used for diagnostic decision-making, assessing learners, or drawing unsupported interpretations. In order to obtain a comprehensive view of an individual, the assessor must combine the results with information gathered from other quantitative or qualitative information that is available about the learner. For example, information obtained from behavioural observation (e.g., presentations), other formal assessments (e.g., quizzes, essays, tests), and through conversations/discussions with the student and other informants. Considering information from a multitude of sources helps the assessor make a more comprehensive and ecologically valid assessment. It is important to consider any factors that may bias results, such as socially desirable responding, misunderstanding of item content, or carelessness in responding. The PSD focuses on specific skills that are amenable to development, and is *not* a personality test that describes static traits of a person.

Respondent Rights

Individuals completing the *Professional Skills Development* tool should:

- be informed about how results will be used and who will have access to them (informed consent), and
- be able to withdraw from the process at any point – all blank responses will be assigned a score of 0 so the program should carefully note any learners who decide not to complete the test, and interpret results with caution.

Before administering the *Professional Skills Development* tool, the purpose of the assessment should be explained clearly to the respondent. Additional details can be provided after the test is completed. It is not sufficient to simply send a link to the test; instead, the administrator should place the assessment in the appropriate context, emphasize confidentiality, clarify that results will not have adverse

consequences, and help ease any concerns respondents may have about participation.

Informed consent should be secured beforehand. This tool should not be used to mislead or pressure participants into sharing information they would prefer to keep private. Informed consent means giving respondents enough relevant information so they can make a conscious and informed decision about whether to proceed. At minimum, respondents should be told the reasons for using the assessment and how their results will be applied. Importantly, providing initial consent does not bind respondents permanently; they should retain the right to withdraw at any time without penalty.

Additional Terms of Use information for Acuity's Terms and Conditions statement.

CHAPTER 2: INTERPRETATION

This chapter describes the intended interpretation process for the Professional Skills Development tool.

What the Professional Skills Development tool Scores Tell You

Overarching Construct Definition

The four dimensions of the *Professional Skills Development* tool all fall under the umbrella of personal and professional skills. More explicitly: the dimensions collectively relate to the ability to effectively reflect on and communicate responses to interpersonal and professional dilemmas using critical reasoning and social interpretation.

Dimensions

To better serve formative purposes, it was important to distill the overarching construct into a set of dimensions and underlying competencies that could be measured and then, based on the results, used to foster development.

After extensive examination of a large number of frameworks (see *Section Framework: Dimensions and Competencies* for additional details), we identified four specific quantifiable dimensions: Intrapersonal Skills, Social & Ethical Responsibility, Interpersonal Skills, and Critical Thinking. Within each dimension, the content coverage includes specific competencies. Definitions for these dimensions and subtending competencies are provided in *Tables 2.1 to 2.4*.

Table 2.1. Interpersonal Skills: Dimension and Competencies Definitions.

Interpersonal Skills
Dimension
Building meaningful connections and working effectively with others by communicating clearly, showing empathy, collaborating toward shared goals, and inspiring others towards positive outcomes.
Competencies
Communication: The ability to effectively listen and exchange information, ideas, and feelings through verbal and nonverbal means.
Empathy: The ability to compassionately respond to others' feelings and perspectives while considering how others are impacted by the context and actions of those involved.
Collaboration: The ability to work effectively with others, leveraging varied perspectives, sharing responsibilities, and contributing to collective success.
Leadership: The ability to inspire, guide, and influence others toward achieving common goals, while fostering a healthy environment for the team.

Table 2.2. Intrapersonal Skills: Dimension and Competencies Definitions.

Intrapersonal Skills
Dimension
Understanding and regulating oneself by recognizing emotions, biases, and behaviors, staying motivated, adapting to challenges, and committing to lifelong learning for personal growth.
Competencies
Self Awareness: The ability to understand one's own emotions, thoughts, biases and behaviors, and how they affect oneself and others.
Lifelong Learning: A continuous commitment to learning throughout life, fostering personal growth, and professional development in a changing context.
Resilience: The ability to adapt effectively by seeking support, adjusting behaviors, and identifying strategies to handle stressful and unfavorable challenges.
Motivation: The drive that regulates human behavior towards achieving personal and professional goals.

Table 2.3. Social & Ethical Responsibility: Dimension and Competencies Definitions.

Social & Ethical Responsibility
Dimension
Recognizing and respecting people's differences, upholding ethical principles, and contributing to the well-being of society through responsible actions.
Competencies
Respect for Others: The ability to show awareness and respect towards personal differences, and effectively interact with people from different backgrounds and experiences. *Ethics*: The ability to act with integrity, honesty, and accountability to uphold professional standards and ethical principles. *Societal Welfare*: The ability to prioritize and recognize opportunities to improve the well-being of society and the community at large.

Table 2.4. Critical Thinking: Dimension and Competencies Definitions.

Critical Thinking
Dimension
The ability to gather information, evaluate options, and find effective solutions to problems while efficiently managing resources and risks.
Competencies
Research & Inquiry: The ability to gather information from multiple sources, while verifying the accuracy and relevance of this information. *Management & Organization*: The ability to organize all the gathered insights, and to prioritize next steps efficiently with time and resources in mind. *Critical Evaluation & Solution Enactment*: The ability to identify effective strategies to solve complex problems, while evaluating the risks, costs, and strengths of multiple alternative solutions.

PSD Interpretation

Responses on the Professional Skills Development tool are rated on a scale from 1-5, where 1 is the lowest possible rating and 5 is the highest. There is one exception, questions that are left blank receive a score of 0. The score for each dimension is then computed using the average rating for all responses for questions that are specific to each domain. *Table 2.5* shows the categorization of the Dimension Scores based on these average scores. These categorizations are based on the distribution of scores observed to date.

Table 2.5. Guidelines for Evaluating Dimension Scores.

Guideline	Range	Score Meaning
High Competency	Score ≥ 4	The learner consistently demonstrates this skill across contexts and is well-prepared to apply it in real-world settings. Further development should focus on refining and extending already well-established abilities.
Moderate Competency	Score ≥ 2.5 and < 4	The learner demonstrates emerging strength in this area but may apply these skills inconsistently. They show an understanding of core principles but would benefit from continued development to improve consistency and depth of application.
Developing Competency	Score < 2.5	The learner is in the early stages of developing this skill and may require additional support and practice. Application may be limited, particularly in more complex situations. Development should focus on building a strong foundation in the underlying principles.

Step by Step Interpretation Guidelines: Learner Development

As programs use PSD as a tool to foster student development on this skillset, the following guidelines for interpretation and use of the tool are provided.

Step 1 - Evaluate the quality and engagement of the learner's responses.

Uniformly low scores (e.g., below 2 across all four domains) may indicate limited engagement with the assessment (e.g., low effort or motivation, especially when coupled with very short response time). In such cases, it is important to explore potential reasons for disengagement as a preliminary step.

However, low scores may also reflect genuinely underdeveloped competencies across domains. Interpretation should therefore consider both possibilities, using additional evidence (e.g., response quality, completeness, or other performance indicators) to distinguish between low engagement and low skill level.

Step 2 - Interpret the dimension scores.

Check the overall dimension score results for the student. Each domain score is categorized as a high competency, moderate competency, or developing competency area. Generally, scores in the lowest performance range are indicative of competency areas that should be targeted for improvement. Invite learners to review their results and feedback and complete the reflection prompts and learning exercises.

Step 3 - Use the qualitative information provided with the reports to gather more insight into their results.

Although scores are only provided at the dimension level, the understanding of these scores can be enriched by leveraging the qualitative feedback provided as part of learners' results. Each domain score consists of a variety of competencies. So, even when an overall domain score is in the high or moderate range, the qualitative information provided can be helpful in identifying specific skill gaps within domains that can further benefit the development of these skills. Also, identifying and fully recognizing areas of strength can be helpful for students in leveraging these skills to promote positive outcomes.

Step 4 - Create a plan for improving these skills.

These results and exercises can be used in multiple ways to support learners in improving these skills.

The learners can work with their advisor and/or supporting faculty to discuss the feedback provided on the report and determine a suitable plan for development. The Professional Skills Development tool provides resources and exercises that cover all the relevant competencies, so it is easy to pick and choose resources that are best suited for specific learners. Alternatively, learners may be allowed to go through all of the exercises so that their skills can be enhanced across all of the domains and competencies.

Step 5 - Compare test results over time.

Maximum benefit can be derived from the tool if the assessment is used multiple times at predetermined intervals. This practice of repeat usage helps track development that occurs over time and further consolidates that experience and knowledge that the assessment provides. Remind learners that this assessment is designed to support development, not evaluate them as “good” or “bad” learners. Growth happens over time.

Program Level Application

The Program results are designed to provide additional information to serve potential needs of the program itself. These results provide aggregative performance levels for the domains, a summary of support recommendations both in tabular as well as graphic form, and individual learner results are presented succinctly and conveniently.

Ideally, these program level results will foster using the PSD to

- Spot trends: Look for areas where learners are excelling or lagging to inform practice.
- Guide individual support: Use the Individual Learner Results to flag students for additional support.
- Integrate into advisory programs: Share competency areas as part of academic advising or mentorship.
- Inform teaching strategies: Use insights to strengthen classroom activities that build targeted competencies.
- Support curriculum planning: Look across cycles to track growth and inform course design.
- Contribute to outcomes and accreditation processes: Use aggregated results to support outcomes review and continuous improvement.

Case Examples

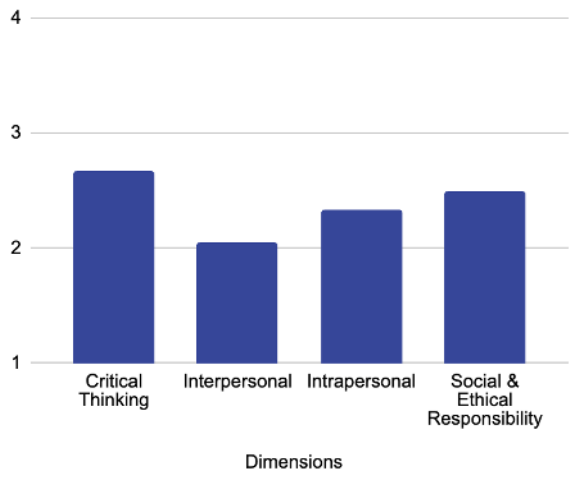
A series of simple case examples is provided to exemplify the use of the Professional Skills Development tool in practice.

Case 1 - Student: Developing/Moderate Results across Domains

Background. Pooja is a first year Engineering student who is just beginning her program. She is excited to be in the program and seemed very interested in the Professional Skills Development tool. The program decided to use the assessment with students near the beginning of the program to emphasize the importance of these skills from the outset of the program, and also to capture initial baseline skill levels for these domains.

Results. Pooja's scores were in the low to middle portion of the moderate range as reflected in *Figure 2.1*.

Figure 2.1. Graphic representation of Pooja's results (Case Study 1)



After the test was completed, a brief preliminary feedback session with Pooja indicated that she enjoyed the exercise and was looking forward to receiving the feedback on his responses to the scenarios.

Interpretation. Pooja's results are very consistent with expectations for a new student entering the program. Her scores in the moderate range indicate that she brings with her a reasonably solid foundation of skill in the area of personal and professional competency. Her lowest score was in the Interpersonal domain, and

through examination of her narrative feedback was able to hone in on empathy as a specific component of interest and for improvement.

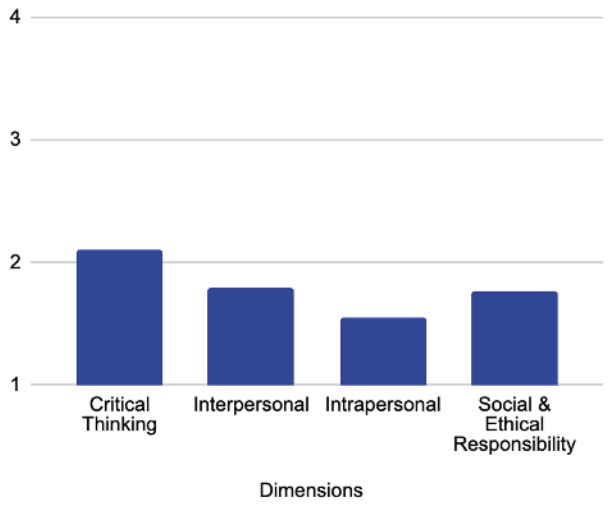
Actionables. Pooja comes to the program with a good foundation for this skill set so there are no particular concerns for her and she embarks on the program. With this being early on in the program, exposure to all of the resources and exercises offered as part of the feedback from the assessment should be valuable. Depending on time available in her schedule, it might make sense for Pooja to start with the specific resources and exercises that target the Interpersonal domain, and Empathy specifically.

Case 2 - Student: All developing competency scores, potentially unengaged with the task

Background. Adele is a third year, top notch, dedicated medical student who is generally near the top of her class in terms of her academic grades. She is friendly and interacts well with her classmates. Although Adele is very adept and interested in the technical elements of the program, she sees personal skills as much less important and prefers to concentrate her efforts on what she sees is most relevant to the medical profession.

Results. Adele's scores on the Professional Skills Development tool are shown in *Figure 2.2*. Surprisingly, given her academic standing, her scores on the assessment were quite low.

Figure 2.2. Graphic representation of Adele's results (Case Study 2)



After the test was completed, a brief preliminary feedback session with Adele indicated that she had been consumed with her course load and, while she recognized that completing the SJT was required, she knew it wasn't part of her formal grade and only gave it a cursory effort.

Interpretation. Adele's results suggest that she needs support in the area of personal and professional skills. The low scores may reflect poor skills or a lack of motivation or focus to properly complete the task at hand.

Actionables. Initial steps may involve reiterating to her the importance of these skills for her development toward her being a well rounded professional. The resources and exercises should be completed and in Adele's case will provide an additional data point regarding where she stands in achieving proficiency in these competency areas. A subsequent administration of the SJT, perhaps at a time when her workload is less, with a more concerted effort would also be valuable to evaluate and update her progress on development of these skills.

Case 3 - Program level application: University of Waterloo, Pharmacy.

Background. Second-year pharmacy students completed the SJT as part of a seminar course activity, with results incorporated into the School's midpoint assessment. This assessment traditionally includes two components designed to mirror licensing examinations: a 100-question multiple-choice examination (MCQ) and a five-station Objective Structured Clinical Examination (OSCE), which also serves as the final assessment for a Professional Practice course.

Beginning in 2024, the program introduced a third component—an in-program SJT—to evaluate its utility in identifying students requiring remediation in non-technical competencies. This application has been described in prior work (Archbell et al., 2025; Bynkoski et al., 2025). The SJT was re-administered in 2025 with similar objectives.

The program's goals were to:

1. Determine the utility of the SJT at the midpoint assessment in identifying students requiring remediation in non-technical domains (e.g., ethical judgment, interpersonal dynamics)
2. Evaluate associations between SJT performance and other in-program assessments to better understand the measurement of professionalism-related competencies
3. Examine student uptake of remediation opportunities related to personal and professional skill development

Results. In 2025, a total of 77 students required no remediation across assessments. Three students were identified for both academic and professionalism-related remediation. Eleven students demonstrated satisfactory academic performance but were identified by the SJT as requiring support in personal and professional competencies.

Interpretation. Remediation based on MCQ and OSCE performance was not associated with SJT outcomes, indicating that students identified for academic or clinical remediation are not necessarily those experiencing challenges in professional judgment and reasoning. This finding is programmatically important, as it demonstrates that the SJT provides incremental value by identifying learners who may otherwise go undetected using traditional assessment approaches.

Actionables. The inclusion of an in-program SJT supports the early identification of learners requiring development in non-technical competencies, particularly in the early years of training. This enables more targeted intervention and a more comprehensive approach to supporting student development.

CHAPTER 3: TEST DEVELOPMENT

In this chapter, the test development process is described beginning with a summary of the test's background and conceptualization, then the original prototype is described along with preliminary psychometric findings. Next, we explain the learnings from the prototype stage and the changes that were implemented as the test was refined into the current version.

Overview

The Professional Skills Development (PSD) tool was developed over several years, with initial work beginning in October 2022. Development progressed through multiple phases, including early conceptualization, prototype development and proof-of-concept testing, and the subsequent creation and evaluation of a revised version intended for early adopters. Following additional refinement, this process culminated in the current iteration of the tool. During the conceptualization phase, several core objectives were established. Chief among these was the need for a psychometrically sound assessment capable of producing scientifically defensible measurement outcomes, while also being designed for seamless integration into existing educational programs.

Background

Personal and professional competencies—such as empathy, collaboration, communication, and ethical reasoning—are central to success in both educational environments and later professional practice (Iqbal et al., 2025). These skills support not only academic achievement but also longer-term career outcomes (McGill, Ali, & Barton, 2020; Saunders & Bajjaly, 2022). Despite their recognized importance, many training programs struggle to assess and cultivate these competencies effectively, often due to constrained resources, limited standardization in assessment approaches, and insufficient opportunities for individualized feedback (Alt, Naamati-Schneider, & Weishut, 2023; Saunders & Bajjaly, 2022). Formative situational judgment tests (SJTs) represent one potential response to these challenges. By presenting learners with structured, contextually rich scenarios, SJTs can assess and support the development of interpersonal and professional skills in a low-stakes, feedback-oriented format (Patterson et al., 2019; Sahota et al., 2023; Cullen et al., 2022). Additionally, SJTs may help identify learners who could benefit from early support or targeted professional development (Cullen et al., 2017, 2020). Within this framework, social and personal intelligence is conceptualized as the capacity to

reflect critically on, and respond effectively to, interpersonal and professional situations.

Within educational contexts, SJTs have been applied across the full training continuum, including admissions, undergraduate education, postgraduate selection, and ongoing professional development. They have been implemented in multiple formats—such as free-text responses, multiple-choice questions, and video-based scenarios—and used for both formative and summative purposes (Ballejos et al., 2024; Lievens & Motowidlo, 2016; Patterson & Driver, 2018; Saxena et al., 2024). A recent systematic review identified eight studies that employed institution-specific, in-house SJTs primarily for formative use (Foucault et al., 2015; Kiessling et al., 2016; Frohlich et al., 2017; Goss et al., 2017; Antes et al., 2020; Graupe et al., 2020; Ludwig et al., 2021; Reiser et al., 2021). Collectively, these studies suggest meaningful associations between SJT performance and indicators of professional behaviour. For example, higher levels of agreeableness, conscientiousness, extraversion, and openness have been linked to stronger SJT performance, whereas neuroticism has been associated with lower scores. Several studies (Antes et al., 2020; Foucault et al., 2015; Goss et al., 2017) explicitly positioned SJTs as instructional tools, though most did not directly assess changes in knowledge or attitudes. In one example, Goss et al. (2017) incorporated feedback for lower-performing students, an approach mirrored in subsequent work. Overall, these findings are consistent with broader perspectives on professional identity formation and support the potential educational value of formative SJTs.

Notably, all formative SJT studies identified in the review relied on close-ended, text-based scenarios using ranking or multiple-choice response formats. However, several studies employing video-based SJTs (Frohlich et al., 2017; Graupe et al., 2020; Reiser et al., 2021) demonstrated promise in supporting the development of desirable professional attributes. Open-ended response formats offer an additional advantage by providing rich qualitative data beyond what can be captured through close-ended items. They also avoid constraining respondents to a single “correct” answer, acknowledging that complex professional dilemmas often allow for multiple defensible responses. For these reasons, the current version utilizes open-ended responses.

Prototype

The prototype version of the PSD assessment shared many conceptual features with later iterations. It consisted of 14 scenarios, ten of which were video-based and featured human actors, with each video lasting approximately one minute. The

remaining four scenarios were shorter (approximately 20 seconds each) and used animated characters rather than live actors. For each scenario, participants responded to a combination of open-ended and multiple-choice questions targeting the constructs of interest. Open-ended responses were scored on a 1–9 scale by experienced raters who had received training on the scoring criteria. Reflecting the developmental focus of the assessment, no time limits were imposed on response completion in the prototype phase.

Initially, the assessment was designed to measure distinct constructs, mirroring those assessed in the Casper test: Collaboration, Communication, Empathy, Fairness, Ethics, Motivation, Problem Solving, Resilience, and Self-Awareness. However, early development work indicated that assessing all constructs individually would result in an assessment of impractical length for most educational contexts. Consequently, a revised strategy was adopted in which these constructs were consolidated into a smaller number of broader components. This approach was intended to strike a balance between obtaining meaningful, interpretable results and maintaining a test duration that would be feasible for routine implementation.

Identified Improvements

Based on the research results (see Appendix B for details), experiences with the prototype, and the feedback from both students and programs, the following refinements were identified for the next version:

- Need to measure a limited number of constructs.
    - Learning from prototype: Explored measuring a large number of constructs but to do so in a psychometrically sound manner would result in an overly lengthy assessment that would severely limit its utility.
    - The improvement: The revised version measures four domains, and the report also provides qualitative feedback on competencies within each of those domains. Also, the revised version is split up into three modules that take about 35 minutes each to complete. This modular set-up provides a reasonable total test time and provides more flexibility to accommodate time constraints by allowing the modules to be completed in separate test sittings.

- Need to revise the framework.
    - Learning from prototype: Although the original framework was well received overall and produced decent psychometric performance, it

was clear that we needed to find a model that would be consistent with a broader range of competency frameworks.
    - The improvement: A thematic analysis was conducted across a wide and comprehensive examination of frameworks (see section entitled Framework: Dimensions and Competencies for more details) leading to the four domain model. The model also can be mapped onto existing frameworks very easily since these four dimensions are general themes from competencies across many different frameworks.

- Need to make the solution easy to integrate into a program's operations.
    - Learning from prototype: From conversations with programs and from trial runs, it is clear that the solution would need to fit into an existing curriculum and not unduly increase the burden on learners, teachers and programs.
    - The improvement: The modular structure provides administration flexibility. In addition, the report includes "ready-made" resources and learning exercises that optimally would be used in conjunction with a facilitator (student advisor, professor, mentor) from within the program, but can also be used with little additional support depending on the needs of specific programs.

- Need to use open-ended responses
    - Learning from prototype: Both open-ended and multiple choice questions were tested, and open-ended responses performed substantially better (e.g., open-ended responses showed far better reliability).
    - The improvement: This revised version contains only open-ended questions for the scenarios.

Current Version of the Professional Skills Development tool

Leveraging the experiences, psychometric results, and user feedback from the prototype, the current version was developed. This development work involved determining the optimal framework, most effective test structure, storyline, scenario, and question development, and the implementation of AI scoring and report generation mechanisms.

Framework: Dimensions and Competencies

To determine common central constructs to use as a focal point for our tool, we reviewed North American competency frameworks across professions and compiled a list of 96 unique competencies across 10 frameworks, including their definitions and sub-competencies. Then, using a thematic analysis approach, we identified key domains that unify competencies based on shared themes (e.g., “interpersonal skills”). The thematic analysis revealed four unifying core non-technical competency domains across multiple frameworks: (1) Intrapersonal skills, (2) Interpersonal skills, (3) Social and Ethical Responsibility, and (4) Critical Thinking (Sitarenios et al., 2026a, 2026b, 2026c; Dore et al., 2025). The identification of the four broad domains offers programs a clear and focused framework for enhancing student learning. By narrowing attention to these core areas, programs can concentrate on shared and essential non-technical competencies, making the development process more efficient and impactful.

Competency Mapping

There are a vast number of frameworks across professional programs. While we examined a variety of frameworks, we focused our analysis on the most commonly used frameworks from four major professional program types: medical education, business education, engineering, and law. Frameworks included:

- Medical Education: AAMC, ACGME, CANMEDS
- Business Education: AACSB, NACE
- Engineering: CEAB, ABET
- Law: LSO, WCCP, ABA

Thematic Analysis

We then applied a thematic analysis to those competencies with the intention of uncovering unifying themes. We hypothesized, discussed, and ultimately identified four broad clusters under which those competencies fell.

Furthermore, we used the competencies under each cluster to inform the definition and development of our dimensions.

Literature Review

As a means of corroborating our determined dimensions, we reviewed the literature and found consistency with, and support for, the dimensions included in the Professional Skills Development tool. For example, three studies emphasized interpersonal components leveraging SJTs for developing communication skills (Kiessling et al. 2016; Ludwig et al. 2021; Reiser et al. 2021). Van de Camp et al. (2004) examined 166 Medline articles and uncovered three primary themes: *Interpersonal* professionalism which includes effective interactions with others in meeting professional demands; *Public* professionalism which reflects the need to account for societal considerations; and *Intrapersonal* professionalism which pertains to elements such as flexibility and lifelong learning. These elements converge strongly with the four dimensions of the Professional Skills Development tool.

Test Structure

Behavioural Tendency Questions

This tool uses behavioural tendency questions to ask applicants what they *would* do in a given situation. This is fundamentally unique from knowledge questions which ask applicants what they *should* do in a given situation (McDaniel et al., 2007). The use of behavioural tendency questions are important for two distinct reasons. First, behavioural tendency questions have shown to correlate more with assessments of personality rather than measures of technical-knowledge abilities (McDaniel et al., 2007). Secondly, behavioural tendency questions have shown to produce lower demographic group differences across applicants of varying gender and race relative to knowledge-type questions (Whetzel et al., 2008).

Open-Ended Responses

The open-ended response (*i.e.,*constructed response) format of this SJT means that applicants are not forced to select a predetermined response, but rather allows for diversity and uniqueness of responses. This response option, which avoids specifying a particular correct response, tends to produce lower demographic differences relative to close-ended response options (Lievens et al., 2019), is less susceptible to faking by applicants (Lievens & Peeters, 2008), and has the ability to better

discriminate between applicant responses relative to close-ended questions (Funke & Schuler, 1998). Open-ended responses also provide rich quantitative information that can be used for reporting purposes and derives more insight into students' responses than could be derived from multiple choice responses.

PSD paradigm Components

The test components are organized in a set of three modules composed of storylines, scenarios, and questions. See *Figure 3.1* which graphically portrays the test set up:

Figure 3.1. Test Structure.

Modules

The PSD test is split into 3 modules, each taking around 30 minutes to complete.

The content for each module targets a combination of two domains:

- one module targets *Interpersonal Skills* and *Social & Ethical Responsibility*
- one module targets *Social & Ethical Responsibility* and *Intrapersonal Skills*
- one module targets *Interpersonal Skills* and *Intrapersonal Skills*

Since Critical Thinking is part of the process that applies to resolving all dilemmas, this component is included in all three modules.

To foster student engagement, all scenarios within a module are linked to a common storyline. For example, one module might discuss the storyline of Jack, who is looking for a new job. One scenario might involve Jack talking to a friend about the job search process. Another scenario in this module might involve Jack applying to a specific job. Another scenario might connect to his feelings during the job interview. And, it might conclude with a scenario about the decision he has to make regarding job acceptance.

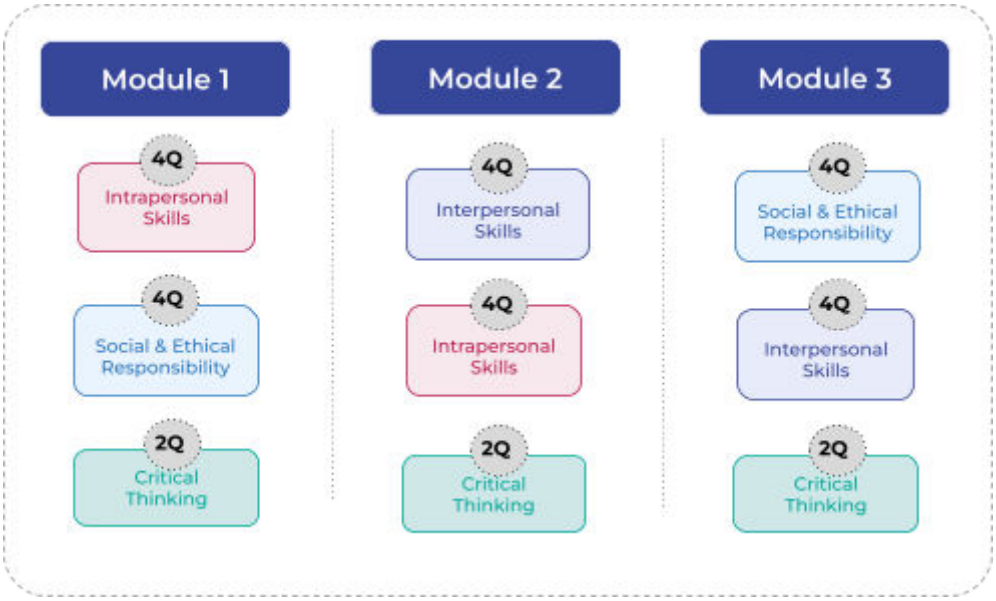
Scenarios

Each module consists of 5 scenarios. Two of the five scenarios are video-based scenarios, and three are word-based. The video-based scenarios feature videos with 3D animations. We intentionally avoided overly profession-specific scenarios to enhance the applicability of the assessment across program types. Highly specific scenarios risk conflating non-technical skill with domain-specific knowledge, thereby weakening the validity of score interpretations (Lievens, Peeters, & Schollaert, 2008). Instead, scenarios are grounded in academically relevant contexts that reflect situations students are likely to encounter in their educational journeys (e.g., use of AI in assignments, navigating academic pathways, interactions with faculty).

Questions

For each scenario, students are asked two questions for which they have seven minutes to complete each question. Across the 3 modules and 15 scenarios, there are a total of 30 questions for the entire test. The setup is depicted in *Figure 3.2*.

Figure 3.2. Question Distribution by Module.



Typed Responses

All questions require typed responses. Learners are given six minutes to respond to the two questions associated with each scenario, and responses are limited to a maximum of 200 words. These constraints are intended to ensure the assessment can be completed within a reasonable timeframe while also providing clear expectations regarding response length.

The time and word limits serve several purposes. First, they help guide learners toward concise, focused responses, reducing the likelihood that they feel compelled to produce overly long or time-intensive answers. Second, the word limit helps minimize variability in response length that could unduly influence scoring. Third, applying a consistent word limit across all test-takers helps maintain comparability in response scope, including for those receiving extended time accommodations.

Storyline, Scenario and Question Development

The creation of a formative SJT is a complex and iterative process that involves content development experts from several development teams at Acuity Insights. Initial drafts of stories, scenarios, and questions were reviewed and re-reviewed before being included in a test. All components went through several iterations before being finalized for inclusion. Once included in a test, scenarios and questions were then evaluated scientifically to establish the psychometric soundness. The

following sections provide an overview of the foundational pieces involved in the content development process.

Storyline development and script writing process for both video and word based scenarios

The Content Production team started by developing a module storyline that served as a backbone for each module and steered the direction of five scenarios in each module. Each storyline was intended to cover a combination of two dimensions that were selected/tagged at the very initial stage. These storylines were then reviewed by the broader Content and Research teams to ensure that the storyline is coherent, nuanced and realistic, and had sufficient layers of complexity to develop five scenarios and their respective questions. Since the dimensions and their respective competencies were proposed in the very initial phase of storyline development, both teams also reviewed the accuracy of initially tagged competencies.

Once the storylines were approved by both teams, the Production team converted the scenario ideas into fully fleshed out scenario scripts for both video and word based scenarios. Representatives from the Research and Content teams then reviewed the scripts and provided feedback (e.g., challenged whether the script is in fact tapping the right construct; assessed if the script is realistic enough; assessed if the scenarios follow educational principles; ensured that each scenario has sufficient layers to create two unique questions, targeting one competency per question; considered whether the script will be of relevance to the students; prevented any potential overlap in the scenarios within and across modules) until a consensus was reached on the scripts' content. The scenario scripts were finalized before video production and question writing took place.

Stimuli production

After scripts were finalized, they moved to the production phase. The video-based scenarios were produced using AI generated actors (i.e., avatars). Beginning in October 2024, we conducted a structured evaluation of different AI video generation tools to determine whether this would provide a viable solution.

While AI-generated avatars accelerated initial video production, careful scripting, post-production editing, and oversight by human experts was required to produce videos of high quality standards. After initial generation of a video, visual and audio post-production editing was necessary to bring the quality up to standard. We refined video characteristics such as format and background sets, as well as Avatar costumes and voices. Advanced editing techniques were applied to fix timing (lip sync, pacing, pauses) and visual artifacts (uncanny expressions, erratic gestures,

glitch frames). Avatars were also recast when needed to maintain consistency and quality. Sometimes, video format redesign was needed to accommodate Avatar limitations.

AI video technology itself is evolving quickly, which means we'll need to keep adapting as new options emerge to improve the video quality. The use of animated videos is also being considered given the challenges in rendering AI generated videos that are of satisfactory quality and do not compromise the effectiveness of test delivery, and the student experience.

Question writing process

As shown in Figure 3.2 above, each module consists of a combination of two dimensions and Critical Thinking. Consider a module targeting *Interpersonal Skills* and *Intrapersonal Skills dimensions*. Questions were constructed such that all competencies of both dimensions were represented at least once in the module, in addition to two questions targeting critical thinking. This ensures a balanced representation of all competencies in the test.

Once the questions were drafted, they went through the same iterative review process where all teams review the questions for their clarity, accuracy in probing for the targeted competency, and ability to generate a nuanced response. All questions were written to be scenario-specific (rather than generic).

Sample storyline, scenario and questions

To illustrate how these components come together within the test itself, this section provides an example of a storyline, scenario and related questions.

Basic Storyline

Two close friends are co-leading a group project together for a major college assignment.

Scenario

Kendra: Hey, I finalized the project schedule so we can meet the deadline. This project is a huge part of our final grade.

Alex: This seems too rigid. The group members want to talk about task division before we finalize the schedule.

Kendra: We're out of time! If we don't start now, we risk missing the deadline and losing marks.

Alex: But we might lose our group's trust by not involving them.
And if you keep ignoring my suggestions, I feel like it's going to affect our friendship.

Questions

Q1: COLLABORATION

How could the co-leads, Alex and Kendra, have collaborated more effectively with their group to create a clear and realistic project plan?

Q2: COMMUNICATION

Imagine that a group member is concerned about the finalized schedule. How should this group member share their concerns with the co-leads?

Scoring and Feedback

We developed a tightly-coupled AI scoring and feedback generation system to evaluate specific, observable behaviours related to the four dimensions: Intrapersonal skills, Social & Ethical Responsibility, Interpersonal skills, and Critical Thinking (Walsh, Ivan, & Sitarenios, 2026; Walsh et al., 2025; Walsh et al., 2026a, 2026b). This AI Scoring and Feedback generation system comprises three distinct components:

1. Feature Extraction
2. Regression
3. Feedback Generation

The outputs from the first (feature extraction) component are used as inputs to the other two components. We describe each of these components in greater detail next.

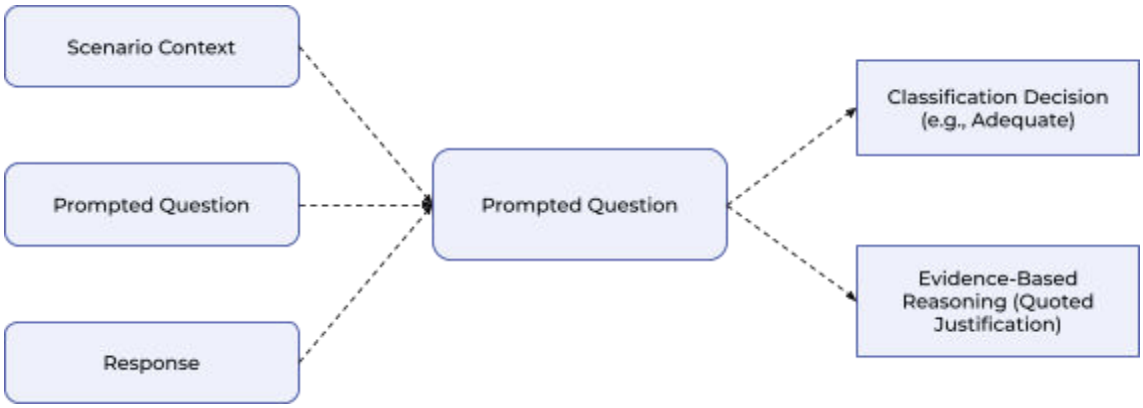
Feature Extraction

The feature extraction component of this system classifies responses according to whether and to what extent they exhibit specified features[1]. We evaluate 44 distinct

[1] The criteria used in content development are also used as features in evaluating the responses. We refer to these criteria as 'features' in the sections concerning the AI models, following ML best practices for these terms.

features of responses across all assessment responses, but responses for any single question are only evaluated on a subset of three to eight features corresponding to the competency assessed by that question. We constructed 44 LLMs-as-Judges, one for each feature, using the architecture shown in *Figure 3.3*. Each LLM-as-a-Judge uses a dynamic prompt template including instructions, constraints, and examples for evaluating the specified feature. Each judge takes as input the scenario context, question, and response and returns a classification decision and associated reasoning.

Figure 3.3. LLMs-as-Judges architecture.



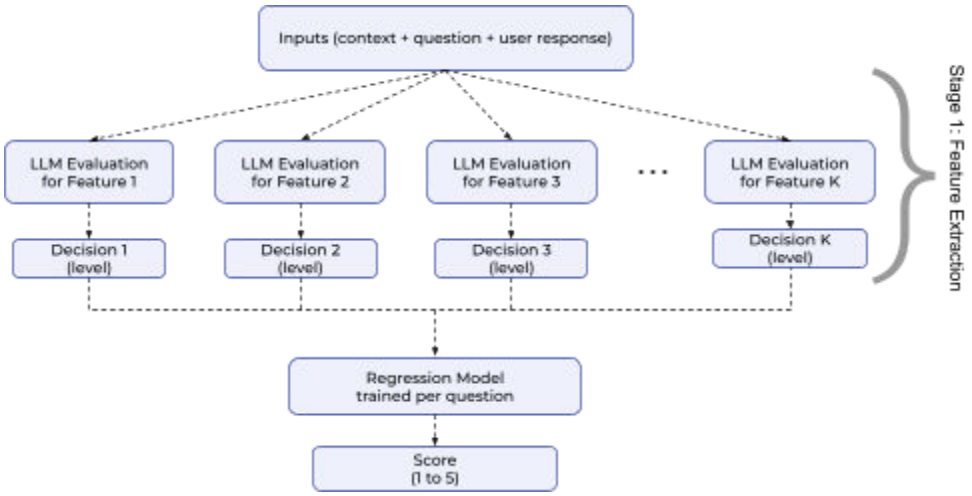
We optimized each judge using ground truth data collected with trained human raters. This ground truth data was collected in a research study outside typical PSD rating operations (approved by Veritas IRB, protocol 2025-3732-22051-3). For each assessment item, human raters evaluated 100 responses according to three to eight features that we identified as being relevant for the competency assessed by the item. We iteratively modified each LLM judge through few-shot prompting and continually evaluated the judges' performances against human raters. Across all 44 features, the judges achieved an average quadratic-weighted Kappa (QWK) agreement of 0.548 with human evaluations. For a subset of responses from each item, two independent human raters provided evaluations producing an average human-human QWK of 0.550. These results indicate that our LLMs-as-Judges agree with human evaluations to the same degree that independent human evaluators agree with each other.

Regression

The second component of our system uses the *feature classification decisions* output from the LLMs-as-Judges as inputs to a regression model (see *Figure 3.4*). We used the extracted features to train a regression model for each assessment item.

The regression model predicts a score between 1 and 5 for a given response and, to preserve as much accuracy as possible, predicted ratings may not be integers as in human rating. Using the LLMs-as-Judges constructed for the first system component, we extracted features from responses for all assessment items from 910 test takers. Separate from the feature-level data collection described above, human raters also evaluated these same responses on a 1-5 Likert scale. We used the extracted features and human ratings to train a regression model for each assessment item. We trained each regression model on a subset of responses and evaluated the models on a separate subset of responses not used for training. This training procedure ensured that our models weighted features in a way that aligned most closely with human raters when assigning numeric scores.

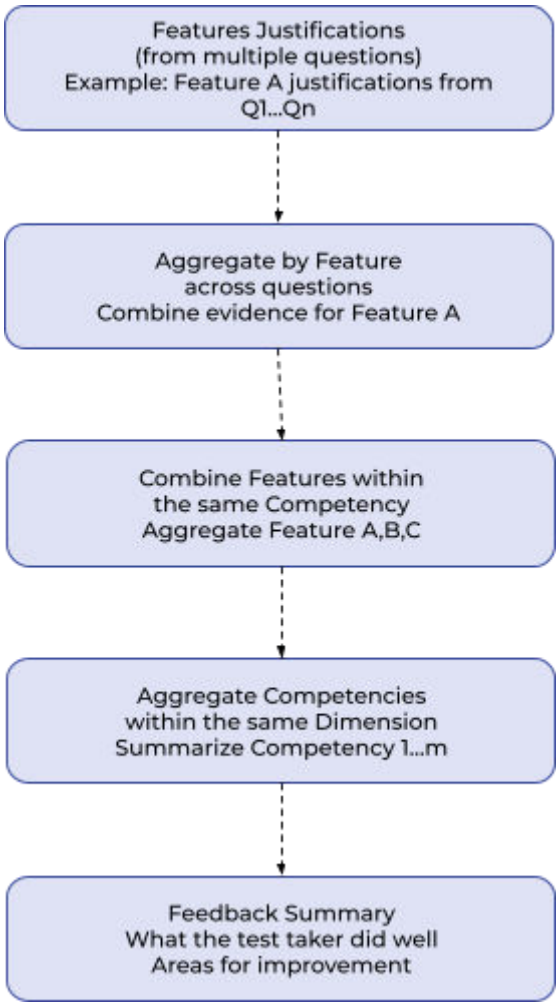
Figure 3.4. Feature classification decisions as inputs to a regression model.



Feedback Generation

The third component of our system, the feedback generation component, takes as input the *evidence-based reasoning* output from the LLMs-as-Judges in the first component and carries out a number of aggregation steps using an LLM summarization pipeline. In the first stage of the pipeline reasonings are summarized by *feature*, pulling out the higher level strengths and weaknesses exhibited across responses according to the specified feature. In the second stage, feature summaries are aggregated by *competency* to extract higher level strengths and weaknesses at the competency level. We repeat this aggregation one more time at the *dimension* level before rephrasing extracted strengths and weaknesses as constructive feedback (i.e., "What you did well" and "Areas for improvement") for the test taker at this dimension-level. This aggregation procedure is displayed in *Figure 3.5*.

Figure 3.5. LLM pipeline aggregation procedure.



We enforce the following characteristics of the feedback shown to test takers:

- Feedback is separated into sections: “What you did well” and “Areas for improvement”
- Each section includes a maximum of three pieces of feedback directly tied to one or more features
    - Each piece of feedback includes 1-2 examples (either direct quotes or paraphrased snippets) from the test taker’s responses
- Feedback is designed to be constructive and actionable.

    - Feedback is framed in terms of what the user said or did, why it was effective and how it connects to the feature(s) discussed (“What you did well” section) or what they could say or do differently in the future (“Areas for improvement” section)
- Feedback ignores any spelling, grammar, or formatting in the user's responses.
- Feedback exclusively uses the second person ("you") to address the user directly.

These characteristics were informed by feedback best practices (Shute, 2008; Mazullo & Bulut, 2025; Mazullo et al., 2025; Mazullo et al., 2026).

Feedback Generation Quality Assurance

The quality of the AI feedback system was evaluated using responses from 841 students in the United States and Canada (Bulut & Walsh, 2026). Prior to beginning the assessment, students provide consent for their data to be used for research purposes. Students’ written responses were first used to generate item-level feedback (via GPT-5 mini), which was then aggregated at the dimension level to produce four feedback summaries for each student. These dimension-level summaries incorporated direct excerpts from student responses and synthesized key strengths along with potential areas for development.

Each dimension-level feedback record was subsequently evaluated by Claude Haiku 4.5 using six criteria drawn from prior research (Mazzullo et al., 2025; Mazzullo & Bulut, 2025): (1) Linguistic quality—clarity, grammar, and punctuation; (2) Factual accuracy—the extent to which feedback accurately reflected the student’s responses and incorporated relevant quotations; (3) Personalization—whether feedback was tailored to the individual rather than generic; (4) Actionability—whether feedback included clear and useful suggestions for improvement; (5) Affective tone—whether tone was balanced and constructive, neither excessively praising nor overly critical; and (6) Second-person language—whether the student was directly addressed using “you” or “your.”

Across all evaluated feedback records, Claude Haiku 4.5 assigned consistently high quality scores regardless of evaluation method. Mean total scores (out of 6) were:

- 5.84 (SD = 0.49) for zero-shot evaluation, in which the model rated feedback directly using the criteria with no examples provided.

- 5.91 (SD = 0.30) for few-shot evaluation, in which the model first reviewed two scored examples (one high quality and one low quality) before rating new feedback.
- 5.82 (SD = 0.53) for chain-of-thought evaluation, in which the model considered each criterion step-by-step before assigning an overall score.

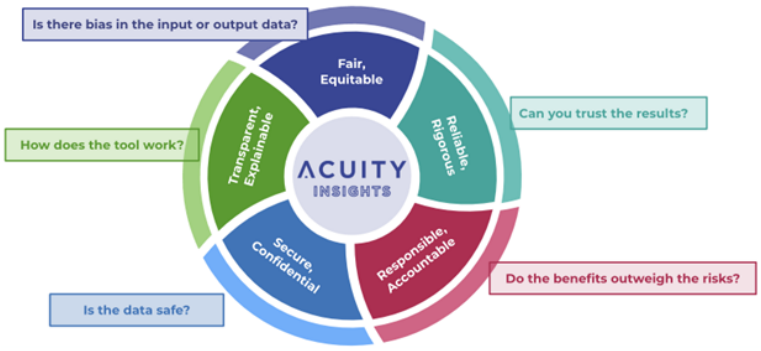
These findings indicate that AI-generated feedback was uniformly high in quality and that results were stable across evaluation methods.

AI Ethical Framework

The use of AI in educational contexts introduces important ethical considerations related to validity, reliability, transparency, fairness, and equity. In particular, algorithmic bias and limited interpretability of AI-driven decisions may reinforce existing inequities and influence assessment results (Bulut et al., 2024).

Therefore, as we began integrating artificial intelligence into our development activities, our first priority was to articulate a clear ethical framework to guide this work. This framework (see Ivan, MacIntosh, Robb, & Walsh, 2024) was intended to establish core principles that would inform and shape all subsequent development decisions. To do so, we examined guidance on responsible AI issued by a range of stakeholders, including regulatory bodies, industry organizations, and postsecondary institutions. Drawing on this review, we synthesized the guidance into five overarching ethical principles (see Figure 3.6).

Figure 3.6. Ethical AI principles and key questions.



As part of our effort to systematically capture the considerations relevant to our use of AI, we developed a set of guiding questions to support the evaluation of both existing and newly developed AI tools. These questions are intended to help determine whether the tools align with our Ethical AI principles. Below, we briefly describe each principle and provide illustrative examples of the types of questions used to support ethical evaluation.

Fair and Equitable

AI systems should be designed to identify, document, and reduce bias in both training data and outputs, while upholding accepted standards of fairness.

- Does the AI system rely on personal or sensitive data as input, and if so, is this necessary or avoidable?
- How does the institution evaluate whether the tool's outputs exhibit bias?
- In what ways does the AI system address or reduce existing biases, and what additional mitigation strategies could be implemented?

Reliable and Rigorous

AI systems should function consistently and accurately, producing dependable results even under less-than-optimal conditions.

- Given the potential for inaccurate or inconsistent outputs, how will the accuracy of results be verified?

- What procedures are in place to respond to concerns about reliability or performance?
- What assumptions and limitations are inherent in the tool, and how might real-world conditions not reflected in the training data influence outcomes?

Responsible and Accountable

AI tools should be aligned with human values and institutional goals, with clear governance structures to address unintended consequences.

- What are the intended objectives of the AI tool, and how do they align with institutional values and priorities?

- What risks or unintended effects might arise from its use, and what mitigation plans are in place?
- Who holds responsibility for interpreting, implementing, and acting on the tool's outputs?

Secure and confidential

AI systems must protect data privacy, comply with relevant data protection regulations, and maintain strong data security practices.

- Are privacy and security policies publicly available for both the AI provider and the institution?
- What data are shared with the tool's developer, what rights do they retain, and do users have the ability to opt out?
- What safeguards are in place to protect stored data, ensure regulatory compliance, and defend against cybersecurity threats?

Transparent and explainable

AI systems should offer clear, accessible information that enables users to understand how decisions are generated.

- Is documentation available that outlines, at a high level, the key inputs and factors influencing the system's outputs?
- Are users and affected parties able to explain the tool's basic functioning and purpose?
- How is feedback gathered from users, stakeholders, and others impacted by the system?

Collectively, these principles and questions serve as a living framework to guide responsible AI adoption, supporting both innovation and accountability. As AI capabilities evolve, this framework helps ensure that ethical considerations remain central to development and implementation decisions.

The development of the PSD scoring system using large language models (LLMs) was guided by the ethical principles outlined in our AI framework, with particular emphasis on fairness, reliability, and transparency. From the outset, development decisions prioritized construct alignment and the minimization of

construct-irrelevant influences on scoring. For example, model evaluation explicitly examined the impact of factors such as response length, writing quality, and irrelevant content to ensure that scores reflected underlying competencies rather than superficial features. This approach supports the principle of fair and equitable use by actively identifying and mitigating potential sources of algorithmic bias, while also reinforcing the reliability and rigor of the scoring process through systematic validation and stress testing.

In addition, the PSD system was designed to support responsible and transparent use in educational contexts. The role of the LLM is clearly defined as augmenting, rather than replacing, human judgment, with final interpretation and decision-making remaining with educators and program administrators. Documentation and validation evidence are provided to enable users to understand the general basis of scoring, while acknowledging the inherent limitations of complex AI systems. Data handling practices and system design also reflect a commitment to security and confidentiality, ensuring that learner responses are protected and used only for their intended educational purposes. Collectively, these practices ensure that the integration of AI into PSD remains aligned with institutional values, supports defensible and development-focused use, and maintains accountability to both learners and stakeholders.

Descriptive Statistics by Program Type

This section summarizes the distributional characteristics of PSD dimension scores overall and by program type. The purpose is to provide a clear picture of the range, central tendency, and shape of scores across the four PSD dimensions. Accompanying frequency histograms for each dimension (Figures 3.7–3.10) provide a visual representation of score distributions.
Table 3.1 presents descriptive statistics for the Interpersonal dimension. Overall, the mean score was 2.69, with a median of 2.56 and standard deviation of 0.79, indicating moderate variability in student scores. The distribution was slightly positively skewed (0.39) and somewhat platykurtic (-0.82), suggesting a modest tail on the higher end of the scale and a slightly flatter distribution than a normal curve.

Table 3.1. **Interpersonal Skills** dimension descriptive statistics overall and by program type.

PROGRAM	N	Mean	Median	SD	Skew	Kurtosis
OVERALL	878	2.69	2.56	0.79	0.39	-0.82
Business Education	550	2.41	2.35	0.59	0.38	-0.41
Engineering	163	2.52	2.46	0.67	0.41	-0.54
Health Sciences	117	3.81	3.91	0.42	-1.31	3.21
Medical Education	48	3.67	3.87	0.62	-1.03	0.43

As shown in Table 3.2, the Intrapersonal dimension had an overall mean of 2.77 and median of 2.64, with SD = 0.78. The overall distribution exhibited slight positive skew (0.40) and negative kurtosis (-0.85), indicating modest asymmetry and a slightly flattened distribution.

Table 3.2. **Intrapersonal Skills** dimension descriptive statistics overall and by program type.

PROGRAM	N	Mean	Median	SD	Skew	Kurtosis
OVERALL	862	2.77	2.64	0.78	0.40	-0.85
Business Education	539	2.46	2.36	0.58	0.46	-0.30
Engineering	161	2.71	2.70	0.63	0.40	-0.39
Health Sciences	116	3.90	3.95	0.40	-1.39	3.24
Medical Education	46	3.75	3.92	0.59	-1.47	2.14

Table 3.3 shows that the Social & Ethical Responsibility dimension had an overall mean of 2.79, median of 2.56, and SD of 0.79. Positive skew (0.53) and slightly negative kurtosis (-0.80) characterize the overall distribution, suggesting more students scored below the mean with a flatter-than-normal distribution.

Table 3.3. **Social & Ethical Responsibility** dimension descriptive statistics overall and by program type.

PROGRAM	N	Mean	Median	SD	Skew	Kurtosis
OVERALL	868	2.79	2.56	0.79	0.53	-0.80
Business Education	541	2.49	2.36	0.61	0.84	0.26
Engineering	165	2.68	2.56	0.61	0.46	-0.31
Health Sciences	116	3.93	3.98	0.42	-0.87	1.48
Medical Education	46	3.83	4.02	0.59	-0.93	-0.08

The Critical Thinking dimension (Table 3.4) had an overall mean of 2.68, median 2.55, and SD of 0.78, with positive skew (0.52) and negative kurtosis (-0.62), reflecting a slightly asymmetric and flatter distribution.

Table 3.4. **Critical Thinking** dimension descriptive statistics overall and by program type.

PROGRAM	N	Mean	Median	SD	Skew	Kurtosis
OVERALL	874	2.68	2.55	0.78	0.52	-0.62
Business Education	549	2.41	2.32	0.61	0.69	0.23
Engineering	162	2.59	2.47	0.69	0.62	-0.37
Health Sciences	116	3.72	3.78	0.51	-0.45	0.46
Medical Education	47	3.52	3.68	0.73	-0.64	-0.57

Across all four PSD dimensions, the descriptive statistics indicate overall moderate variability in scores across the student population. Higher scores were observed for some program types over others. Slight deviations from normality (skew and kurtosis), particularly in programs with higher mean scores, suggesting clustering of learners at the upper end of the scales. The accompanying frequency histograms (Figures 3.7–3.10) provide a visual representation of these distributions, illustrating differences across programs and highlighting the range of scores observed in the learner population.

Figure 3.7. Overall Interpersonal Score

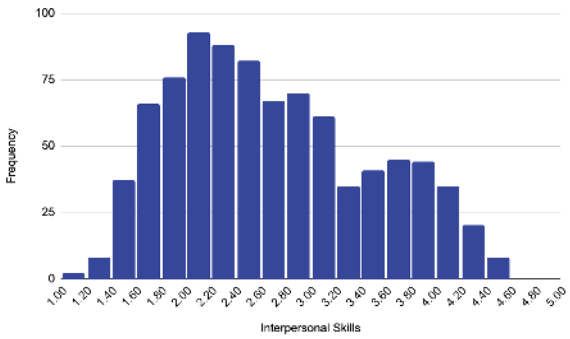


Figure 3.8. Overall Intrapersonal Score

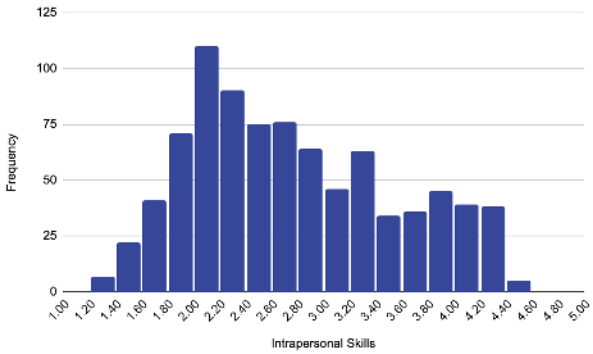


Figure 3.9. Overall Social/Ethical Score

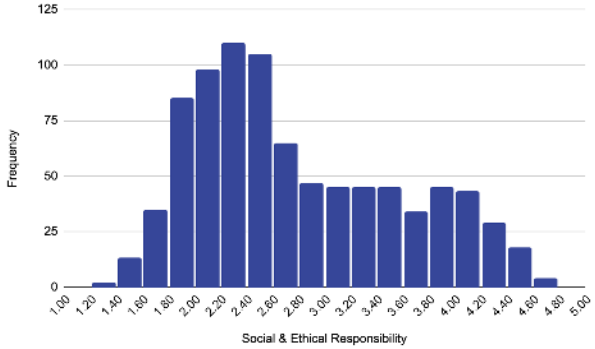
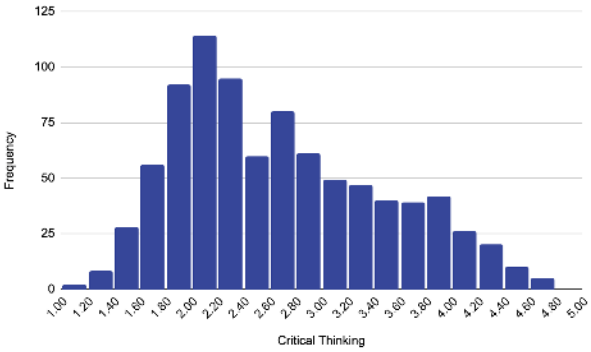


Figure 3.10. Overall Critical Thinking Score



CHAPTER 4: RELIABILITY

This chapter describes research evidence supporting the reliability of the current version of the Professional Skills Development tool.

Generally speaking, reliability refers to the consistency of a measure. A test demonstrates reliability to the extent that it produces similar scores across sources of potential variance (*e.g.,* different editions of the test, different raters, etc.). Reliability is a foundational component of any test (American Educational Research Association [AERA] et al., 2014). There are several methods for which the reliability of a test can be measured, each employed to assess a different source of potential variance, and each subsequently providing a unique piece of information (AERA et al., 2014; Cortina, 1993).

Reliability Summary

The Professional Skills Development tool has been evaluated extensively and has yielded strong supportive evidence for the reliability across several different varied samples. Each facet of reliability is outlined briefly below with more comprehensive discussions available in the following subsections.

Internal Consistency Reliability. Across all samples, the Professional Skills Development tool has consistently demonstrated strong levels of internal consistency reliability with overall alpha values ranging from α = 0.85 to α = 0.91 across dimensions, and omega values ranging from Ω = .86 to Ω = .92 across dimensions.

Standard Error of Measurement. The Standard Error of Measurement (SEM) for the PSD dimension scores ranged from .24 to .30 scale points. Corresponding 95% confidence intervals ranged from approximately ± .48 to ± .60 points around an observed score. These values indicate a relatively high level of measurement precision, and also provide guidance for practical use of the scaling.

Parallel-Forms Reliability. Alternate PSD forms demonstrated strong and comparable internal consistency (α = .85–.91; ω = .86–.93). Mean differences across forms were small (.04–.13 scale points), with nearly identical standard deviations (.77–.85). Skew and kurtosis values were similar across versions, indicating stable distributional characteristics. Overall, findings support the equivalence and interchangeability of alternate PSD forms.

Inter-Rater Reliability. AI scoring of the PSD demonstrates strong agreement with trained human raters. Item-level QWK values were predominantly moderate to very good (.35–.74), and AI–Human MAE values (.653–.790) were comparable to, and in some cases lower than, Human–Human MAE (.717–.780). Overall, AI scoring performs at a level consistent with expert human judgment.

Internal Consistency Reliability

Evidence to show that PSD scenarios and questions measure the same construct.

The internal consistency of a test refers to the extent that the items consistently measure the same construct. This type of reliability can be estimated using coefficient alpha (also frequently referred to as Cronbach's alpha), which provides an estimate of the mean correlation of all split-half reliabilities (Cortina, 1993). Coefficient alpha ranges from α=0.00 to α=1.00 with larger values indicative of greater levels of reliability. Omega (McDonald, 1999) is generally considered to be a superior metric particularly in cases where the fundamental assumptions of Cronbach's alpha are not met (Sitarenios, 2022; Deng & Chan, 2017) which is likely the case with these data. For instance, questions for dimensions occur with common modules, scenarios and storylines which means they have some shared error variance. Reliability values of .70 and above are often cited as the minimum threshold for an adequate level of internal-consistency (Cortina, 1993). The internal reliability of this Professional Skills Development tool was tested for three different samples.

Reliability values for each scale are shown Overall (Table 4.1), by program type (Table 4.2), and by gender (4.3). The results are based on data from 993 across a variety of different institutions and program types (e.g., Business Education, Engineering, Health Sciences, and Medical Education). Demographic information was optional for students, but from the responses provided were indicative of a demographically diverse sample (Male: 60.3%, Female: 38.7%, Other: 1.0%; White/European: 54.8%, Hispanic, Latinx, or Spanish origin: 14.9%, Black, African, Caribbean, African American: 9.9%, Asian: 14.6%, Other: 5.8%).

Table 4.1. PSD Overall Reliability (N = 993).

	Alpha	Omega	Alpha CI	Med R
Critical Thinking	0.85	0.86	.84 - .87	0.49
Interpersonal	0.90	0.92	.89 - .91	0.53
Intrapersonal	0.89	0.91	.88 - .90	0.50
Social and Ethical	0.91	0.92	.90 - .91	0.53

Table 4.2. PSD Reliability by Program Type.

Group	N	Alpha				Omega			
		CT	Inter	Intra	SE	CT	Inter	Intra	SE
Business Education	576	0.77	0.82	0.80	0.84	0.79	0.87	0.84	0.88
Engineering	248	0.80	0.86	0.83	0.81	0.85	0.89	0.86	0.86
Health Sciences	117	0.64	0.75	0.63	0.70	0.76	0.81	0.71	0.76
Medical Education	52	0.79	0.84	0.82	0.83	0.89	0.89	0.87	0.90

*CT=Critical Thinking; Inter=Interpersonal; Inta=Intrapersonal; SE=Social/Ethical

Table 4.3. PSD Reliability by Gender.

Group	N	Alpha				Omega			
		CT	Inter	Intra	SE	CT	Inter	Intra	SE
Female	212	0.82	0.90	0.87	0.89	0.87	0.92	0.90	0.91
Male	291	0.82	0.86	0.86	0.86	0.89	0.88	0.88	0.89

*CT=Critical Thinking; Inter=Interpersonal; Inta=Intrapersonal; SE=Social/Ethical

For all dimensions, reliability results were good to excellent indicating overall strong reliability across dimensions and that results remain stable regardless of program type or gender.

Standard Error of Measurement / Confidence Intervals

Evidence that applicants' PSD scores are very similar to their true scores.

The standard error of measurement (SEM) is an estimate of the amount of error in the obtained scores. An applicant's true score can never genuinely be known as every test contains some level of error, however, observed scores from tests can be evaluated to determine how much they likely differ from the true score. SEMs are a measure of reliability in that lower SEMs suggest that test scores are more precise (*i.e.*, less dispersed around the true score). Alternatively, higher SEMs indicate that the observed scores vary widely around the true score and are a less precise measure. Test scores for an assessment will fall within +/- one standard error of the individual's true score 68% of the time, and within +/- two standard errors 95% of the time. As can be seen in Table 4.4, PSD dimensions have relatively small SEMs, and associated confidence intervals. These values indicate a relatively high level of measurement precision. In practical terms, a student's observed dimension score is expected to fall within roughly half a scale point of their true score 95% of the time. The narrow confidence bands suggest that PSD dimension scores are stable enough to support individual-level feedback while also being sufficiently precise for cohort-level comparisons and longitudinal tracking.

Table 4.4. PSD Standard Error of Measurement, and Confidence Intervals (N = 993).

	SEM	95% Confidence Interval
Critical Thinking	0.30	+/- 0.60
Interpersonal	0.25	+/- 0.50
Intrapersonal	0.26	+/- 0.52
Social & Ethical Responsibility	0.24	+/- 0.48

Parallel Forms Reliability

Evidence that PSD scores remain consistent across alternate test versions.

Parallel forms reliability was evaluated by comparing internal consistency and score distributions across two alternate PSD versions (herein labeled Form A and Form B) using the same participant sample (N = 993). The structure of the two forms was

identical but each form has its own scenarios and questions. Internal consistency estimates (Cronbach's alpha and McDonald's omega) are presented in Table 4.9, and distributional characteristics are shown in Table 4.10.

Alternate Test Form Internal Consistency Comparability

Reliability coefficients were strong and highly comparable across forms. Cronbach's alpha ranged from .85 to .91 for Test 1 and from .86 to .91 for Test 2. McDonald's omega ranged from .86 to .92 for Test 1 and from .88 to .93 for Test 2.

Across all four dimensions, differences in reliability between forms were minimal (≤ .03 in most cases). No systematic decrease in reliability was observed in the alternate form. These results indicate that both versions of the PSD demonstrate equivalent internal consistency and measurement precision.

Table 4.9. Reliability Values for Two Different Versions of the PSD test.

	Alpha		Omega	
Dimension	**Test 1**	**Test 2**	**Test 1**	**Test 2**
Critical Thinking	0.85	0.86	0.86	0.90
Interpersonal	0.90	0.91	0.92	0.93
Intrapersonal	0.89	0.86	0.91	0.88
Social & Ethical Responsibility	0.91	0.90	0.92	0.91

Distributional Comparability

Score distributions were similarly stable across forms. Mean differences between Test 1 and Test 2 were small, ranging from .04 to .13 scale points across dimensions. Standard deviations were nearly identical (generally between .77 and .85), indicating comparable score dispersion.

Quartile values showed consistent interquartile ranges across forms, suggesting similar differentiation among learners. Skewness values were modest and positive in both versions (approximately .39 to .73), indicating slight clustering toward higher scores. Kurtosis values were consistently negative (approximately −.23 to −.85),

reflecting relatively flat distributions without excessive peakedness. These patterns were stable across dimensions and test versions. Note that if the same students take the test on separate occasions, scores generally do improve on the second test sitting due to student development and learning.

Table 4.10. Distributional Characteristics for Two Different Versions of the PSD test.

Dimension	Version	Mean	Median	Q1	Q3	SD	Skew	Kurtosis
Critical Thinking	Test 1	2.68	2.55	2.07	3.25	0.78	0.52	-0.62
	Test 2	2.64	2.48	2.02	3.02	0.79	0.73	-0.36
Interpersonal	Test 1	2.69	2.56	2.05	3.23	0.79	0.39	-0.82
	Test 2	2.63	2.41	2.06	3.10	0.83	0.67	-0.49
Intrapersonal	Test 1	2.77	2.64	2.14	3.35	0.78	0.40	-0.85
	Test 2	2.69	2.55	2.08	3.12	0.77	0.70	-0.23
Social & Ethical Responsibility	Test 1	2.79	2.56	2.16	3.40	0.79	0.53	-0.80
	Test 2	2.66	2.47	1.97	3.31	0.85	0.61	-0.66

Taken together, the consistency in reliability coefficients and distributional characteristics provides strong evidence of parallel forms reliability. The two PSD versions produce comparable score levels, variability, and internal consistency, supporting their interchangeable use for retesting, longitudinal tracking, and program evaluation.

Inter-Rater Reliability

PSD Inter-rater Reliability with Human Expert Judgment

To evaluate the performance of the AI scoring system, we benchmarked its agreement with human expert ratings against a second, important reference point: the level of agreement observed between human raters themselves. Including both human–AI and human–human comparisons allows for a more meaningful interpretation of model performance. Specifically, human–human agreement provides an upper-bound estimate of expected variability in scoring, reflecting the degree of subjectivity present even among trained experts. Human–AI agreement can then be interpreted relative to this benchmark to determine whether AI performance falls within the range of normal human scoring variation.

Inter-rater agreement between AI scoring and human expert scoring was examined using (i) Quadratic Weighted Kappa (QWK)—a measure of agreement between two

raters for ordinal scores, where disagreements are weighted more heavily the farther apart the ratings are (e.g., a 1 vs. 5 disagreement counts more than 3 vs. 4); and (ii) Mean Absolute Error (MAE), which quantifies the average absolute difference in scores.

All item-level models were trained individually using an 80/20 train-test split: for each item, responses (with associated human ratings) were randomly partitioned such that 80% were used for training with five-fold cross-validation. The remaining 20% of responses formed a held-out test set. Model predictions on this unseen subset were compared directly to human ratings, and results are reported on this held-out data to reflect expected generalization performance.

Quadratic Weighted Kappa (QWK)

Table 4.11 shows the results for QWK. QWK values above .80 are indicative of near perfect agreement, values between .61 to .80 are considered as indicating very good agreement, values from .40 to .60 indicate moderate agreement and values less than .40 indicate fair to poor agreement (Landis & Koch, 1977). QWK is reported at the question level, which represents a stringent test of agreement. When scores are aggregated across questions to form dimension-level scores, discrepancies tend to average out, typically resulting in stronger overall correspondence.

Results indicate moderate to very good agreement between AI scoring and human expert judgment across most questions.

Table 4.11. Automated Scoring x Human Inter-rater Agreement

	N Pairs	QWK
Interpersonal Dimension		
Collaboration Q1	200	0.560
Collaboration Q2	196	0.599
Communication Q1	200	0.629
Communication Q2	200	0.441
Empathy Q1	200	0.661
Empathy Q2	202	0.521
Leadership Q1	198	0.498
Leadership Q2	198	0.704
Critical Thinking	201	0.517
Critical Thinking Q1	203	0.464
Critical Thinking Q2	199	0.741
Critical Thinking Q3	196	0.597
Critical Thinking Q4	196	0.615
Critical Thinking Q5	200	0.349
Social and Ethical Responsibility		
Ethics Q1	199	0.541
Ethics Q2	197	0.539
Ethics Q3	196	0.561
Respect for Others Q1	199	0.637
Respect for Others Q2	197	0.511
Respect for Others Q3	195	0.589
Societal Welfare Q1	198	0.600
Societal Welfare Q2	193	0.591
Intrapersonal Dimension		
Lifelong Learning Q1	202	0.604
Lifelong Learning Q2	189	0.352
Motivation Q1	200	0.591
Motivation Q2	195	0.716
Resilience Q1	197	0.518
Resilience Q2	195	0.518
Self-awareness Q1	191	0.575
Self-awareness Q2	200	0.522

Mean Absolute Error

Table 4.12 presents MAE values comparing: Human-to-Human scoring, and AI-to-Human scoring. Although there are no universal cutoffs for interpreting MAE values, they are meaningfully interpreted relative to scale and comparative benchmarks. Here, AI–Human MAE values are directly compared with Human–Human MAE values.

Across all four PSD dimensions, AI–Human MAE values are highly comparable to Human–Human MAE values. In some cases (e.g., Critical Thinking), AI–Human MAE is lower than Human–Human MAE. These findings indicate that AI scoring performs at a level comparable to trained human raters with respect to average scoring deviation.

Table 4.12. MAE values comparing AI and Human Scoring.

Dimension	Human-Human	AI-Human
Critical Thinking	0.750	0.653
Interpersonal	0.734	0.733
Intrapersonal	0.717	0.786
Social & Ethical Responsibility	0.780	0.790

Across correlational analyses, QWK agreement statistics, and MAE comparisons, AI-generated PSD scores demonstrate strong correspondence with trained human expert judgment. Correlations are high across dimensions, inter-rater agreement is predominantly moderate to very good at the item level, and AI scoring deviation is comparable to human-to-human scoring differences.

Taken together, these results provide strong evidence that the AI scoring system produces evaluations consistent with expert human judgment.

Reliability Conclusions

As evidenced by the aforementioned results, it is clear that the Professional Skills Development tool is a reliable measure of personal and professional skills. Consistently high and uniform levels of internal-consistency provide evidence that all scenarios and questions of the test work together to measure the constructs represented in the test. The low SEM values indicate that scores derived from this

Professional Skills Development tool are reflective of their true scores. Results of parallel-forms analysis provide evidence that test scores remain consistent across time periods and across different variations of the test. In sum, the consistency of the scores across different analyses provide support for the Professional Skills Development tool as a reliable measure of personal and professional skills.

CHAPTER 5: VALIDITY

Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing tests and evaluating tests. The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores for proposed uses that are evaluated, not the test itself (AERA, APA, & NCME, 2014).

The intended purposes of the PSD test were outlined at the outset of this manual (see *Intended Purposes and Use* section). These purposes included the ability to

(i) **provide quantifiable measurement** for competency areas related to personal and professional skills (specifically interpersonal skills, intrapersonal skills, social & ethical responsibility, and critical thinking).
(ii) **meaningfully differentiate between performance levels** for individuals at different skill levels.
(iii) **effectively capture changes in development**.

This chapter, therefore, includes sections that provide supporting evidence for each of these stated purposes.

Validity Summary

The sections that follow present multiple sources of validity evidence aligned with the intended uses of the Professional Skills Development (PSD) assessment. Each source of evidence is summarized briefly below, with more detailed analyses provided in the sections that follow.

Provide Quantifiable Measurement (Construct Validity)
Internal Structure (Confirmatory Factor Analysis). Confirmatory factor analysis supported the hypothesized four-dimension structure of the PSD assessment. Model fit indices indicated excellent fit (CFI = .99, TLI = .99, RMSEA = .02). All competency indicators loaded significantly onto their intended dimensions (loadings = .63–.88), and the four dimensions were positively correlated but empirically distinguishable.

Relations with External Measures (Casper Scores). PSD scores showed small-to-moderate correlations with Casper situational judgment test scores (r =

.24–.37). These relationships are consistent with expectations given that both assessments measure aspects of social and professional competence while differing in structure and administration.

AI Scoring Validity. AI-generated PSD scores showed strong correspondence with human expert ratings (r = .82–.88). Additional analyses showed systematic declines in scores when responses were evaluated outside their intended scenario–question context, indicating that the scoring system is sensitive to construct alignment.

Meaningfully Differentiate Between Performance Levels (Discriminative Validity)
PSD scores differentiated between groups expected to vary in professional competencies. Fourth-year engineering learners scored higher than first-year learners across dimensions, and learners in a selective enhanced program consistently scored higher than those in a standard program. Observed differences were generally in the small-to-moderate range.

Effectively Capture Changes in Development (Sensitivity to Change).
PSD scores demonstrated sensitivity to developmental change over time. In one program, learners showed consistent improvements across dimensions following a short instructional interval. In a second program, changes varied by dimension but included substantial gains in intrapersonal competencies.

PSD Performance Across Groups (Generalizability Across Groups).
Exploratory subgroup analyses examined score patterns across several demographic characteristics. Most differences were small, particularly for accommodation status, community background, language, and socioeconomic indicators. Moderate differences were observed in some gender and race/ethnicity comparisons, though patterns varied across dimensions. Overall, results suggest that PSD scores function comparably across many demographic groups.

Evidence that PSD provides quantifiable measurement of the targeted constructs (Construct Validity)

Evidence of construct validity refers to the degree to which an assessment meaningfully and accurately measures the theoretical constructs it is intended to assess. In line with contemporary validity frameworks, construct validity is supported through multiple sources of evidence, including the internal structure of the measure, relationships with external variables, and the degree to which observed scores reflect the intended latent constructs (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; Messick, 1995). In this section, we present evidence that the PSD assessment provides quantifiable and interpretable measurement of the targeted competency domains.

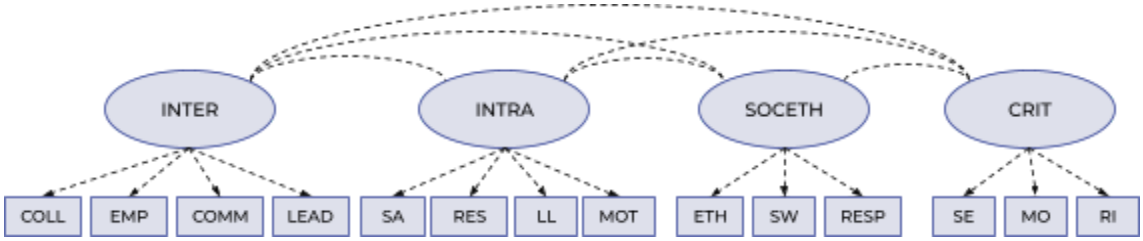
Confirmatory Factor Analysis: Evidence Supporting the Validity of the PSD Structure

Structural validity refers to the degree to which the internal structure of an assessment aligns with the theoretical construct it is intended to measure (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Factor analytic methods are commonly used to evaluate this aspect of validity. Exploratory factor analysis (EFA) is typically used in early stages of test development to identify potential underlying factor structures without imposing strong a priori assumptions. In contrast, confirmatory factor analysis (CFA) is used to test hypothesized models by specifying relationships between observed indicators and latent constructs. In CFA procedures, the primary objective is to evaluate the extent to which the proposed model adequately fits the observed data; poor fit may indicate the need for model refinement or reconsideration of the underlying structure. Interpretation of model fit followed commonly used guidelines that emphasize evaluation across multiple indices rather than reliance on a single cutoff, with acceptable fit often indicated by CFI and TLI values ≥ .90 and RMSEA ≤ .10 (Bentler, 1990; Byrne, 2006; Browne & Cudeck, 1993).

To test the dimensional structure of the PSD test, questions from the same competency were combined to create indicators for the four Dimensions. In this CFA model, the competency subcomponents of each domain are taken into account. That is, items from the same competency (within a domain) are treated as a common unit in the model. Then, these competency level indicators are mapped onto their respective domains. Furthermore, since all of the dimensions are commonly related to personal and professional skills, the model specifications

included intercorrelations between the dimensions. Figure 5.1 displays the model that was tested.

Figure 5.1. Diagrammatic Representation of CFA model testing PSD structure.



Note: INTER = Interpersonal Skills; INTRA = Intrapersonal Skills; SOCETH = Social & Ethical Responsibility; CRIT = Critical Thinking; COLL = Collaboration; EMP = Empathy; COMM = Communication; LEAD = Leadership; SA = Self Awareness; RES = Resiliency; LL = Lifelong Learning; MOT = Motivation; ETH = Ethics; SW = Societal Welfare; RESP = Respect for Others; SE = Critical Evaluation & Solution Enactment; MO = Management & Organization Skills; RI = Research & Inquiry.

CFA Results

Model Fit

Overall model fit statistics are presented in Table 5.1.

Table 5.1. Goodness of fit Indicators.

Fit Indicator	Value	Interpretation
CFI	0.99	Excellent Fit
TLI	0.99	Excellent Fit
RMSEA	0.02	Excellent Fit

Confirmatory factor analysis supported the hypothesized four-dimension structure of the PSD assessment. Model fit indices indicated excellent fit (CFI = .99, TLI = .99, RMSEA = .02). Across all indices, results exceeded the cited benchmarks, providing strong support for the proposed measurement model.

Factor Loadings

Standardized parameter estimates for the CFA model are summarized in Table 5.2. In line with common guidelines for interpreting standardized factor loadings (e.g., ≥ .50 = moderate, ≥ .70 = strong; see e.g., Hair et al., 2019), all component indicators loaded

significantly onto their respective latent dimensions ($p < .001$), indicating that each indicator contributes meaningfully to the measurement of its intended construct. Within the Critical Thinking dimension, factor loadings ranged from 0.630 to 0.786, suggesting moderate to strong relationships between the latent construct and its component indicators (Solution Enactment, Management/Organization, and Research Inquiry). For the Interpersonal dimension, loadings ranged from 0.644 to 0.798 across Collaboration, Empathy, Communication, and Leadership, indicating consistently moderate to strong associations with the underlying construct. Within the Intrapersonal dimension, factor loadings ranged from 0.669 to 0.878, with Resilience demonstrating the strongest association with the latent construct, while Motivation, Self-Awareness, and Lifelong Learning also showed strong and statistically significant relationships. For the Social and Ethical Responsibility dimension, factor loadings ranged from 0.705 to 0.775 across Ethics, Societal Welfare, and Respect for Others, indicating consistently strong contributions from each component to the overall construct.

Table 5.2. CFA Parameter Estimate Summary.

Parameter Estimates	Estimate	SE	Significance
Critical Thinking Dimension			
Solution Enactment	0.630	0.024	$p < .001$
Management / Organization	0.786	0.026	$p < .001$
Research Inquiry	0.746	0.025	$p < .001$
Interpersonal Dimension			
Collaboration	0.798	0.026	$p < .001$
Empathy	0.644	0.024	$p < .001$
Communication	0.787	0.025	$p < .001$
Leadership	0.763	0.025	$p < .001$
Intrapersonal Dimension			
Motivation	0.699	0.023	$p < .001$
Self-awareness	0.690	0.025	$p < .001$
Resilience	0.878	0.028	$p < .001$
Lifelong Learning	0.669	0.022	$p < .001$
Social & Ethical Responsibility			
Ethics	0.705	0.022	$p < .001$
Societal Welfare	0.775	0.025	$p < .001$
Respect for Others	0.775	0.024	$p < .001$

Overall, the magnitude and consistency of the factor loadings support the interpretation that the proposed component indicators function effectively as measures of their respective dimensions.

Dimension Relationships

All correlations between latent dimensions were large (> .95) positive and statistically significant ($p < .001$), indicating that the dimensions are strongly related to one another. This pattern is consistent with the conceptualization of PSD competencies as interconnected aspects of professional and personal development.

At the same time, the CFA model retains the dimensions as distinct latent constructs, allowing each domain to capture a specific aspect of learner development while acknowledging their shared underlying foundations.

Overall, the CFA results provide strong empirical support for the PSD measurement model. The excellent model fit indices, significant and substantial factor loadings, and theoretically consistent relationships among the dimensions indicate that the four-dimension structure offers a valid representation of the PSD framework and its component indicators.

Evidence Based on Relations with Other Variables: Casper Scores

Evidence based on relations with other variables was examined by evaluating the relationship between scores on the PSD and scores from the Casper Test. Casper is a situational judgment test used in admissions processes to assess applicants' social intelligence and professionalism. Unlike the PSD, which produces separate scores across multiple competency domains, Casper yields a single overall score intended to reflect social intelligence and professionalism.

One program provided Casper scores for consenting learners in addition to having them complete the Professional Skills Development tool. This extra data point provided an opportunity to look at the relationship between scores derived from these two assessments. Given their structural similarities, it is hypothesized that there would be a relationship between the two. The 64 learners reported here were a subset of 116 students from the Pharmacy program of the University of Waterloo who took the SJT as part of the midyear assessment. The subset was restricted to students who provided consent for their SJT results and academic results to be used for research purposes. Although demographic information is not available specifically for the subset, the overall group had the following characteristics. The mean reported age for these students was 21.3 (SD = 8.0), 68% were female, and 39.2% reported as

being “White/Caucasian,” with the remaining ethnic representation being spread out across categories.

Correlations between Casper scores and the PSD dimension scores were calculated and are presented in Table 5.3. Because the sample consisted of students who had already been admitted to the program, correlations were corrected for range restriction to better approximate the relationships that would be expected in a broader applicant population.

Table 5.3. PSD Correlations with Casper.

Dimension	Correlation	Significance
Critical Thinking	.36	$p = .003$
Interpersonal	.24	$p = .056$
Intrapersonal	.37	$p = .003$
Social/Ethical Responsibility	.30	$p = .016$

As shown in Table 5.3, correlations between Casper scores and the PSD dimensions ranged from .24 to .37, indicating small-to-moderate positive relationships. The strongest relationships were observed for the Intrapersonal and Critical Thinking dimensions, followed by Social and Ethical Responsibility, while the Interpersonal dimension demonstrated a smaller but positive association with Casper scores.

The magnitude of these relationships is consistent with expectations. Although both assessments are designed to measure aspects of social and professional competence, they differ in structure, scoring, and intended use. Casper produces a single composite score derived from responses to open-ended scenarios, whereas the PSD provides separate scores across multiple competency domains. In addition, Casper was administered approximately two years prior to the PSD for learners in this sample. Differences in measurement format and the passage of time would reasonably be expected to attenuate the strength of the observed relationships.

Overall, the observed pattern of correlations is consistent with theoretical expectations and provides supporting evidence for the interpretation of PSD scores as indicators of social and professional competencies.

Sensitivity of AI Scores to Scenario–Question Alignment

This section documents how AI-generated scores vary as a function of alignment between a response and its intended scoring context. Specifically, we examined the extent to which AI scores decline when responses are evaluated outside their original scenario, question, competency, or dimension (Walsh & Ivan, 2026). These analyses provide evidence that the AI scoring system is context-sensitive, producing the highest scores when responses are evaluated in their intended context and progressively lower scores as contextual mismatch increases. The AI scoring model is trained to evaluate responses relative to a specific scenario and question, which jointly define the intended competency and performance dimension being assessed. When a response is scored against an alternative context, the semantic and construct alignment between the response and the scoring prompt decreases. If the AI system is functioning as intended, this misalignment should result in systematic score attenuation, rather than random or inflated scores.

Table 5.4 presents mean AI scores under four increasingly misaligned scoring conditions. Each condition represents a deliberate deviation from the response's original scenario–question pairing.

Table 5.4. Mean AI Scores by Scenario–Question Alignment Condition.

Scoring Condition	Description	Mean Score
Correct match	Response scored against its original scenario and question	**3.11**
Same competency	Different question, same competency	**2.38**
Same dimension	Different question, different competency, same dimension	**1.95**
Different dimension	Different question, different dimension	**1.64**

These results demonstrate a clear monotonic decline in mean scores as the degree of contextual mismatch increases.

To further characterize this pattern, Table 5.5 summarizes the incremental score decreases associated with each additional level of mismatch relative to the previous condition.

Table 5.5. Incremental Decrease in Mean Score by Type of Mismatch.

Comparison	Mean Score Decrease
Correct match → Same competency	−0.73
Same competency → Same dimension	−0.43
Same dimension → Different dimension	−0.31

The largest decrease occurs when responses are scored against a different question, even when the competency remains the same. Additional decreases are observed when the competency and, subsequently, the dimension no longer align.

These findings support the interpretation that AI scores are meaningfully anchored to the intended assessment framework and dimensions.

Correspondence to Scores based on Human Expert Judgement

The PSD scoring system is derived from an AI-driven large language model (LLM) algorithm (see *Scoring and Feedback* section for full details). To evaluate the validity of the automated scoring approach, PSD scores were compared with human expert judgments. This analysis evaluates the extent to which AI scoring produces results that align with expert human evaluation.

PSD Score correlation with Human Expert Judgment

PSD responses were independently scored by the AI algorithm and trained human raters. Human raters were trained to evaluate responses by selecting appropriate categorizations across multiple criteria and then assigning an overall dimension score based on these categorizations.

The same sample of 993 learners described in the *Internal Consistency Reliability* section was used for this analysis.

Table 5.6. AI scores correlations with Human Expert Ratings by PSD Dimension.

Dimension	Correlation (*r*)	Significance
Critical Thinking	0.88	$p < .001$
Interpersonal	0.85	$p < .001$
Intrapersonal	0.82	$p < .001$
Social & Ethical	0.85	$p < .001$

As shown in Table 5.6. All correlations are strong (r = .82 – .88), indicating substantial correspondence between AI-generated scores and human expert ratings across all PSD dimensions.

Evidence that PSD effectively distinguishes between performance levels (Discriminative Validity)

Discriminative validity refers to the extent to which an assessment is able to meaningfully differentiate between individuals who are expected to vary in their level of the targeted constructs (American Educational Research Association et al., 2014). In this section, we examine the extent to which the PSD assessment distinguishes between individuals with expected differing levels of competency across the targeted domains.

First year v Fourth year students

143 students from an Engineering program in the midwest region of the U.S. were included in this analysis: 99 were first year students, and 44 were 4th year students. This analysis inspects differential performance of first and fourth year students. The students were predominantly White/European (91.5%), and Male (78.5%).

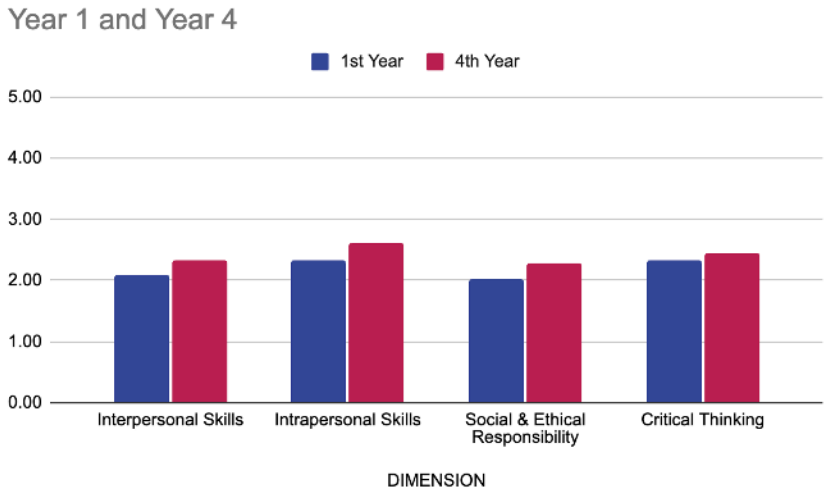
Fourth year students had higher scores across dimensions with the largest differences being found in the Intrapersonal and Social & Ethical Responsibility dimensions, as shown in Table 5.7 and Figure 5.1. The effect sizes were generally in the small to moderate range. This finding is consistent with expectations that 4th year students improve in their personal and professional skills as they mature and as they learn about the principles associated with these skills from what is conveyed

through the program. This supports PSD validity since the test appropriately captures the superior performance expected from the 4th year students (Cohen, 1992).

Table 5.7. By Year comparison – Mean (SD).

Domain	First Year Students (N = 99)	Fourth Year Students (N = 44)	Effect Size (Cohen's d)	Effect Size Guideline	Sig
Critical Thinking	2.32 (.74)	2.46 (.68)	.21	Small	*p* = .287
Interpersonal	2.10 (.91)	2.33 (.95)	.14	Negligible -Small	*p* = .171
Intrapersonal	2.32* (.81)	2.61* (.84)	.33	Small-Moderate	*p* = .052
Social & Ethical	2.03* (.83)	2.28* (.72)	.47	Moderate	*p* = .086

Figure 5.1. By Year comparison.



Known-Groups Validity

Another source of discriminative validity evidence for the PSD tool comes from its ability to differentiate between groups that are expected, a priori, to differ in the constructs being measured. To examine this, a known-groups design was implemented comparing standard introductory business students with students enrolled in a selective living–learning community.

Using data from a U.S. midwestern business school, two groups of undergraduate students enrolled in the same *Introductory Business* course completed the PSD assessment at two time points approximately two months apart. Standard Introductory Business students representing a typical first-year business cohort – we refer to these students as the "Standard Program" group. The other group was previously selected for a special program based on their enhanced competencies. We refer to these students as the "Enhanced Program" group since this special program also provides structured skill development, personalized feedback, and targeted instruction. Students were tested at two time points: Test 1 involved 72 Standard Program students, and 28 Enhanced Program students. Test 2 involved 66 Standard Program students, and 24 Enhanced Program students. The sample sizes dropped slightly with fewer of the students completing the second test sitting.

Based on program design and selection criteria, the "Enhanced Program" students were expected to demonstrate stronger performance across the PSD dimensions.

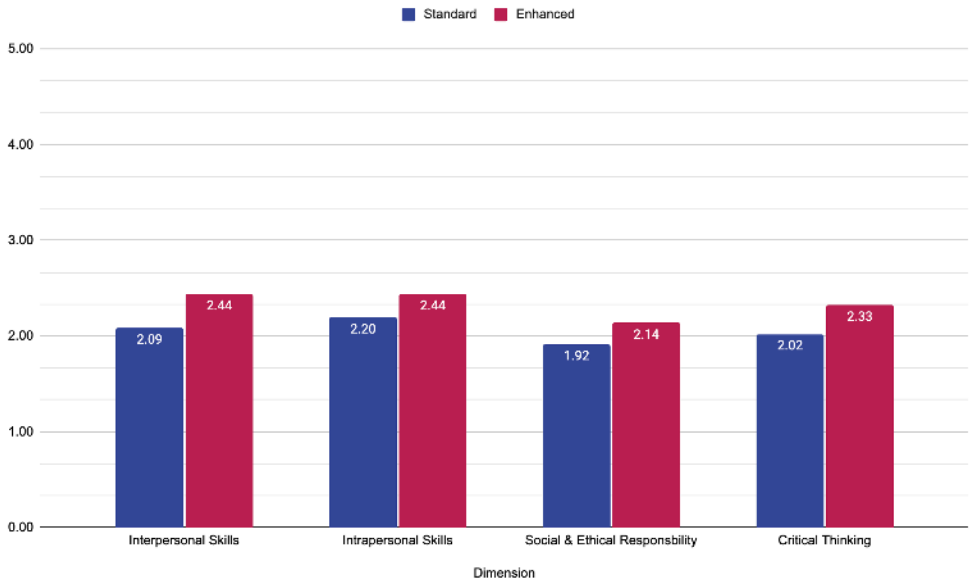
Group Comparisons at Test 1

Table 5.7 presents mean scores, standard deviations, effect sizes (Cohen's *d*), and significance tests for the four PSD dimensions at Test 1. The findings are also displayed in Figure 5.2.

Table 5.7. Group Differences by PSD dimension (Test 1).

	Standard Program	Enhanced Program		
Dimension	**MEAN (SD)**	**MEAN (SD)**	**Cohen's D**	**p**
Interpersonal Skills	2.09 (.63)	2.44 (.77)	0.49	0.044
Intrapersonal Skills	2.20 (.56)	2.44 (.69)	0.38	0.108
Social & Ethical Responsibility	1.92 (.61)	2.14 (.69)	0.34	0.147
Critical Thinking	2.02 (.52)	2.33 (.61)	0.53	0.022

FIgure 5.2. Group Differences by PSD dimension (Test 1).



At Test 1, students in the "Enhanced Program" scored higher than standard program students across all four dimensions. Statistically significant differences were observed for Interpersonal Skills and Critical Thinking, with effect sizes in the small-to-moderate range (Cohen, 1992). Differences for Intrapersonal Skills and Social & Ethical Responsibility were in the expected direction but did not reach conventional levels of statistical significance, likely reflecting modest sample sizes.

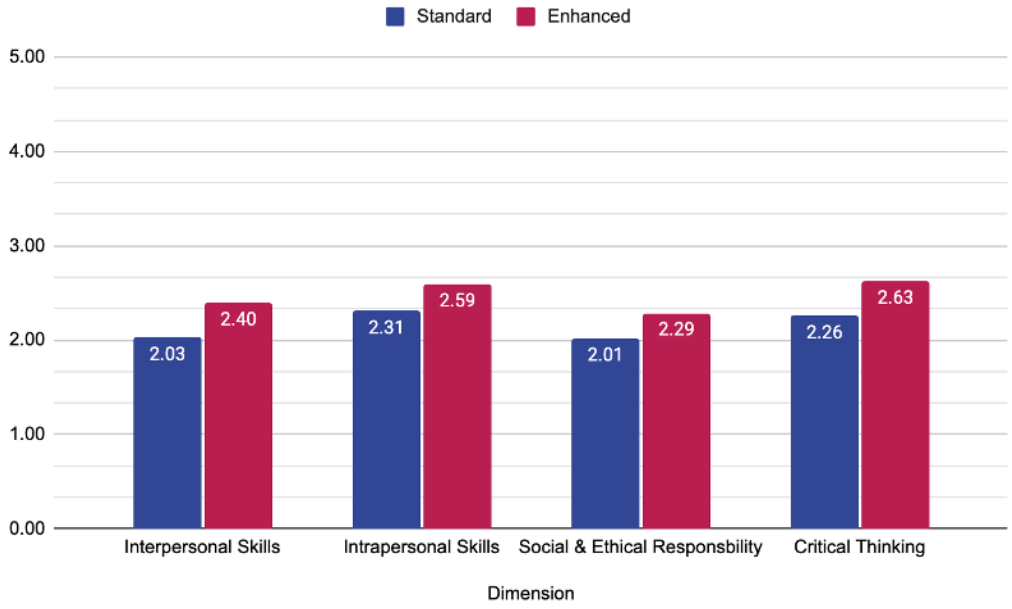
Group Comparisons at Test 2

Table 5.8 and Figure 5.3 show the corresponding results at Test 2.

Table 5.8. Group Differences by PSD dimension (Test 2).

	Standard Program	Enhanced Program		
Dimension	**MEAN (SD)**	**MEAN (SD)**	**Cohen's D**	**Sig**
Interpersonal Skills	2.03 (.67)	2.40 (.73)	0.52	*p* = .036
Intrapersonal Skills	2.31 (.79)	2.59 (.68)	0.38	*p* = .104
Social & Ethical Responsibility	2.01 (.71)	2.29 (.68)	0.39	*p* = .094
Critical Thinking	2.26 (.75)	2.63 (.79)	0.48	*p* = .053

FIgure 5.3. Group Differences by PSD dimension (Test 2).



Results at Test 2 mirrored the Test 1 findings. The students in the Enhanced Program again demonstrated higher mean scores across all PSD dimensions. Differences remained statistically significant for Interpersonal Skills, with Critical Thinking approaching significance. Effect sizes were consistent with those observed at Test 1, indicating stable group separation over time.

Across two administrations, the PSD tool consistently differentiated between groups. The observed differences were directionally consistent, and stable across time. These findings provide supporting evidence for known-groups validity, demonstrating that PSD scores meaningfully reflect expected differences in students' personal,

interpersonal, and professional skill development. Importantly, the results also suggest that the PSD tool is sensitive enough to detect performance differences associated with structured developmental programming and enhanced feedback, reinforcing its intended use as a formative assessment instrument.

Evidence that PSD is appropriately sensitive to developmental change (Sensitivity to Change)

One central application of the PSD assessment is to track changes in learners' social and professional competencies as they progress through their educational programs. Accordingly, an important element of validity is demonstrating that PSD scores are appropriately sensitive to developmental change. If the assessment is functioning as intended, scores should reflect improvements in competencies when learners receive instruction, practice, or feedback designed to develop these skills.

Data from two early adopter programs were used to examine this question. In both cases, the PSD tool was administered twice to the same group of students, with a relatively short interval of approximately two months between administrations. During this interval, students continued their normal coursework and were provided with resources and feedback following the first administration designed to support the development of PSD competencies.

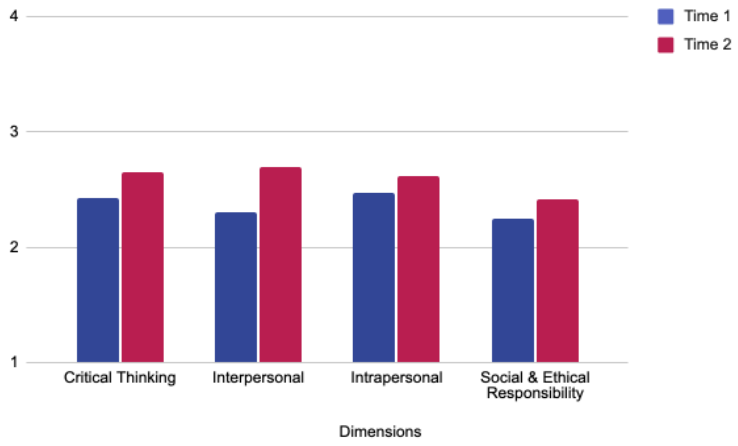
Because the interval between administrations was relatively brief, only modest improvements were expected. Nevertheless, if the PSD tool is sensitive to developmental change, scores should show detectable shifts over time in at least some of the competency domains.

In the first program (here referred to as Program A), a total of 62 students completed the PSD assessment at both time points. Results are presented in Table 5.9 and Figure 5.4.

Table 5.9. PSD results Time 1 v Time 2 (Program A, N = 62).

Variable	T1_Mean (SD)	T2_Mean (SD)	Cohen's D	Sig
Critical Thinking	2.42 (0.73)	2.66 (0.75)	.36	*p* = .006
Interpersonal	2.30 (0.95)	2.69 (1.11)	.44	*p* < .001
Intrapersonal	2.47 (0.82)	2.61 (0.88)	.22	*p* = .084
Social & Ethical Responsibility	2.25 (0.80)	2.41 (0.74)	.26	*p* = .047

Figure 5.4. PSD results Time 1 v Time 2 (Program A, N = 62).



Across the four PSD dimensions, mean scores increased from Time 1 to Time 2, indicating consistent improvement across domains. The largest increase was observed in the Interpersonal dimension, where the mean score increased from 2.30 to 2.69, corresponding to a moderate effect size (Cohen's d = .44) that was statistically significant (p < .001). Improvements were also observed for Critical Thinking, which increased from 2.42 to 2.66 (d = .36, p < .01), and Social & Ethical Responsibility, which increased from 2.25 to 2.41 (d = .26, p < .05). The Intrapersonal dimension also showed a positive increase (from 2.47 to 2.61), although the effect size was smaller (d = .22) and only marginally significant (p < .10).

Overall, the results from Program A demonstrate consistent directional improvement across all PSD dimensions within a relatively short developmental interval. Although the magnitude of change was modest—as expected given the brief time between

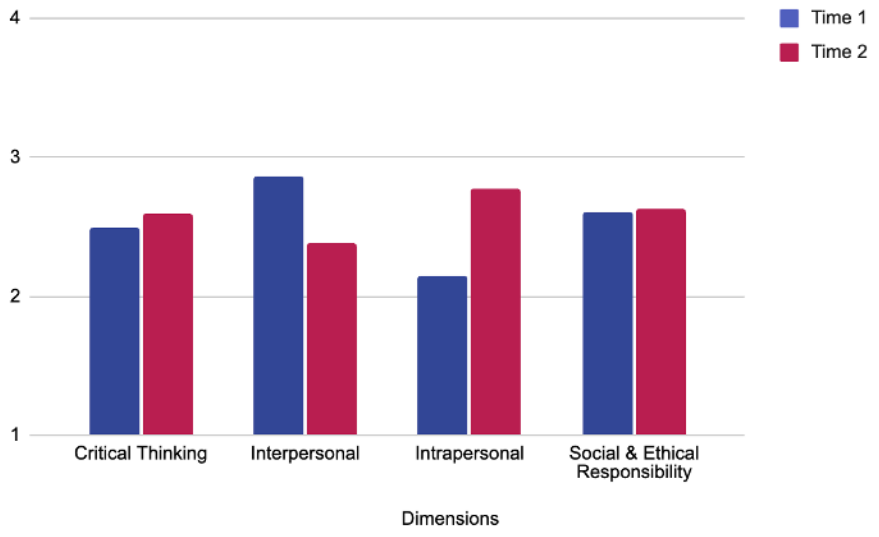
administrations—the pattern of results suggests that PSD scores are responsive to short-term developmental experiences.

In the second program (Program B), 73 students completed the PSD at both time points. Results are summarized in Table 5.10 and Figure 5.5.

Table 5.10. PSD results Time 1 v Time 2 (Program B, N = 73).

Variable	T1_Mean (SD)	T2_Mean (SD)	Cohen's D	Sig
Critical Thinking	2.50 (.87)	2.59 (.75)	.12	p = .240
Interpersonal	2.86 (.80)	2.39 (.80)	-.59	p < .001
Intrapersonal	2.15 (.72)	2.77 (.77)	.84	p < .001
Social & Ethical Responsibility	2.61 (.81)	2.63 (.78)	.02	p = .830

Figure 5.5. PSD results Time 1 v Time 2 (Program B, N = 73).



Unlike Program A, the pattern of results in Program B varied across dimensions. The Intrapersonal dimension showed a substantial increase from Time 1 to Time 2, rising from 2.15 to 2.77. This change corresponded to a large effect size (d = .84) and was statistically significant (p < .001). In contrast, the Interpersonal dimension decreased from 2.86 to 2.39 (d = –.59, p < .001). This decrease may reflect differences in

instructional emphasis, cohort characteristics, or student response patterns between administrations.

Changes in the remaining dimensions were small. Critical Thinking increased slightly from 2.50 to 2.59 (*d* = .12), while Social & Ethical Responsibility remained essentially unchanged (2.61 to 2.63; *d* = .02). Neither of these changes reached statistical significance.

Taken together, the results from Program B suggest that the PSD is capable of detecting meaningful changes when they occur, while also reflecting dimension-specific patterns of development that may differ across instructional contexts.

Across the two early adopter programs, the PSD demonstrated sensitivity to developmental change over time. In Program A, improvements were observed across all dimensions, whereas Program B showed a more differentiated pattern of change across competencies. The magnitude of observed changes was generally modest, which is consistent with the short interval between test administrations and the early stage of competency development. Overall, these findings provide support that PSD scores can capture changes in students' social and professional competencies, supporting the intended use of the assessment as a tool for monitoring developmental progress within educational programs.

Fairness: Evidence Regarding Score Comparability Across Demographic Groups (Generalizability Across Groups).

An important aspect of validity for any assessment used in educational settings is evidence that scores function comparably across demographic subgroups. Large systematic differences between groups may raise concerns about potential bias or construct-irrelevant variance. Accordingly, subgroup score patterns were examined across several demographic characteristics to evaluate whether PSD scores show substantial differences between groups.

Analyses were conducted using data from six programs that implemented the Early Adopter (EA) version of PSD with a total sample size of 943 although group sizes differed depending on the demographic characteristic being examined.

To ensure stable estimates, subgroup comparisons were limited to groups with sample sizes of at least 30 participants. For each demographic variable, mean PSD scores were computed and Cohen's d effect sizes were calculated relative to the

subgroup with the largest sample size within that demographic category. Standardized mean difference values (d) are often interpreted such that difference values of 0.20, 0.50, and 0.80 correspond to small, moderate, and large effect sizes, respectively (Cohen, 1992).

Accommodation

Table 5.11 presents subgroup comparisons based on self-reported accommodation or disability status. Students reporting ADD/ADHD, a long-term mental health condition, or another form of disability (captured in a general "Any Condition" category) were compared with students reporting no disability or chronic condition, which served as the reference group.

Table 5.11. Subgroup comparisons for accommodation demographic variable.

		Critical Thinking		**Interpersonal**		**Intrapersonal**		**Social/ Ethical**	
Group	N	Mean	D	Mean	D	Mean	D	Mean	D
No Disability or Chronic Condition*	396	2.67 (0.74)	–	2.67 (0.75)	–	2.75 (0.73)	–	2.75 (0.74)	–
I have ADD/ADHD	60	2.66 (0.75)	.00	2.70 (0.72)	-.05	2.82 (0.72)	-.09	2.85 (0.69)	-0.14
I have a long-term mental health condition	32	2.68 (0.66)	-.02	2.76 (0.71)	-.13	2.84 (0.72)	-.13	2.88 (0.66)	-0.18
Any Condition[1]	95	2.73 (0.79)	-.11	2.76 (0.75)	-.18	2.92 (0.77)	-.27	2.94 (0.75)	-0.30

*used as comparison group; [1]This category includes individuals who specified any disability status other than "No Disability or Chronic Condition."

Across all PSD dimensions, differences between groups were minimal and, where present, slightly favored students reporting a disability or chronic condition. Mean scores were highly comparable across groups, and associated effect sizes were consistently near zero, indicating negligible practical differences. Overall, these findings suggest that PSD scores operate similarly for students regardless of accommodation status within the current sample, providing no evidence of meaningful group-based disparities.

Community Size

This section examines potential differences in PSD scores across categories of community size. Community background is treated as a demographic variable that may reflect differences in educational context, access to resources, and lived experience. To explore whether PSD performance varies across these contexts, subgroup comparisons are conducted across five categories: large towns, medium towns, major urban centers, small towns, and rural areas. For analytical clarity, these categories are grouped into two broader clusters—students from large towns, medium towns, and major urban centers versus those from small towns and rural areas—with the former serving as the reference group. Comparisons are reported across the four PSD dimensions: Critical Thinking, Interpersonal, Intrapersonal, and Social/Ethical.

Table 5.12 presents the subgroup comparisons for the community size demographic variable.

Table 5.12. Subgroup comparisons for community size demographic variable.

		Critical Thinking		**Interpersonal**		**Intrapersonal**		**Social/ Ethical**	
Group	N	Mean	D	Mean	D	Mean	D	Mean	D
Large town, Medium Town, Major urban center*	347	2.64 (0.75)	–	2.63 (0.76)	–	2.73 (0.75)	–	2.75 (0.75)	–
Small town, Rural	156	2.65 (0.72)	.01	2.65 (0.72)	.03	2.76 (0.70)	.05	2.74 (0.70)	-0.01

*used as comparison group

Across PSD dimensions, differences between community size groups were consistently negligible. Mean scores were highly similar across categories, and the associated effect sizes were uniformly close to zero, indicating minimal practical differences. These findings suggest that PSD scores are not meaningfully associated with community size, and that students from rural and urban backgrounds perform comparably across all four domains within the current sample.

Gender

This section examines potential differences in PSD scores across gender. Gender is treated as a demographic variable that may be associated with differences in communication styles, educational experiences, and socialization patterns that could, in turn, relate to performance on PSD dimensions.

Table 5.13. Subgroup comparisons for the gender demographic variable.

		Critical Thinking		**Interpersonal**		**Intrapersonal**		**Social/ Ethical**	
Group	N	Mean (SD)	D	Mean (SD)	D	Mean (SD)	D	Mean (SD)	D
Man*	320	2.43 (0.71)	–	2.45 (0.68)	–	2.33 (0.67)	–	2.44 (0.66)	–
Woman	260	2.86 (0.74)	0.51	2.84 (0.77)	0.49	2.76 (0.75)	0.56	2.99 (0.77)	0.64

*used as comparison group

Across all PSD dimensions, women obtained higher mean scores than men. The observed differences were consistent across domains and were accompanied by effect sizes in the moderate range, indicating meaningful group differences in performance. These findings suggest that PSD scores vary by gender within the current sample, with women demonstrating comparatively higher scores across all four dimensions.

Language

This section examines potential differences in PSD scores across language backgrounds. Language proficiency is considered a demographic variable that may reflect differences in educational exposure, communication patterns, and familiarity with the language of assessment, all of which could influence performance on PSD dimensions. To explore these potential differences, subgroup comparisons are conducted between students identifying as fluent English speakers and those identifying as non-fluent English speakers

Table 5.14 presents comparisons between students reporting Native and non-Native English language.

Table 5.14. Subgroup comparisons for language demographic variable.

		Critical Thinking		Interpersonal		Intrapersonal		Social/ Ethical	
Group	N	Mean	D	Mean	D	Mean	D	Mean	D
Fluent English Speaker*	381	2.72 (0.75)	–	2.71 (0.75)	–	2.82 (0.73)	–	2.83 (0.74)	–
Non-fluent English Speaker	91	2.54 (0.74)	-0.24	2.50 (0.73)	-0.28	2.57 (0.71)	-0.35	2.57 (0.72)	-0.35

*used as comparison group

Across all PSD dimensions, fluent English speakers obtained slightly higher mean scores than non-fluent English speakers. The observed differences were small in magnitude, with effect sizes indicating modest but consistent advantages for fluent English speakers across the four domains. These findings suggest that language background is associated with PSD performance, although the differences are relatively minor and should have limited practical impact on PSD scores.

Parental Income

This section examines potential differences in PSD scores across parental income. Parental income is used as a proxy indicator of socioeconomic background, which may be associated with differences in educational opportunities, access to resources, and prior academic experiences that could influence performance on PSD dimensions. To assess these potential differences, subgroup comparisons are conducted between students reporting parental income of $100,000 or more and those reporting less than $100,000.

Table 5.15. Subgroup comparisons for the parental income demographic variable.

		Critical Thinking		Interpersonal		Intrapersonal		Social/ Ethical	
Group	N	Mean	D	Mean	D	Mean	D	Mean	D
100K or more*	183	2.65 (0.72)	–	2.64 (0.78)	–	2.76 (0.74)	–	2.78 (0.75)	–
Less than 100K	182	2.65 (0.75)	-0.01	2.63 (0.69)	0.02	2.72 (0.69)	0.06	2.72 (0.69)	0.08

*used as comparison group

Across all PSD dimensions, mean scores were highly similar between the parental income groups. Differences were minimal in magnitude, and effect sizes were consistently negligible, indicating no meaningful variation in performance across income categories. Overall, these findings suggest that PSD scores are broadly comparable across students from different socioeconomic backgrounds within the current sample, with no evidence of practically significant differences associated with parental income.

Race / Ethnicity

This section examines potential differences in PSD scores across race and ethnicity. Race and ethnicity are treated as demographic variables that may reflect differences in lived experiences, cultural context, and educational opportunities, which could in turn relate to performance on PSD dimensions. To explore these potential differences, subgroup comparisons are conducted across four categories: White or European, Hispanic or Latinx, Black, African, Caribbean, or African American, and South Asian or East Asian students (including Bangladeshi, Bhutanese, Indian, Nepali, Pakistani, Sri Lankan, and East Asian backgrounds).

Table 5.16 presents the subgroup comparisons for the race/ethnicity demographic variable.

Table 5.16. Subgroup comparisons for race/ethnicity demographic variable.

		Critical Thinking		**Interpersonal**		**Intrapersonal**		**Social/ Ethical**	
Group	N	Mean	D	Mean	D	Mean	D	Mean	D
White or European*	283	2.64 (0.73)	–	2.64 (0.74)	–	2.73 (0.73)	–	2.76 (0.72)	–
Hispanic or Latinx	84	2.44 (0.59)	0.29	2.44 (0.59)	0.28	2.57 (0.54)	0.25	2.56 (0.55)	0.29
Black, African, Caribbean, or African American	51	2.33 (0.56)	0.43	2.31 (0.59)	0.47	2.44 (0.55)	0.42	2.42 (0.49)	0.44
South Asian or East Asian	82	3.08 (0.81)	-0.60	3.01 (0.78)	-0..50	3.13 (0.79)	-0.54	3.13 (0.86)	-0.50

*used as comparison group

Across PSD dimensions, differences between racial and ethnic groups were observed, with effect sizes ranging from small to moderate in magnitude. Compared

to the reference group, Hispanic or Latinx students and Black, African, Caribbean, or African American students tended to have lower mean scores across all dimensions. In contrast, students identifying as South Asian or East Asian demonstrated higher mean scores relative to the reference group, with this pattern being consistent across PSD domains. Overall, the findings indicate meaningful variation in PSD scores across race and ethnicity categories within the current sample.

Regression Analysis of Group Comparisons

Building on the subgroup comparisons summarized using Cohen's d, multiple linear regression analyses were conducted to further examine whether demographic variables uniquely predicted PSD performance across competency domains when considered simultaneously. Whereas Cohen's d provided an initial indication of the magnitude of pairwise subgroup differences, the regression models estimate the unique association of each demographic variable while holding the other variables in the model constant, thereby clarifying whether a predictor contributes independently to performance after accounting for shared variance with other demographic factors.

Separate models were estimated for each domain (Critical Thinking, Interpersonal, Intrapersonal, and Social & Ethical Responsibility) using multiple imputed datasets. Multiple imputation addresses missing data by creating several complete datasets in which missing values are replaced with statistically plausible estimates based on observed response patterns; parameter estimates are then pooled across datasets to reflect uncertainty due to missingness. This approach typically yields results comparable to complete-data analyses when missingness is limited, while reducing bias and preserving sample size. Partial eta squared (ηp^2) values were computed to estimate effect sizes.

For the Critical Thinking dimension (see Table 5.17), gender was a statistically significant predictor of performance ($b = -0.46$, $p < .001$, $\eta^2 p = 0.10$), indicating a modest association consistent with the moderate subgroup differences observed in the Cohen's *d* analyses. All other demographic variables—including community size, English fluency, income proxy, and race/ethnicity indicators—were not statistically significant (all $p > .05$), and their effect sizes were negligible ($\eta_p^2 \leq 0.01$), aligning with the small subgroup differences previously observed.

Table 5.17. Regression Results: **Critical Thinking** Dimension.

Parameter	estimate	std.error	Sig	Partial Eta-Squared
(Intercept)	2.98	0.31	$p < .001$	–
Gender	-0.46	0.09	$p < .001$	0.10
Community Size	-0.08	0.16	$p = .645$	0.01
English Fluency	-0.06	0.30	$p = .830$	<0.01
Parental Income	-0.01	0.10	$p = .897$	<0.01
Race/Ethnicity 1*	-0.29	1.50	$p = .847$	<0.01
Race/Ethnicity 2*	-0.43	0.70	$p = .536$	<0.01
Race/Ethnicity 3*	0.63	0.70	$p = .371$	<0.01

*1 = Black, African, Caribbean, or African American; 2 = Hispanic or Latinx; 3 = South Asian or East Asian

In the Interpersonal dimension (see Table 5.18), gender again emerged as a statistically significant predictor (b = -0.51, p < .001, $\eta_p^2 = 0.11$), reinforcing the presence of meaningful subgroup differences identified in Cohen's *d* analyses. No other demographic variables were statistically significant, and all effect sizes were small ($\eta_p^2 \leq 0.02$), consistent with the overall pattern of minimal subgroup differences across most demographic categories.

Table 5.18. Regression Results: **Interpersonal Skills** Dimension.

Parameter	estimate	std.error	Sig	Partial Eta-Squared
(Intercept)	3.25	0.34	p < .001	–
Gender	-0.51	0.10	p < .001	0.11
Community Size	-0.16	0.16	$p = .352$	0.02
English Fluency	-0.30	0.35	$p = .411$	<0.01
Parental Income	0.09	0.10	$p = .357$	<0.01
Race/Ethnicity 1*	0.17	1.50	$p = .911$	<0.01
Race/Ethnicity 2*	-0.29	0.70	$p = .679$	<0.01
Race/Ethnicity 3*	0.19	0.70	$p = .792$	<0.01

*1 = Black, African, Caribbean, or African American; 2 = Hispanic or Latinx; 3 = South Asian or East Asian

For the Intrapersonal dimension (see Table 5.19), gender was again a significant predictor (b = -0.57, $p < .001$, $\eta_p^2 = 0.14$), representing the strongest observed effect across the analyses. This finding is consistent with the moderate subgroup differences identified in Cohen's *d* results. All other demographic variables were not statistically significant (all $p > .05$), and their effect sizes remained negligible ($\eta_p^2 \leq 0.02$).

Table 5.19. Regression Results: **Intrapersonal Skills** Dimension.

Parameter	estimate	std.error	Sig	Partial Eta-Squared
(Intercept)	3.14	0.40	$p < .001$	–
Gender	-0.57	0.09	$p < .001$	0.14
Community Size	-0.13	0.14	$p = .374$	0.01
English Fluency	-0.04	0.34	$p = .911$	<0.01
Parental Income	0.07	0.10	$p = .470$	<0.01
Race/Ethnicity 1*	0.65	1.46	$p = .656$	<0.01
Race/Ethnicity 2*	-0.68	0.68	$p = .318$	<0.01
Race/Ethnicity 3*	0.26	0.68	$p = .703$	<0.01

*1 = Black, African, Caribbean, or African American; 2 = Hispanic or Latinx; 3 = South Asian or East Asian

In the Social dimension (see Table 5.20), gender was also a statistically significant predictor (b = -0.50, $p = .001$, $\eta_p^2 = 0.12$). No other demographic variables reached statistical significance, and their associated effect sizes were very small ($\eta_p^2 \leq 0.01$), again consistent with the small subgroup differences observed in Cohen's *d* analyses for most demographic variables.

Table 5.20. Regression Results: **Social & Ethical Responsibility** Dimension.

Parameter	estimate	std.error	Sig	Partial Eta-Squared
(Intercept)	3.14	0.30	$p < .001$	NA
Gender	-0.50	0.11	$p = .001$	0.12
Community Size	-0.09	0.09	$p = .306$	0.01
English Fluency	-0.06	0.31	$p = .858$	<0.01
Parental Income	0.04	0.10	$p = .702$	<0.01
Race/Ethnicity 1*	1.34	1.46	$p = .360$	<0.01
Race/Ethnicity 2*	-0.70	0.68	$p = .300$	<0.01
Race/Ethnicity 3*	-0.17	0.68	$p = .803$	<0.01

*1 = Black, African, Caribbean, or African American; 2 = Hispanic or Latinx; 3 = South Asian or East Asian

Taken together, the regression results are broadly consistent with the subgroup comparisons reported using Cohen's *d*. Across all competency domains, gender emerged as the only demographic variable with a consistent and statistically significant association with PSD performance, with small to moderate effect sizes observed across domains. In contrast, community size, English fluency, income proxy, and race/ethnicity variables did not demonstrate statistically significant unique contributions to performance when considered simultaneously, and their effect sizes were uniformly small.

Overall, the combined evidence from both Cohen's *d* and regression analyses suggests that PSD scores demonstrate reasonable subgroup comparability across most demographic variables, with limited and dimension-specific differences primarily associated with gender. These findings support the interpretation that PSD measures the intended competencies in a broadly consistent manner across diverse student populations.

Differential Item Functioning (DIF) by Group

Differential item functioning (DIF) analyses were conducted across all four PSD dimensions—Critical Thinking, Interpersonal Skills, Intrapersonal Skills, and Social & Ethical Responsibility—to evaluate whether items functioned differently across demographic groups. DIF analyses examine whether individuals from different groups, but with comparable standing on the underlying construct being measured, have different probabilities of receiving the same item score. As such, DIF provides an important indicator of item fairness and measurement equivalence across subgroups.

Logistic regression-based DIF methods were used, examining both uniform (ΔR^2-U) and non-uniform DIF (ΔR^2-NU) effects. Uniform DIF refers to a consistent difference between groups across all levels of the underlying ability. In this case, one group systematically performs better or worse on an item regardless of their overall standing on the construct being measured. Non-uniform DIF occurs when the magnitude or direction of group differences varies across levels of the underlying ability. This indicates an interaction between group membership and ability, such that the item may favor different groups at different points along the ability continuum (Camilli et al., 1994; Zumbo, 1999).

The results are interpreted in terms of commonly accepted benchmarks for interpreting DIF effect sizes ($\Delta R^2 < .02$ = negligible; .02–.13 = small; > .13 = moderate to large) These thresholds are consistent with established guidelines for evaluating the practical significance of DIF effects (e.g., see Zumbo, 1999).

For the Critical Thinking dimension (see Table 5.21), results indicated that all detected DIF effects were negligible in magnitude across items and demographic groups. Although some statistically detectable differences were observed, all ΔR^2 values fell well below the threshold for small effects ($\Delta R^2 < .02$), indicating minimal practical impact. This pattern was consistent across all six items in this dimension and all demographic variables examined, including gender, community size, language background, parental income, and race/ethnicity. No items met criteria for small or

moderate DIF, and all items were classified as negligible (T1) for both uniform and non-uniform DIF. Overall, these findings provide strong evidence that the Critical Thinking items function equivalently across demographic groups, supporting the fairness and measurement invariance of this dimension within the current sample.

Table 5.21. DIF Results: **Critical Thinking** Dimension.

	Gender		**Community Size**		**English Fluency**		**Parental Income**		**Race / Ethnicity 1***		**Race / Ethnicity 2***		**Race / Ethnicity 3***	
	ΔR^2-U	ΔR^2-NU	ΔR^2-U	ΔR^2-NU	ΔR^2-U	ΔR^2-NU	ΔR^2-U	ΔR^2-NU	ΔR^2-U	ΔR^2-NU	ΔR^2-U	ΔR^2-NU	ΔR^2-U	ΔR^2-NU
Q1	<.001	<.001	.005	<.001	.003	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001
Q2	<.001	<.001	.006	<.001	.001	<.001	<.001	<.001	<.001	.002	.003	<.001	.006	.001
Q3	.001	<.001	<.001	<.001	<.001	<.001	.001	<.001	.001	.001	.002	<.001	.002	<.001
Q4	<.001	.004	<.001	<.001	<.001	<.001	<.001	.001	.006	<.001	.004	.001	<.001	.002
Q5	.001	.004	<.001	<.001	.001	.001	.001	.001	<.001	<.001	.001	<.001	.001	<.001
Q6	<.001	<.001	<.001	<.001	.001	<.001	.001	<.001	<.001	.001	.002	.002	.003	<.001
T1**	6/6		6/6		6/6		6/6		6/6		6/6		6/6	
T2**	0/6		0/6		0/6		0/6		0/6		0/6		0/6	
T3**	0/6		0/6		0/6		0/6		0/6		0/6		0/6	

*1 = Black, African, Caribbean, or African American; 2 = Hispanic or Latinx; 3 = South Asian or East Asian
**DIF categorization T1 = Negligible (<.02), T2 = Small (.02 - .13) ; 3 = Medium/Large (>.13).

For the Interpersonal Skills dimension (see Table 5.22), results indicated that all DIF effects were negligible in magnitude across items and demographic groups. While some variation in ΔR^2 values was observed, all values remained below the threshold for small effects ($\Delta R^2 < .02$), indicating no meaningful practical impact. This pattern was consistent across all eight items in this dimension and all demographic variables examined, including gender, community size, language background, parental income, and race/ethnicity. All items were classified as negligible (T1) for both uniform and non-uniform DIF, with no items meeting criteria for small or moderate DIF. Overall, these findings provide strong evidence that the Interpersonal Skills items function equivalently across demographic groups, supporting the fairness and measurement invariance of this dimension within the current sample.

Table 5.22. DIF Results: **Interpersonal Skills** Dimension.

	Gender		Community Size		English Fluency		Parental Income		Race / Ethnicity 1*		Race / Ethnicity 2*		Race / Ethnicity 3*	
	ΔR^2-U	ΔR^2-NU	ΔR^2-U	ΔR^2-NU	ΔR^2-U	ΔR^2-NU	ΔR^2-U	ΔR^2-NU	ΔR^2-U	ΔR^2-NU	ΔR^2-U	ΔR^2-NU	ΔR^2-U	ΔR^2-NU
Q1	<.001	<.001	<.001	<.001	.002	.002	.004	<.001	<.001	<.001	.001	<.001	<.001	.002
Q2	.013	<.001	<.001	<.001	.001	<.001	.001	<.001	<.001	.002	.004	.004	.002	.003
Q3	<.001	<.001	.001	.001	.002	.001	.006	.001	.002	<.001	<.001	<.001	<.001	<.001
Q4	.001	.004	<.001	.001	<.001	<.001	.001	<.001	<.001	.001	<.001	<.001	<.001	<.001
Q5	<.001	.004	<.001	<.001	.001	<.001	.006	<.001	.002	<.001	.003	.001	<.001	<.001
Q6	.005	<.001	<.001	.003	.001	.003	.002	<.001	<.001	<.001	.002	<.001	<.001	.002
Q7	.001	.004	.002	<.001	<.001	<.001	.001	<.001	.003	.001	.001	<.001	<.001	<.001
Q8	<.001	<.001	<.001	.001	.001	<.001	<.001	.004	<.001	<.001	<.001	.001	<.001	<.001
T1**	8/8		8/8		8/8		8/8		8/8		8/8		8/8	
T2**	0/8		0/8		0/8		0/8		0/8		0/8		0/8	
T3**	0/8		0/8		0/8		0/8		0/8		0/8		0/8	

*1 = Black, African, Caribbean, or African American; 2 = Hispanic or Latinx; 3 = South Asian or East Asian
**DIF categorization T1 = Negligible (<.02), T2 = Small (.02 - .13) ; 3 = Medium/Large (>.13).

For the Intrapersonal Skills dimension (see Table 5.23), all observed DIF effects were negligible in magnitude across items and demographic groups. Although small variations in ΔR^2 values were present, all values remained well below the threshold for small effects ($\Delta R^2 < .02$), indicating no meaningful differences. This pattern was consistent across all eight items of this dimension and all demographic variables examined, including gender, community size, language background, parental income, and race/ethnicity. All items were classified as negligible (T1) for both uniform and non-uniform DIF, with no items meeting criteria for small or moderate DIF. Overall, these findings indicate that the Intrapersonal Skills items function equivalently across demographic groups, further supporting the fairness and measurement invariance of this dimension within the current sample.

Table 5.23. DIF Results: **Intrapersonal Skills** Dimension.

	Gender		Community Size		English Fluency		Parental Income		Race / Ethnicity 1*		Race / Ethnicity 2*		Race / Ethnicity 3*	
	ΔR^2-U	ΔR^2-NU	ΔR^2-U	ΔR^2-NU	ΔR^2-U	ΔR^2-NU	ΔR^2-U	ΔR^2-NU	ΔR^2-U	ΔR^2-NU	ΔR^2-U	ΔR^2-NU	ΔR^2-U	ΔR^2-NU
Q1	<.001	<.001	<.001	.001	.001	<.001	<.001	.001	<.001	<.001	.005	.006	.001	<.001
Q2	.002	<.001	<.001	<.001	<.001	<.001	.001	<.001	.001	.001	.001	<.001	<.001	<.001
Q3	.001	.001	<.001	.001	<.001	<.001	.001	.002	<.001	<.001	.002	.001	<.001	<.001
Q4	.002	.001	<.001	.001	<.001	.001	<.001	.001	<.001	.007	<.001	.001	.003	<.001
Q5	<.001	.004	.001	<.001	.001	.001	.001	.003	<.001	<.001	.002	<.001	.002	<.001
Q6	<.001	.001	.003	<.001	<.001	.001	.001	.001	<.001	.004	.001	<.001	.002	<.001
Q7	<.001	<.001	.004	.002	<.001	.001	.008	.001	<.001	<.001	.003	<.001	.003	<.001
Q8	.003	<.001	<.001	<.001	.001	<.001	<.001	<.001	<.001	<.001	.003	<.001	.001	.001
T1**	8/8		8/8		8/8		8/8		8/8		8/8		8/8	
T2**	0/8		0/8		0/8		0/8		0/8		0/8		0/8	
T3**	0/8		0/8		0/8		0/8		0/8		0/8		0/8	

*1 = Black, African, Caribbean, or African American; 2 = Hispanic or Latinx; 3 = South Asian or East Asian
**DIF categorization T1 = Negligible (<.02), T2 = Small (.02 - .13) ; 3 = Medium/Large (>.13).

For the Social & Ethical Responsibility dimension (see Table 5.24), all observed DIF effects were negligible in magnitude across items and demographic groups. While some small variations in ΔR^2 values were observed, all values remained below the threshold for small effects ($\Delta R^2 < .02$), indicating no practically meaningful differences. This pattern was consistent across all eight items in this dimension and all demographic variables examined, including gender, community size, language background, parental income, and race/ethnicity. All items were classified as negligible (T1) for both uniform and non-uniform DIF, with no items meeting criteria for small or moderate DIF. Overall, these findings indicate that the Social & Ethical Responsibility items function equivalently across demographic groups, providing further evidence of measurement fairness and invariance for this dimension within the current sample.

Table 5.23. DIF Results: **Social & Ethical Responsibility** Dimension.

	Gender		Community Size		English Fluency		Parental Income		Race / Ethnicity 1*		Race / Ethnicity 2*		Race / Ethnicity 3*	
	ΔR^2-U	ΔR^2-NU	ΔR^2-U	ΔR^2-NU	ΔR^2-U	ΔR^2-NU	ΔR^2-U	ΔR^2-NU	ΔR^2-U	ΔR^2-NU	ΔR^2-U	ΔR^2-NU	ΔR^2-U	ΔR^2-NU
Q1	.001	.001	<.001	<.001	<.001	<.001	.004	.001	<.001	<.001	.003	.001	.002	.001
Q2	.001	.002	.001	<.001	.004	<.001	.001	<.001	<.001	<.001	.002	.001	.009	.001
Q3	.004	.001	<.001	<.001	.002	<.001	<.001	<.001	<.001	.003	.004	<.001	<.001	.002
Q4	<.001	<.001	<.001	<.001	.001	.001	<.001	.003	<.001	<.001	.008	.008	<.001	<.001
Q5	.002	<.001	<.001	<.001	.004	<.001	.002	<.001	.006	.012	.001	.002	.001	.002
Q6	.002	<.001	.002	<.001	.001	.001	.001	<.001	.010	<.001	.003	.002	.002	.003
Q7	.003	.001	<.001	<.001	<.001	.001	.004	<.001	.002	.001	.001	.002	.002	.001
Q8	.003	.001	<.001	.001	.002	<.001	<.001	.001	<.001	.002	.001	.001	<.001	.002
T1**	8/8		8/8		8/8		8/8		8/8		8/8		8/8	
T2**	0/8		0/8		0/8		0/8		0/8		0/8		0/8	
T3**	0/8		0/8		0/8		0/8		0/8		0/8		0/8	

*1 = Black, African, Caribbean, or African American; 2 = Hispanic or Latinx; 3 = South Asian or East Asian
**DIF categorization T1 = Negligible (<.02), T2 = Small (.02 - .13) ; 3 = Medium/Large (>.13).

Across all four PSD dimensions, results were highly consistent. Although some statistically detectable differences were observed, all DIF effects were negligible in magnitude and fell below the threshold for small effects ($\Delta R^2 < .02$). No items met criteria for small or moderate DIF for either uniform or non-uniform effects across any demographic grouping examined, including gender, community size, language background, parental income, and race/ethnicity.

Overall, these findings provide strong evidence that PSD items function equivalently across demographic groups, supporting the fairness, comparability, and measurement invariance of scores within the current sample.

References

AACSB International. (2020). *AACSB Accelerators: Seven Competencies of a Societal Impact Leader*. Tampa, FL: AACSB International.

Acuity Insights. (2026). *Casper technical manual*. https://acuityinsights.com/wp-content/uploads/2024/07/Casper-Technical-Manual.pdf

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Alt, D., Naamati-Schneider, L., & Weishut, D. J. N. (2023). Competency-based learning and formative assessment feedback as precursors of college students' soft skills acquisition. *Studies in Higher Education*, *48*(12), 1901–1917. https://doi.org/10.1080/03075079.2023.2217203

Antes, A. L., Dineen, K. K., Bakanas, E., Zahrli, T., Keune, J. D., Schuelke, M. J., & DuBois, J. M. (2020). Professional decision-making in medicine: Development of a new measure and preliminary evidence of validity. *PLoS ONE*, *15*(2), e0228450. https://doi.org/10.1371/journal.pone.0228450 pubmed.ncbi.nlm.nih.gov+1nottingham-repository.worktribe.com+1

Archbell, K., Bynkoski, K., Richard, C., Sitarenios, G., & Ivan, R. (2025). *Implementing a formative Situational Judgement Test to support the development of professionalism among pharmacy students* (April 30, 2025; University of Waterloo Teaching and Learning Conference, Waterloo, ON, CAN)

Ballejos, M. P., Cestone, C. Copeland, H. L., Dunleavy, D. M., Geiger, T., & Patel, D. M. S. (2024). Predicting Medical Student Performance with a Situational Judgment Test in Admissions. *Academic Medicine*, *99*(2),175-182. https://doi.org/10.1097/ACM.0000000000005516

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Browne, M. W., & Cudeck, R. (1993). *Alternative ways of assessing model fit*. In K. A. Bollen & J. S. Long (Eds.), Testing structural equation models (pp. 136–162). Sage.

Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., Ormerod, C., Fabiyi, D. G., Ivan, R., Walsh, C., Rios, O., Wilson, J., Yildirim-Erbasli, S. N., Liu, X., Wongvorachan, T., Liu, J. X., Tan, B., & Morilova, P. (2024). Applications of artificial intelligence in educational measurement: Opportunities and ethical challenges. *Chinese/English Journal of Educational Measurement and Evaluation*, *5*(3), Article 4. https://doi.org/10.59863/SFEB5996

Bulut, O., & Walsh, C. (2026). *Closing the feedback loop: Automated quality assurance for LLM-generated feedback*. In *Proceedings of the International Conference on Artificial Intelligence in Education (AIED 2026)*.

Bynkoski, K., Archbell, K., Richard, C., Sitarenios, G., & Ivan, R. (2025). *Implementing a formative Situational Judgement Test to support the development of professionalism among pharmacy students* (June 17, 2025; Canadian Pharmacy Education and Research Council, Niagara Falls, ON, CAN)

Byrne, B. M. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming (2nd ed.)*. Lawrence Erlbaum Associates.

Camilli, Gregory, & Shepard, Lorrie A. (1994). Methods for identifying biased test items. *Educational Measurement: Issues and Practice*, *13*(4), 12–19.

Cortina, J. M., (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98-104. doi: 10.1037/0021-9010.78.1.98

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cohen J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge Academic.

Cullen, M. J., Zhang, C., Marcus-Blank, B., Braman, J. P., Tiryaki, E., Konia, M., Hunt, M. A., Lee, M. S., Van Heest, A., Englander, R., & Sackett, P. R. (2020). Improving our ability to predict resident applicant performance: Validity evidence for a situational judgment test. *Teaching and Learning in Medicine*, *32*(5), 508–521. https://doi.org/10.1080/10401334.2020.1760104

Cullen, M. J., Zhang, C., Sackett, P. R., Thakker, K., & Young, J. Q. (2022). Can a situational judgment test identify trainees at risk of professionalism issues? A multi-institutional, prospective cohort study. *Academic Medicine*, *97*(10), 1494–1503. https://doi.org/10.1097/ACM.0000000000004756

Cullen, M. J., Konia, M. R., Borman-Shoap, E. C., Braman, J. P., Tiryaki, E., Marcus-Blank, B., & Andrews, J. S. (2017). Not all unprofessional behaviors are equal: The creation of a checklist of bad behaviors. *Medical Teacher*, *39*(1), 85–91. https://doi.org/10.1080/0142159X.2016.1231917

Dagilyte, E. & Coe, P. (2014). "Professionalism in higher education: important not only for lawyers". *The Law Teacher, 48*(1), 33-50, doi: 10.1080/03069400.2013.875303

Deng, L., & Chan, W. (2017). Testing the Difference Between Reliability Coefficients Alpha and Omega. *Educational and Psychological Measurement*, *77*(2), 185–203. https://doi.org/10.1177/0013164416658325

Dore, K., Ivan, R., Iqbal, M. Z., Sitarenios, G., Bynkoski, K., Archbell, K., Petersen, K. H., Beck, A., & Reiter, H. (2025, September 11). *In-program formative situational judgment test: Competency frameworks' analysis and pilot program findings*. Poster presented at ChangeMedEd, Chicago, IL, United States.

Foucault, A., Dubé, S., Fernandez, N., Gagnon, R., & Charlin, B. (2015). Learning medical professionalism with the online Concordance-of-Judgment learning tool (CJLT): A pilot study. *Medical Teacher*, *37*(10), 955–960. https://doi.org/10.3109/0142159X.2014.970986 md.umontreal.ca+1pubmed.ncbi.nlm.nih.gov+1

Fröhlich, M., Kahmann, J., & Kadmon, M. (2017). Development and psychometric examination of a German video-based situational judgment test for social competencies in medical school applicants. *International Journal of Selection and Assessment*, *25*(1), 94-110.

Funke, U., & Schuler, H. (1998). Validity of stimulus and response components in a video test of social competence. *International Journal of Selection and Assessment*, *6*(2), 115-123.

Goss, B. D., Ryan, A. T., Waring, J., Judd, T., Chiavaroli, N. G., O'Brien, R. C., Trumble, S. C., & McColl, G. J. (2017). Beyond selection: The use of situational judgment tests in the teaching and assessment of professionalism. *Academic Medicine*, *92*(6), 780–784. https://doi.org/10.1097/ACM.0000000000001591

Graupe, T., Fischer, M. R., Strijbos, J. W., & Kiessling, C. (2020). Development and piloting of a Situational Judgement Test for emotion-handling skills using the Verona Coding Definitions of Emotional Sequences (VR-CoDES). *Patient Education and Counseling*, *103*(9), 1839–1845. https://doi.org/10.1016/j.pec.2020.04.001

Hagen, M. & Bouchard, D., (2016). Developing and improving student non-technical skills in IT education: A literature review and model. *Informatics*, *3*(2), 7. https://doi.org/10.3390/informatics3020007

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis (8th ed.)*. Cengage.

Institute for the Advancement of the American Legal System. (2020). *Foundations for practice: The whole lawyer and the character quotient*. University of Denver.

International Engineering Alliance. (2013). *Graduate Attributes and Professional Competencies*. Washington Accord.

Iqbal, M. Z., Ivan, R., Bynkoski, K., Archbell, K, Richard, C., & Petersen, K, H. (2025). Formative SJT: Developing student personal and professional competencies, *The Score*, *47*(3), https://www.apadivisions.org/division-5/publications/score/index

Ivan, R., MacIntosh, A., Robb, C., & Walsh, C. (2024). Call to action for ethical AI in education. *The Score*, *46*(1), 5-9.

Jackson, D. & Chapman, E. (2012). Non-technical competencies in undergraduate business degree programs: Australian and UK perspectives. *Studies in Higher Education, 37*(5), 541-567. doi: 10.1080/03075079.2010.527935

Kiessling, C., Bauer, J., Gartmeier, M., Iblher, P., Karsten, G., Kiesewetter, J., Moeller, G. E., Wiesbeck, A., Zupanic, M., & Fischer, M. R. (2016). Development and validation of a computer-based situational judgement test to assess medical students' communication skills in the field of shared decision making. *Patient Education and Counseling*, *99*(11), 1858–1864. https://doi.org/10.1016/j.pec.2016.06.006

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Lievens, F., & Motowidlo, S. J. (2016). Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. *Industrial and Organizational Psychology*. *9*(1), 3-22. http://doi.org/10.1017/iop.2015.71

Lievens, F., & Peeters, H. (2008). Impact of elaboration on responding to situational judgment test items. *International Journal of Selection and Assessment, 16*(4), 345-55. https://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=6579&context=lkcsb_research

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review*, *37*(4), 426–441. https://doi.org/10.1108/00483480810877598

Lievens, F., Sackett, P. R., Dahlke, J. A., Oostrom, J. K., DeSoete, B. (2019). Constructed response formats and their effects on minority-majority differences and validity. *Journal of Applied Psychology, 104*(5), 715-726. https://psycnet.apa.org/doi/10.1037/apl0000367

Ludwig, S., Behling, L., Schmidt, U., & Fischbeck, S. (2021). Development and testing of a summative video-based e-examination in relation to an OSCE for measuring communication-related factual and procedural knowledge of medical students. *GMS Journal for Medical Education*, *38*(3), Doc70. https://doi.org/10.3205/zma001466

Mazzullo, E., & Bulut, O. (2025). *Automated feedback generation for open-ended questions: Insights from fine-tuned LLMs*. In S. Li, Z. Cui, J. Lu, D. Harris, & S. Jing (Eds.), Proceedings of Machine Learning Research: Vol. 264. Large foundation models for educational assessment (pp. 103–120). PMLR.

Mazzullo, E., Bulut, O., Jerez, D., Vo, K., Walsh, C., Sitarenios, G., & MacIntosh, A. (2026). *Is it worth the effort? A comparison of automatic educational feedback generated by base and fine-tuned LLMs*. Education and Information Technologies. https://doi.org/10.1007/s10639-026-13988-0

Mazzullo, E., Bulut, O., Walsh, C., Sitarenios, G., & MacIntosh, A. (2025). *Fine-tuning GPT-3.5-Turbo for automatic feedback generation*. In Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing (SAC '25) (pp. 40–47). ACM. https://doi.org/10.1145/3672608.3707735

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L. (2007). Situational judgement tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60*(1), 63-91. https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1744-6570.2007.00065.x?casa_token=H3yIazT_5csAAAAA:xdjdFTYgH-PED9PrGYPsC_iJt3extBOuZs7KRlfF2GEW5lFKTgOxc6loBjgUb6KfZs4KOicl_8GyhKG2tw

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.

McGill, C. M., Ali, M., & Barton, D. (2020). Skills and competencies for effective academic advising and personal tutoring. *Frontiers in Education*, 5, 135. https://doi.org/10.3389/feduc.2020.00135

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

National Association of Colleges and Employers. (2021). *Career Readiness Competencies*. Bethlehem, PA: NACE.

Patterson F., & Driver R. (2018). Situational judgement tests (SJTs). In F. Patterson & L. Zibarras (Eds.), *Selection and recruitment in the healthcare professions: Research, theory and practice* (pp. 79-112). Palgrave Macmillan.

Patterson, F., Galbraith, K., Flaxman, C., & Kirkpatrick, C. M. (2019). Evaluation of a situational judgement test to develop non-academic skills in pharmacy students. *American Journal of Pharmaceutical Education*, *83*(10), 7074.

Reiser, S., Schacht, L., Thomm, E., Bauer, J., Figalist, C., Janssen, L., Schick, K., Berberat, P. O., Gartmeier, M., & Dörfler, E. (2021). A video-based situational judgement test of medical students' communication competence in patient encounters: Development and first evaluation. *Patient Education and Counseling*, *105*(6), 1283–1289. https://doi.org/10.1016/j.pec.2021.10.012

Sahota, G. S., Fisher, V., Patel, B., JuJ, K., & Taggar, J. S. (2023). The educational value of situational judgement tests (SJTs) when used during undergraduate medical training: A systematic review and narrative synthesis. *Medical Teacher*, *45*(9), 997–1004. https://doi.org/10.1080/0142159X.2023.2168183

Saunders, L., & Bajjaly, S. (2022). Direct instruction and assessment of personal and professional skills across disciplines: Faculty perspectives. *International Journal of Teaching and Learning in Higher Education*, *33*(3), 374–384.

Saxena, A., Desanghere, L., Dore, K., & Reither, H. (2024) Incorporating a situational judgment test in residency selections: Clinical, educational and organizational outcomes. *BMC Medical Education*, 24(339). https://doi.org/10.1186/s12909-024-05310-8

Shute, V. J. (2008). Focus on Formative Feedback. *Review of Educational Research*, *78*(1), 153-189.

Sitarenios, G. (2022). Short versions of tests: Best practices and potential pitfalls. *Journal of Pediatric Neuropsychology*, *8*(3), 101–115.

Sitarenios, G. (2024a). *Phase 2 ACE Pilot Research Analysis - 1*. Technical Report. Acuity Insights.

Sitarenios, G. (2024b). *Phase 2 ACE Pilot Research Analysis - 2*. Technical Report. Acuity Insights.

Sitarenios, G., Ivan, R., Iqbal, M. Z., Walsh, C., & Moskowitz, J. (2026a). *In-program formative situational judgment test: Competency framework analysis and pilot study findings*. Proceedings of the National Council on Measurement in Education (NCME) Annual Meeting.

Sitarenios, G., MacIntosh, A., Ivan, R., Iqbal, M. Z., & Walsh, C. (2026b). *Formative situational judgment testing for professional competency development: Framework analysis and pilot findings*. Paper presented at the International Conference on Assessment in Medical Education (ICAM), Ottawa, Canada, April 16–19, 2026.

Sitarenios, G., Ivan, R., Iqbal, M. Z., & Walsh, C. (2026c). *Formative situational judgment testing for professional competency development: Framework analysis and pilot findings*. Paper presented at the International Conference on Assessment in Medical Education (ICAM), Ottawa, Canada, April 16–19, 2026.

Van de Camp, K., Vernooij-Dassen, M. J. F. J., Grol, R. P. T. M., & Bottema, B. J. A. M. (2004). How to conceptualize professionalism: A qualitative study. *Medical Teacher*, *26*(8), 696–702. https://doi.org/10.1080/01421590400019542

van der Vleuten, C. P. M., & Schuwirth, L. W. T. (2005). Assessing professional competence: From methods to programmes. *Medical Education*, *39*(3), 309–317. https://doi.org/10.1111/j.1365-2929.2005.02094.x

Walsh, C., & Ivan, R. (2026). *Measuring what matters—or what's convenient?: Robustness of LLM-based scoring systems to construct-irrelevant factors*. arXiv. https://doi.org/10.48550/arXiv.2603.25674

Walsh, C., Ivan, R., & Sitarenios, G. (2026). *AI Scoring and Feedback Generation on a Formative Situational Judgment Test*. Proceedings of NCME.

Walsh, C., Ivan, R., Iqbal, M. Z., & Robb, C. (2025). *Using LLMs to identify features of personal and professional skills in an open-response situational judgment test*. In Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Full Papers (pp. 221-230).

Walsh, C., Ivan, R., Sitarenios, G., MacIntosh, A., Iqbal, M. Z., & Robb, C. (2026). *Scaling personalized feedback on professional competencies: Large language models for open-response SJTs in medical education*. Paper presented at the International Conference on Assessment in Medical Education (ICAM), Ottawa, Canada, April 16–19, 2026.

Walsh, C., Suresh, S., & Ivan, R. (2026). Best of both worlds: Combining LLMs and traditional ML for automated scoring of an open-response situational judgment test. *Frontiers in Education*, *11,* Article 1756673. https://doi.org/10.3389/feduc.2026.1756673

Whetzel, D. L., McDaniel, M. A., & Nguyen, N. T. (2008). Subgroup differences in situational judgement test performance: A meta-analysis. *Human Performance, 21*(3), 291-309. https://doi.org/10.1080/08959280802137820

Wilson, A., Åkerlind, G., Walsh, B., Stevens, B., Turner, B., & Shield, A. (2013). Making 'professionalism' meaningful to students in higher education. *Studies in Higher Education*, *38*(8), 1222-1238. doi: 10.1080/03075079.2013.833035 https://www.tandfonline.com/doi/full/10.1080/03075079.2013.833035

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. Ottawa, ON: National Defense Headquarters.

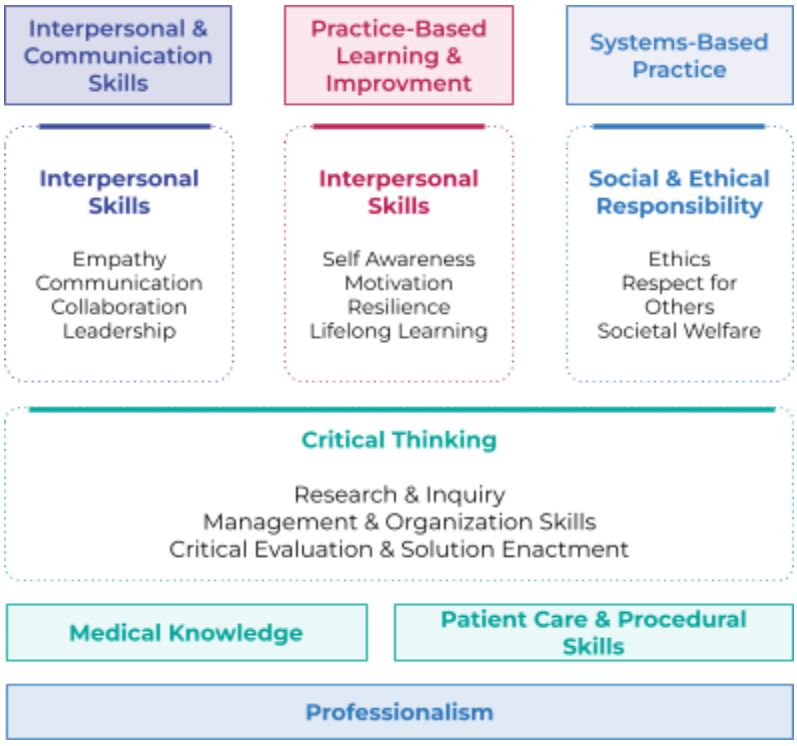
Appendix A - Crosswalk Between PSD Dimension model and Across Program Competency Frameworks

Medical Education

AAMC/AACOM/ACGME Foundational Competencies for Undergraduate Medical Education

Figure A.1 illustrates the conceptual alignment between the 2024 Foundational Competencies for Undergraduate Medical Education (AAMC, AACOM, & ACGME, 2024) and the PSD dimensional model. While the PSD assessment does not target technical medical skills directly, subcompetencies within Medical Knowledge and Patient Care & Procedural Skills show conceptual overlap with the Critical Thinking dimension. The domain of Professionalism spans all four PSD dimensions, reflecting its integrative nature and reliance on interpersonal, intrapersonal, and ethical competencies alongside critical reasoning.

Figure A.1. AAMC/AACOM/ACGME Foundational Competencies for Undergraduate Medical Education.

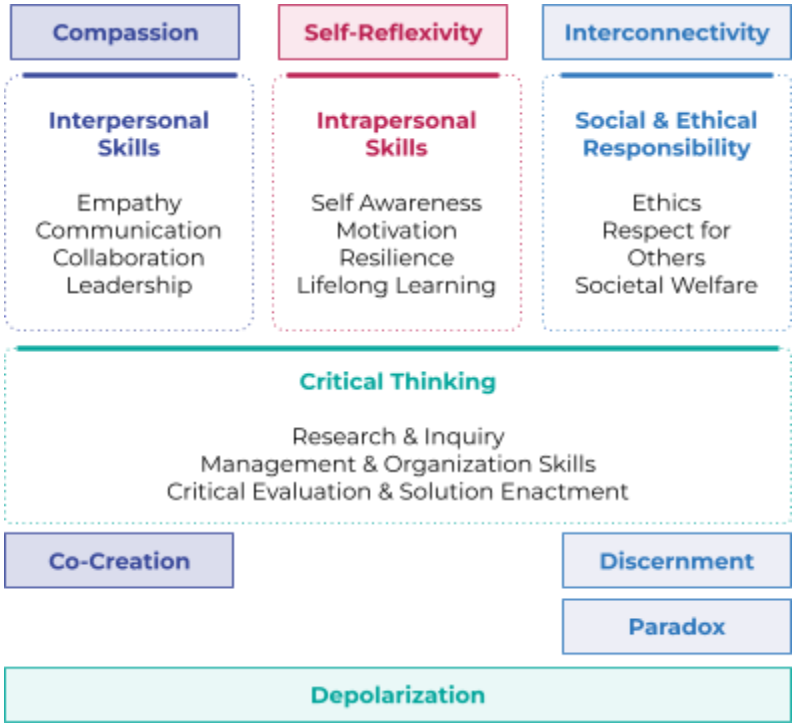


Business Education

AACSB Accelerators

Figure A.2 below illustrates the conceptual alignment between the AACSB Accelerators competency framework ("Seven Competencies of a Societal Impact Leader": AACSB International, 2020) and the PSD dimensional model. Most competencies align primarily with the Social Consciousness dimension, reflecting the framework's emphasis on ethical leadership and societal impact. The competency of Depolarization spans multiple domains, drawing on skills across all four PSD dimensions and highlighting the integrative nature of professional effectiveness in complex social contexts.

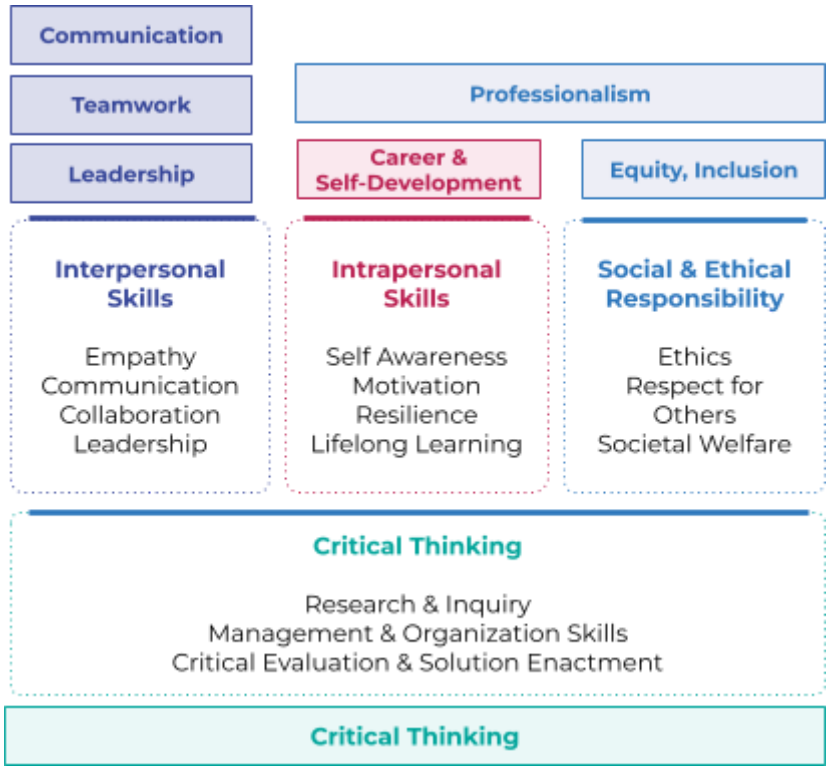
Figure A.2. AACSB Accelerators competencies mapped to the PSD dimensional model.



NACE Career Readiness Competencies

The figure below illustrates the conceptual alignment between the NACE Career Readiness Competencies and the PSD dimensional model. As outlined by National Association of Colleges and Employers (2021), most competencies map onto one or more PSD dimensions. The "Technology" competency reflects technical skills that fall outside the scope of the present formative SJT. In contrast, the "Professionalism" competency spans multiple domains, aligning most closely with the Intrapersonal and Social & Ethical Responsibility dimensions and highlighting the integrative nature of professional behavior.

Figure A.3. NACE competencies mapped to the PSD dimensional model.

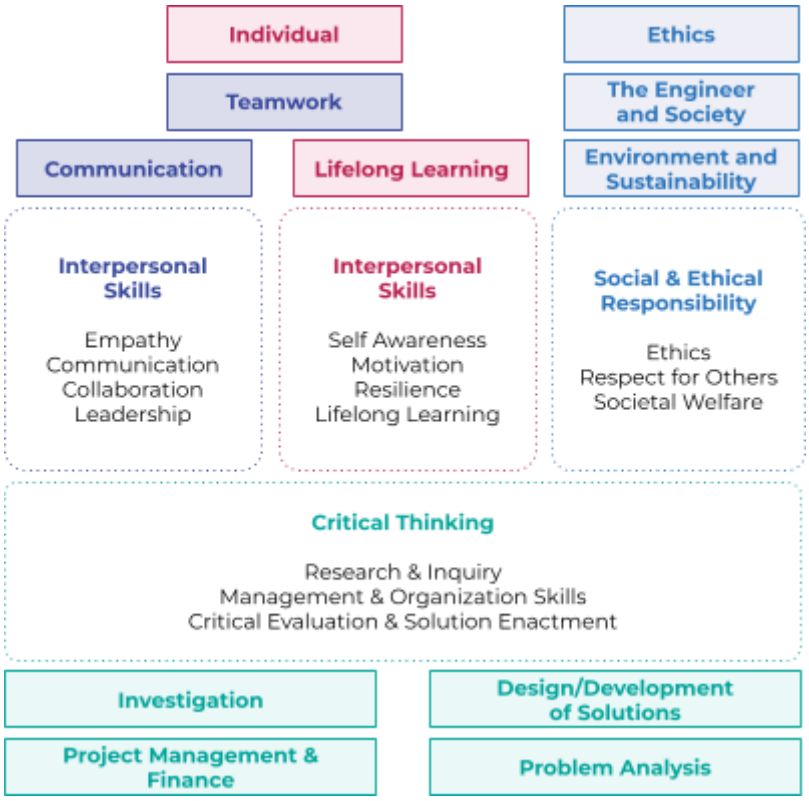


Engineering

Washington Accord Graduate Attributes

Figure A.4 illustrates the conceptual alignment between the Washington Accord Graduate Attributes and the PSD dimensional model. As described by the International Engineering Alliance (2013), the 12 attributes map across multiple PSD dimensions. Most attributes align primarily with a single dimension; however, the “Individual and Team Work” attribute spans both the Interpersonal and Intrapersonal domains, reflecting the combined importance of personal responsibility and collaborative effectiveness.

Figure A.4. Washington Accord Graduate Attributes competencies mapped to the PSD dimensional model.

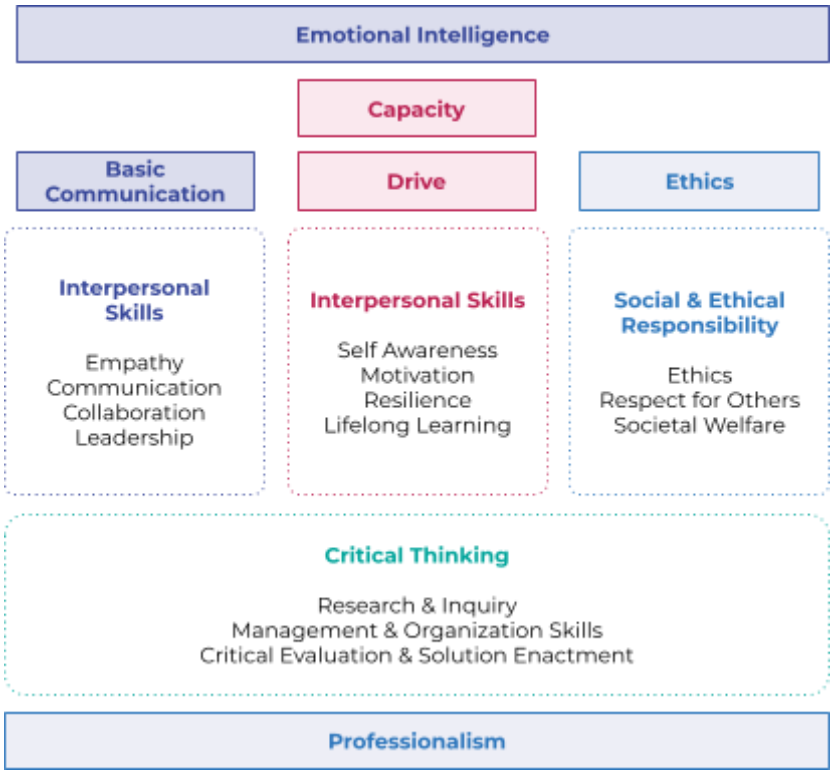


Law

IAALS Foundations

Figure A.5 illustrates the conceptual alignment between the IAALS Foundations competencies and the PSD dimensional model. As outlined by the Institute for the Advancement of the American Legal System (2020), these competencies map across multiple PSD dimensions, with particular emphasis on professional judgment, interpersonal effectiveness, and ethical responsibility. Several competencies span more than one dimension, reflecting the integrated nature of legal practice skills.

Figure A.5. IAALS Foundations competencies mapped to the PSD dimensional model.



Appendix B - Prototype Research Results

Initial Psychometric Findings: Reliability

Initial technical reports (Sitarenios, 2024a, 2024b) and conference paper presentations (Sitarenios et al., 2026a, 2026b, 2026c; Dore et al., 2025) provided early evidence supporting the reliability and construct validity of the PSD prototype. For the open-ended items, domain-level reliability estimates ranged from .79 to .92. Multiple choice questions tapping each construct were also included in the prototype to test out both response formats. In contrast, reliability for the aggregated multiple-choice score was lower at .63.

Aggregate data (N = 338) from two programs who participated in the pilot phase of the development of the prototype was used to further test the viability of PSD. A total of 219 students included students from the Class of 2027 from New York Medical College's School of Medicine MD program, and an additional 119 second year students from University of Waterloo, School of Pharmacy program.

Using these data, internal consistency reliability was examined and the results are shown in (Table B.1) In this analysis, all four domain scores showed strong internal consistency, with *alpha* values ranging from .85 to .89 and *omega* values from .88 to .92.

Table B.1. PSD prototype reliability by domain score - open ended responses.

PROTOTYPE DOMAINS	**# items**	**Alpha**	**Omega**	**Median Item *r***
Reading the Situation	7	.86	.90	.49
How it affects You	7	.85	.88	.45
Effective Interacting	10	.89	.92	.45
Resolution enactment	8	.88	.91	.44

Prototype: Validity

Correlations with Other Metrics

For 61 students from the University of Waterloo Pharmacy, PSD prototype data was connected to variables provided from the program. These students provided consent for the data to be used for this research study. These data included Casper scores which provided an opportunity to look at the relationship between scores derived from these two assessments. Given their structural similarities, it was hypothesized that there would be a relationship between the two. However, the findings showed only a mild correlation between the two, with correlations ranging from .11 to .21 between the Casper score and the various PSD prototype scores (see Table B.2).

Table B2. Casper x Prototype correlations.

PROTOTYPE DOMAINS	**Casper Correlation**	**Sig**
Reading the Situation	.21	p = .104
How it affects You	.20	p = .122
Effective Interacting	.20	p = .122
Resolution enactment	.11	p = .399