CASPER TECHNICAL MANUAL

ACUITY INSIGHTS

This technical manual was developed in accordance with the Standards for Educational and Psychological Testing (2014) and is intended to provide readers with a comprehensive documentation of Casper test-related material and supporting evidence to date. This manual will be updated annually to ensure the most up to date information is available.

Preface

Casper was born out of the need to have a valid and reliable way to evaluate applicants' personal and professional characteristics. Dr. Kelly Dore and Dr. Harold Reiter (the Casper test creators) recognized that of the thousands of applicants to medical schools only a small percentage of them were being evaluated for these skills prior to interview. While reference letters, CVs and personal statements all try to give insight into the applicant beyond "book smarts," they fail to give a reliable differentiation between candidates in the admissions setting. The advent of generative AI has further reduced the value of reference letters and personal statements - many admissions leaders are questioning whether the effort of reviewing this material is worth the burden to evaluate.

Casper is a unique situational judgment test (SJT) that provides test takers with a series of hypothetical scenarios and assesses the individual's response to the situation using open response, as opposed to closed or fixed response like multiple choice or choose the best option test designs. Open response is important as it encourages respondents to explore the underlying reasoning for deciding what to do and embraces the idea that there are often many intelligent viewpoints and courses of actions for a complex situation. Casper also includes both video and text scenarios. The use of video scenarios most closely parallels how dilemmas occur in real life and enhances the ecological validity of the test paradigm. The unique test format illuminates and differentiates applicants in a rich fashion not possible with closed response assessments.

Originally developed for use within medical schools, the Casper test has since found broad application and has been tailored for use in graduate medical education and across programs relating to health science including, physical therapy, occupational therapy, nursing, dentistry, veterinary medicine, pharmacy, physician assistant programs and more. Increasingly Casper is being used to assess skills for programs outside of health sciences including teachers education, business education, engineering education, and more.

Our entire Acuity team continues to work with a steadfast focus to offer products, services, and insights that uncover everyone's full potential and nurture their holistic success. Together with our academic program partners, subject matter experts, and applicants, we continue to find ways to provide admissions decision makers with a higher-fidelity view to allow for clearer and more defensible decision-making.

To date, Casper has been completed by over 900,000 applicants worldwide, for over 600 individual programs across six countries, and in three languages. I am excited

that we now have available this technical manual which documents the enormous amount of both internal and external analyses conducted on Casper to date (over 150) across psychometric elements such as reliability, validity and equity. Our goal is to create the best assessments possible, and this technical manual is reflective of our commitment to quality, fairness, and continuous improvement. This body of support, gathered over nearly 20 years, is truly what distinguishes Casper from other measures of personal and professional characteristics.

While this technical manual underscores the vast and solid research activities that back Casper to date, we continue to look for ways to evolve Casper and our related products, services and insights. We welcome your reactions and comments on this material - feedback, different interpretations of data, or suggestions for further collaborative research. In the end, improving this assessment, through whatever means, benefits us all.

Matt Holland

CEO, Acuity Insights

Introduction

Professional programs desire applicants who not only possess the technical abilities to complete their studies successfully, but also those who demonstrate exceptional personal and professional qualities (Mahon et al., 2013). To provide academic institutions with a reliable and valid assessment of these qualities, the team at Acuity Insights has meticulously crafted the content of the Casper test to ensure a harmonious alignment with the core competencies outlined by major organizations such as the Association of American Medical Colleges (2021), the Royal College of Physicians and Surgeons of Canada (2021) and the Australian Institute for Teaching and School Leadership (2017).

Initially developed by Dr. Kelly Dore and Dr. Harold Reiter, Casper provides a holistic impression of applicants by providing academic programs with a reliable and valid measure of social intelligence and professionalism to complement the technical skills that are typically assessed during the application process. Throughout its lifespan, Casper's development has been supported by several organizations including the Medical Council of Canada, the Stemmler Fund of the National Board of Medical Examiners, and the Faculty of Health Sciences at McMaster University. To date, Casper has been implemented in over 600 unique programs, across 6 countries, in 3 languages, and has been completed by more than 900,000 applicants.

Acuity Insights is driven to create a world served by exceptional professionals. We do this by empowering higher educational institutions to look beyond book smarts and to identify, select, and nurture exceptional professionals by using unique integrated data to uncover insights and understand the actions most effective at driving the outcomes programs seek. Each team within Acuity Insights works together in unique ways to achieve this collective organizational mission. For the Research Team, this means creating socially responsible assessments that have strong reliability and validity, and that work to further equity and fairness. Our data-driven team is open-minded and agile; we value and facilitate creative exploration and innovation by integrating multiple perspectives to help identify exceptional professionals and foster their full potential.

The purpose of this manual is to provide readers with a comprehensive overview of Casper test-related material and supporting evidence to date. This document provides a singular location where readers can learn about the construction of the test, the rich quantitative and qualitative foundations of Casper, the psychometric properties, and program specific outcomes.

It is important to disclose that this document was prepared by the Research Team at Acuity Insights. All members of the Research Team are employed by Acuity Insights who has proprietary ownership of the Casper test. However, a significant portion of data presented in this manual has been assessed under the careful review of several

academic partners, through conference proceedings, and through peer-reviewed journals.

Intended Purpose and Use

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 students, but also the non-technical skills. Scholars have highlighted the importance of assessing and developing professionalism across higher education including, but not limited to, psychology, law, business, international relations, etc. (Wilson et al., 2013 & Dagilyte & Coe, 2014). With evidence that unprofessional behaviour in medical school, for example, is strongly associated with future physician disciplinary action, it is evident that there is a particular need to assess personal characteristics during the admissions stage (Papadakis et al., 2005). While reference letters and personal statements have traditionally been used as a measure of personal and professional competencies, these metrics have often been found to be unreliable and invalid sources of evidence (Albanese et al., 2003).

The Casper test was designed to fill this specific gap in the admissions process. The overarching goal of Casper is to provide academic programs with reliable and valid information regarding applicants' social intelligence and professionalism. The intent is to provide information early in the admissions process so that it can be incorporated alongside measures of pertinent academic knowledge and technical skills (*e.g.,* GPA, MCAT, GMAT, GRE, SAT, etc.) during the decision-making process. The format of the Casper test is similar to that of the multiple-mini interview (MMI; Eva et al., 2004b), a common metric used for assessing personal and professional attributes of applicants to professional programs. Casper was by no means designed to replace the MMI, but rather designed to provide incremental value by bringing strong candidates to the interview process, as MMIs are both time and resource intensive, making it infeasible to offer to every applicant. Although initially developed for medical and health science fields, the content of the Casper test is not domain-specific, and therefore can be used to assess core aspects of professionalism across a variety of programs including, but not limited to, business (Jackson & Chapman, 2012), law (Dagilyte & Coe, 2014), and information technology (Hagen & Bouchard, 2016). Regardless of the field, professionals are expected to demonstrate the very aspects that the Casper test targets: strong communication skills, the ability to empathize with others, ethics, problem solving, motivation and resilience, self awareness, collaboration, equity, and professionalism. Casper is an accessible, affordable, reliable, and valid measure of social intelligence and professionalism that enriches the portrayal of applicants and provides programs with a better holistic understanding of each applicant from the start.

A Note on the Changing Structure of the Casper Test

The Casper test, as of the 2023-2024 cycle, consists of 14 scenarios: 8 of which require a typed response and 6 of which require a video response. However, prior to the 2022-2023 application cycle, Casper consisted of 12 scenarios for which applicants were required to type their responses to three open-ended questions. With evidence to suggest that adding a video-response component may enhance the equity and fairness of the test, Casper was extended to 15 scenarios for the 2022-2023 application cycle: 9 of which required a typed response and 6 of which required a video response. While it was mandatory for applicants to complete all 15 scenarios of the test, only select programs were provided access to the video-response scores for research purposes, thus, only the scores from the typed-response section of the Casper test were provided to programs. The decision to supply programs with only the scores from the typed-response section was made to ensure that only scores with strong psychometric properties and proven validity evidence were used in high-stakes decision making. With evidence to support the addition of the video-response sections, the video-response component was officially incorporated into the Casper test beginning in the 2023-2024 application cycle.

All information in the Casper Technical Manual (unless otherwise stated) will include information only on the scores that programs were provided for decision-making purposes. That is, all information prior to the 2023-2024 application cycle reflects scores that included only a typed response format.

Test Structure

The content of each test is unique and consists of a variety of scenarios which present applicants with either a video prompt (*i.e.,* actors playing out an interaction) or a text prompt (*i.e.,* a short written statement) for them to consider. After watching or reading the scenario, applicants are presented with a series of open-ended questions that they are asked to respond to in a short time-frame. As discussed above, the format of the Casper test has changed over the years to further enhance the psychometric properties, equity, fairness, and applicant experience. For ease of reference, the table below outlines the details of the former and current structure of the Casper test.

	Typed Response Scenarios			Video Response Scenarios		
Application Cycle	**Number of Scenarios**	**Number of Questions Per Scenario**	**Time to Respond**	**Number of Scenarios**	**Number of Questions Per Scenario**	**Time to Respond**
2021-2022 (and prior)	12	3	5 minutes total for each set of 3 questions	NA	NA	NA
2022-2023	9	3	5 minutes total for each set of 3 questions	6	3	1 minute each question
2023-2024 (and after)	8	3	5 minutes total for each set of 3 questions	6	2	1 minute each question

What a Casper Score Tells You

Casper is a situational judgment test (SJT) that measures social intelligence and professionalism, which are understood as a collection of skills and behaviours which everyone uses each time a new situation is presented. Casper presents applicants with scenarios and questions that allow them to demonstrate these particular aspects – and, therefore, the extent to which they can respond to challenging scenarios in a professional and socially intelligent manner. These aspects include empathy, communication, collaboration, self-awareness, resilience, equity, motivation, problem solving, and ethics. Casper incorporates all of these aspects into the test, thereby providing an overall measure of how professional and socially intelligent an applicant is likely to be.

A standard definition for each aspect was developed prior to scenario creation based on extensive literature and stakeholder review. These definitions are provided below. For each aspect, we have also outlined a set of associated demonstrable behaviours, but these are reserved for internal use only to ensure test security.

- **Collaboration**: Functions interdependently by balancing mutual and individual goals; demonstrates openness to others' perspectives and input; reaches consensus in service of a larger mission.
- **Communication**: Effectively interacts with the intent of understanding and being understood in different contexts.
- **Empathy:** Takes perspective of others; considers others' feelings and context in a given situation.
- **Equity**: Acknowledges, appreciates, and respects individual and cultural values, preferences, experiences, and needs of others.

- **Ethics**: Maintains moral principles that dictate personal and professional behaviour; prioritizes integrity, honesty, justice, and respect for personal autonomy.
- **Motivation**: Reflects upon methods of improvement; actively and persistently applies oneself to achieving one's personal best.
- **Problem Solving**: Recognizes and defines problems; develops process to approach and solve problems; evaluates approaches for efficacy.
- **Resilience**: Successfully adapts to change; learns from adversity.
- **Self-Awareness**: Actively identifies and stores information about one's self; candidly reflects upon and explores this information.
- **Professionalism**: Demonstrates and maintains high personal standards of accountability and thoughtfulness; respectfully behaves according to regulations.

In addition to the construct measured by way of these definitions and scoring guidelines, the Research Team at Acuity Insights recently conducted a large scale mixed-methods study to examine the implicit factors of social intelligence and professionalism that are captured within Casper scores. Using both qualitative and quantitative approaches, this study identified several subtle and nuanced factors of social intelligence and professionalism that are reflected in the scores. This study found that Casper scores provide insight into the extent to which an applicant can:

- Reflexively use examples to demonstrate key aspects of social intelligence and professionalism
- Understand and interpret social dilemmas in nuanced and complex ways
- Thoughtfully justify their approach to the presented situation
- Provide novel or creative ideas or solutions to the presented situations
- Apply critical thinking to situations that require a professional approach
- Demonstrate sound professional and ethical judgment
- Carefully consider multiple perspectives

The details of this study are currently being drafted to be submitted to a peer-reviewed journal.

Overview of Situational Judgement Tests and the Unique Components of Casper

Situational Judgement Test (SJT)

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 three theoretical underpinnings of SJTs: 1) behavioural consistency theory, 2) implicit trait policies theory, and 3) theory of planned behaviour. 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). Finally, the theory of planned behaviour proposes that an individual's behaviour in particular situations is impacted by their unique values and attitudes (Ajzen, 1985). 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.

Behavioural Tendency Questions

The Casper test 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 Casper 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).

Use of Independent Rater Scores

Each section of an applicant's Casper test is scored by a unique rater, a method which provides several benefits. First, this practice ensures that applicants' total score (the average of all independent ratings) reflects perspectives from a wide range of individuals who are likely representative of the population the applicant would interact with in their given field. This practice also assists in diluting any personal (explicit or implicit) bias a single rater may have. Additionally, the use of independent scores mitigates the potential for a "halo effect" to occur and reduces the influence of context specificity. The halo effect refers to the notion that one's first impression of another person will influence subsequent judgements (Nisbett & Wilson, 1977). For example, if a rater has a negative impression on an applicants' first section, they may inadvertently provide lower marks throughout the remainder of the test. Since a rater can only rate one scenario of an applicant's score, this phenomenon is not of concern. Taken together, averaging applicants' scores from several unique raters provides a well-rounded and holistic impression of each applicant; something that cannot be achieved from individual ratings.

Time Limited Responses

The time limit for each scenario is employed in order to evoke genuine responses from applicants. Unlike personal statements where applicants have weeks or even months to revise responses, Casper requires applicants to think on their feet just like they would in real life. This time restraint makes it more likely that applicants will provide an authentic response rather than craft a response that they believe would be attractive to raters or programs. This approach often results in a wider variety of responses which makes it easier for programs to really differentiate between applicants.

The Scenario Questions

For each scenario, applicants are presented with unique open-ended questions related to that particular scenario which probe for the various aspects evaluated in the Casper test. Each typed-response scenario is accompanied by three unique open-ended questions while each video-response scenario is accompanied by two unique open-ended questions. Using multiple questions for each scenario allows raters to assess the depth and consistency of applicants' overall response. Applicants are scored globally on these questions, resulting in a single score for each scenario.

CHAPTER 1: RELIABILITY

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, and it is particularly important for tests that are used for high-stakes decisions such as program admission, for which the need for strong reliability increases considerably (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

Casper has been evaluated extensively and has yielded strong supportive evidence for the reliability of the test across several potential sources of variance. Each facet of reliability is outlined briefly below with more comprehensive discussions available in the following subsections. When applicable, comparisons are also drawn to similar measures to provide a meaningful frame of reference for the values which are discussed.

Internal Consistency Reliability. Across all verticals, Casper has consistently demonstrated strong levels of internal consistency reliability with alpha values ranging from α=0.78 to α=0.87.

Standard Error of Measurement. The standard error of measurement (SEM) has been quite low and uniform across application cycles indicating that Casper provides precise estimates of applicants' true score. On average, annual estimates of SEM in z-scores range from 0.36 to 0.46.

Inter-Rater Reliability. Using a variety of methods for estimating inter-rater reliability (IRR), Casper consistently evidences strong consistency between raters indicating that applicants would receive similar scores regardless of which raters score their test. This has been examined by comparing groups of raters (mean IRR=0.88; range:0.86-0.89) and by comparing independent raters to the average of other raters at the scenario and test level (mean IRR=0.82; range:0.54-0.95).

Test-Retest & Parallel-Forms Reliability. Although applicants are not permitted to write the Casper test more than once during the same application cycle within the same vertical, there are two circumstances in which they can write the test more than once: (1) applicants can write Casper tests in the same vertical in different application cycles and (2) applicants can write Casper tests in the same application cycle, for different verticals. These circumstances subsequently provide data from which evaluations of parallel-forms reliability and test-retest reliability can be conducted. On average, an intra-class correlation coefficient (ICC) of 0.75 is observed

for tests in different application cycles, with a similar ICC of 0.76 evidenced in Casper tests taken within the same year.

Generalizability. The Casper test has produced high generalizability coefficients since its inception in 2009. Full test g-coefficients have ranged from 0.82 to 0.85 with most recent analyses indicating that the largest source of variance (22.6%) in Casper scores is applicant ability.

Internal Consistency Reliability- *Evidence to show that Casper scenarios 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. A coefficient alpha of α=0.70 is often cited as the minimum threshold for an adequate level of internal-consistency (Cortina, 1993), but high-stakes assessments (such as Casper), often aim to attain reliability values in the 0.80 to 0.90 range.

Casper has demonstrated excellent internal consistency across its lifespan. As demonstrated in Table 1 and Figure 1, data compiled from a total of 861,329 applicants across 1,049 unique test instances shows that Casper's average alpha values have remained continuously high each year.

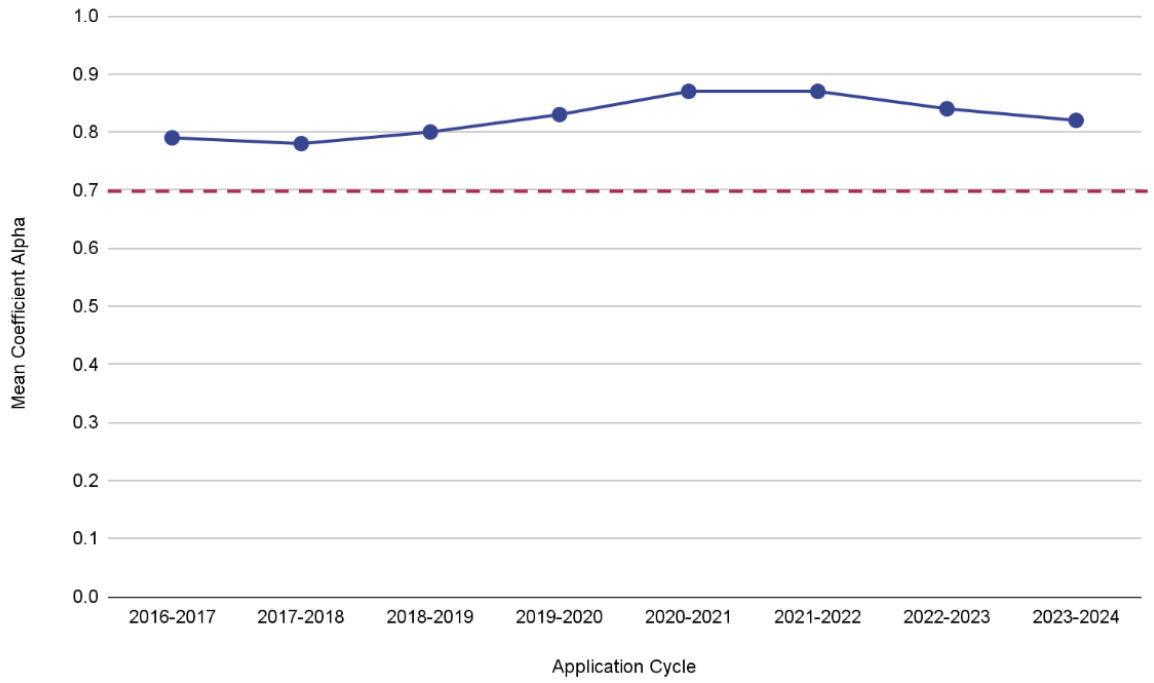
Table 1

Mean Coefficient Alpha Values Across All Verticals

Application Cycle	Number of Test Instances	Total Applicants	Mean Coefficient Alpha
2016-2017	*n*=66	*n*=30,693	α=0.79
2017-2018	*n*=86	*n*=56,387	α=0.78
2018-2019	*n*=106	*n*=88,599	α=0.80
2019-2020	*n*=119	*n*=105,995	α=0.83
2020-2021	*n*=175	*n*=156,838	α=0.87
2021-2022	*n*=182	*n*=151,193	α=0.87
2022-2023	*n*=160	*n*=138,487	α=0.84
2023-2024	*n*=155	*n*=133,137	α=0.82

Figure 1

Mean Coefficient Alpha Values Across All Application Cycles To Date



Internal consistency reliability remains consistent across languages. Importantly, coefficient alpha values are similar across both languages for which the Casper test is available: English and French. An examination of the reliability of 10 French language tests in the 2021-2022 application cycle, evidenced an average coefficient alpha value of α=0.85 (ranged from 0.81 to 0.92). This is similar to the results of the 2020-2021 application cycle for which an average coefficient alpha of α=0.82 was observed (range: 0.76 to 0.85) across 11 French language Casper tests. These results indicate that the Casper test is effective across both languages and that the product and processes can be translated while maintaining high-quality reliability.

Internal consistency reliability relative to other metrics in the same area of assessment. These estimates of internal consistency are much higher than those often produced by Multiple Mini Interviews (MMIs) which are also commonly used to assess applicants' personal and professional skills. Briefly, MMIs consist of several short interview stations, each with an independent rater. This structure was designed to dilute the impact that personal bias from interviewers may have on an individual (Eva et al., 2004). An evaluation of MMIs that are comparable to Casper (*e.g.,* those with a minimum of 8 stations) evidenced coefficient alpha values that were often lower than the α=0.70 threshold and consistently lower than values produced by the Casper test (Dowell et al., 2012; Jerant et al., 2012; O'Brien et al., 2011). Specifically, estimates of internal consistency ranged from α=0.68 (Jerant et al., 2012) to α=0.73

(O'Brien et al., 2011). A slightly higher coefficient alpha (α=0.83) has been reported for an MMI used in a small sample of applicants (Oliver et al., 2014), but the associated value was still lower than that produced by Casper across several application cycles.

The consistency of high alpha values demonstrated across numerous years, programs, countries, and languages indicates that all scenarios of the Casper test are consistently measuring the same construct.

Standard Error of Measurement- *Evidence that applicants' Casper 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 2, the average SEM of Casper z-scores based on data from 861,329 applicants has been uniformly low across all application cycles. Overall, Casper's consistently low SEM across application cycles, countries, and programs further supports the reliability of the instrument indicating that applicants' test scores will typically be very close to their true scores.

Table 2

Mean SEM Values Across All Verticals

Application Cycle	Number of Test Instances	Total Applicants	Mean z score SEM
2016-2017	*n*=66	*n*=30,693	0.44
2017-2018	*n*=86	*n*=56,387	0.46
2018-2019	*n*=106	*n*=88,599	0.44
2019-2020	*n*=119	*n*=105,995	0.40
2020-2021	*n*=175	*n*=156,838	0.36
2021-2022	*n*=182	*n*=151,193	0.36
2022-2023	*n*=160	*n*=138,487	0.40
2023-2024	*n*=155	*n*=133,137	0.42

Inter-Rater Reliability- *Evidence that Casper scores are consistent across different raters.*

Inter-rater reliability (IRR) assesses the consistency in scores between raters (Price, 2017), with high levels of IRR indicating that applicants would receive a similar Casper score across different groups of raters. Intra-class correlation coefficients (ICC) are often used to measure the consistency of scores as it provides an estimate of both the agreement and correlation between raters (Koo & Li, 2016). ICC values are often interpreted as follows (Koo & Li, 2016).

- 0.00 to 0.50 = poor
- 0.50 to 0.75 = moderate
- 0.75 to 0.90 = good
- 0.90 to 1.00 = excellent

Casper's unique rating structure (using a unique rater for each scenario of an applicant's test) must be taken into account when calculating and evaluating rater agreement. To date, the team at Acuity Insights has examined the IRR of Casper in a variety of ways.

To begin, one of the ways IRR has been evaluated by the team at Acuity Insights was by examining rater agreement *within* an applicant's test. That is, each rater's score for a single applicant was compared to the average score from all other raters for that same applicant. In the examination of 367 raters from the 2022-2023 application cycle who scored the typed-response scenarios, we saw an average correlation of 0.54 indicating a moderately strong relationship between raters for a single applicant across unique scenarios.

To evaluate IRR even further, and from a different perspective, the IRR of Casper has also been assessed by randomly oversampling 10% of applicants from a variety of tests and re-rating all responses (12 scenarios at the time) by one or two other raters. The ICC was then calculated using the average score from the 12 scenarios. This approach was adopted to provide a representation of IRR at the test level when different raters are used. Across both the 2019-2020 and 2020-2021 application cycles, the average IRR was very high with ICC values of 0.89 (n=1,020) and 0.86 (n=26,974), respectively.

Further, during the Casper pilot study conducted in 2009, rater consistency was examined by comparing the average score for each individual rater to the average score of all other raters. Analyses were conducted across audio-recorded responses and typed responses, both of which produced high IRR values of 0.82 and 0.81, respectively (Dore et al., 2009). The second part of the pilot study produced similarly impressive IRR estimates for the full test (0.95), video-response scenarios (0.92), and typed-response scenarios (0.90). Later on during the 2018-2019 application cycle, rater consistency was similarly high as evidenced from an internal G-study (ICC=0.85 for the full test).

Inter-rater reliability relative to other metrics in the same area of assessment. Comparatively, although estimates of IRR for MMIs are scant (as many use a single interviewer per station; Rees et al., 2016), the data that is available indicates that Casper often produces substantially higher IRR relative to the MMI. In one of the few studies which used two raters per station for medical school applicants (*n*=444), Sebok et al. (2014) reported correlations between raters ranging between 0.41 and 0.69. Similar to Casper raters, the raters of the Sebok et al. study used a 9-point Likert scale for scoring; however, estimates of IRR were noticeably lower than what has often been produced from Casper raters over the last several years. IRR from the MMI used at the University of Calgary Veterinary medicine program was similar to that of the Sebok et al. study, producing an IRR of 0.52 for their 103 applicants (Hecker et al., 2009).

Parallel Forms Reliability & Test-Retest Reliability- *Evidence that Casper scores remain consistent across time and across different variations of the test.*

Parallel forms reliability measures the consistency of an individual's score (or in the case of Casper, an individual's rank) across distinct versions of a test (Price, 2017). The underlying notion of parallel forms reliability is that although each version of the test contains unique items, they are interchangeable in that they each measure the same constructs, are administered using the same format, are similar in terms of difficulty, and produce similar score distributions (AERA et al., 2014). Test-retest reliability measures the consistency of an individual's scores at different times (Weir, 2005). In the case of Casper, high parallel forms reliability metrics indicate that an applicant is likely to receive a similar score on different versions of the Casper test. In the same vein, high test-retest reliability metrics indicate that an applicant is likely to receive a similar score at another point in time.

For the Casper test, parallel forms reliability and test-retest reliability are evaluated in tandem. For each application cycle, multiple test dates are available for each program, but applicants are only permitted to write one test per vertical, per application cycle. This one test limit, in conjunction with the fact that over 100 unique versions of the Casper test are administered each year make it impractical to conduct traditional assessments of parallel forms reliability for every single version. However, several effective methods for evaluating parallel forms reliability are employed. The following analyses are discussed in turn: (1) comparing first and second attempts at Casper for the *same* vertical in *different* years and (2) comparing first and second attempts at Casper for *different* verticals in the *same* year. Thus, when applicants write a second version of the test, we are able to assess how consistent their scores were across different versions of the test (*i.e.,* parallel forms reliability) and at different points in time (*i.e.,* test-retest reliability).

There are two initial steps that are taken to ensure high test equivalency and subsequently promote high parallel forms reliability. The first is the meticulous and

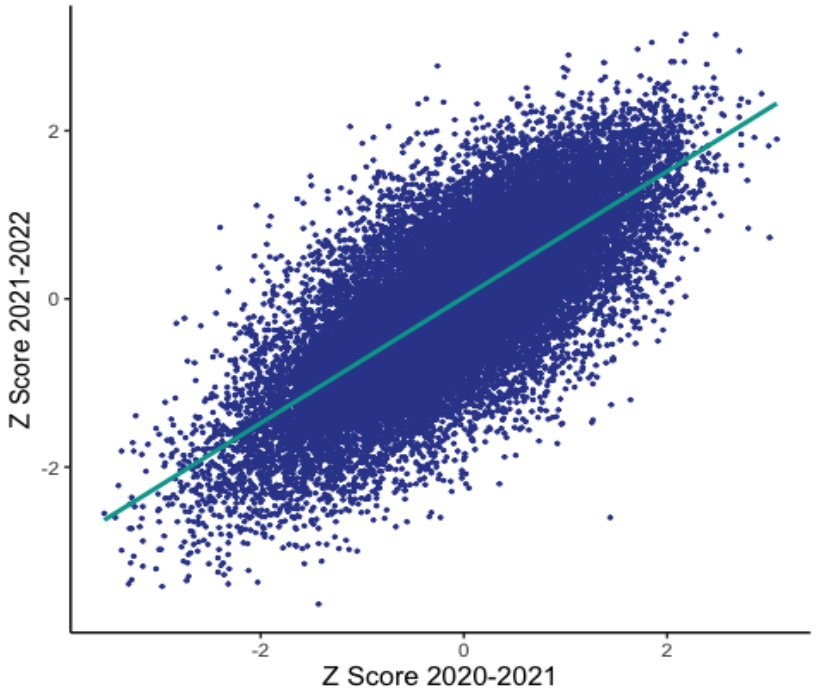
iterative content development and revision process that every Casper item undergoes (details available in the Content Validity section). The second, is the use of z-scores to help mitigate any between-test differences. The use of z-scores within a single test instance ensures that applicants are only being assessed relative to others who wrote the exact same test, thus mitigating any potential between-test differences in scores that may be attributed to test difficulty. Another positive attribute of z-scores is that they allow fair comparison across different samples since the scores are transformed onto a common scale, all of which have a mean of zero and a standard deviation of one (Cohen, 2013).

Comparing Casper Scores Across the Same Verticals in Different Application Cycles. Casper has demonstrated strong relationships between applicants' scores on their first and second tests within the same vertical, but different years. The most recent analyses examined *n*=24,147 applicants who wrote the Casper test twice within the same vertical: once during the 2020-2021 application cycle and again during the 2021-2022 application cycle. These results showed a strong relationship between applicants' scores on the first (mean z-score=-0.09) and second (mean z-score=-0.05) iteration of the test with an ICC of 0.85 (see Figure 2).

Positive results were also found in previous analyses which examined a sample of *n*=3,548 applicants who wrote Casper twice, once in 2018 and once in 2019, for the Canadian Health Science 2 vertical. These analyses produced a moderately strong relationship between applicants' scores on the first (mean z-score= -0.10) and second (mean z-score= 0.17) iteration of the test (ICC= 0.65). Again, these results suggest that applicants' scores are consistent across alternate versions of the test and across different time periods.

Figure 2

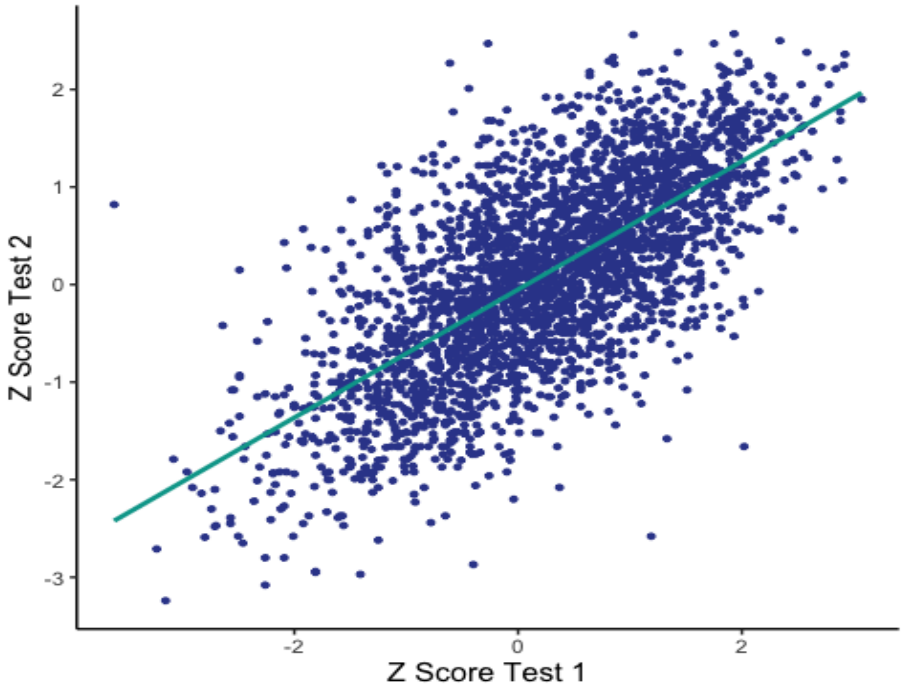
Relationship Between First And Second Casper Tests For Applicants Who Wrote For The Same Vertical In Separate Application Cycles



Comparing Casper Scores Across Different Verticals in the Same Application Cycle. Often, applicants will apply to programs that fall under different verticals. For example, applicants applying to Canadian Medical programs may also apply to United States Medical programs in the same year. Since these programs are classified as two separate verticals, it is possible for applicants to write two unique Casper tests within the same year. As can be seen in Figure 3, an analysis of z-scores for applicants (*n*=2,831) who wrote the Casper test twice in the same application cycle (2021-2022), but for different verticals demonstrated a strong relationship (ICC=0.81). These results are consistent with previous analyses which explored the z-scores of applicants (*n*=2,432) who wrote the Canadian Health Science 2 test and the US Health Science 2 test in the same year (2015, 2016, or 2017). These previous results also demonstrated a strong relationship (ICC=0.71). The time difference between two tests written in the same application cycle (*e.g.,* several weeks apart) is drastically smaller than for tests written across different application cycles (*e.g.,* approximately one year apart), thus reducing the extent to which natural learning or maturation could occur. The strong relationship between applicants' scores on tests within the same application cycle demonstrates both strong parallel forms reliability and strong test-retest reliability.

Figure 3

Relationship Between Applicant Casper Scores From Unique Verticals Within The Same Year



It should be noted that there is a possibility that these results are impacted by practice effects. However, generally, it appears that practicing or receiving coaching on the Casper test has minimal impact on test scores. A detailed discussion of these findings is available in the 'Casper and Test Preparation & Practice' section of the manual.

Parallel forms/test-retest reliability relative to other metrics in the same area of assessment. To put these values into perspective, they can be compared to information collected from similar measures such as the MMI. Results from an analysis of 17 medical school applicants who repeated the MMI at the same institution in 2006 and 2007 indicate that the difference in applicants' scores across the two iterations was 0.72 (Cohen, 2013). Another study showed that correlations between applicants' first and second attempts at an MMI ranged between 0.65 to 0.72 (Dore et al., 2009), results which are very similar to, albeit smaller than, those produced by Casper.

Generalizability- *Evidence that Casper is generalizable.*

Derived from generalizability theory, generalizability studies are another method of assessing the reliability of a test. Results from these studies produce g-coefficients which quantify the generalizability of the test. Similar to other reliability coefficients,

g-coefficients range from 0.00 to 1.00 with a value of 0.70 representing the threshold for an acceptable level (Price, 2017). An additional benefit of a generalizability study is that it allows researchers to examine the total variability of test scores and determine how much of this variability is accounted for by each potential source of variance. Casper has demonstrated strong generalizability throughout its lifetime, from initial conception to more recent application cycles.

Most recent analyses examined responses from 2,495 US medical school applicants for the 2018-2019 test cycle who had their responses rated twice. The overall Casper g-coefficient was 0.85 while the video-based scenarios had a g-coefficient of 0.79 (8 scenarios) and the word-based scenarios had a g-coefficient of 0.67 (4 scenarios). Notably, the discrepancy in the number of scenarios may have deflated the g-coefficient for the word-based scenarios. Results from a follow-up decision study indicated that if each scenario-type were equal in frequency, the word-based g-coefficient would increase to 0.81. This study also showed that the largest source of variance in Casper scores was the difference in the ability of applicants (22.6%). This indicates that differences in test scores among applicants is largely the result of differences in their abilities. The second largest source of total variance in scores was the difference amongst raters (20.7%) followed by applicant and scenario interaction (9.3%) and the difference in difficulty of scenarios (0.6%). These results are similar to that of the first evaluation of Casper generalizability which occurred with a two-part pilot study (Dore et al., 2009).

The first part of the pilot study examined response-type variations of Casper (audio-recorded responses and typewritten responses) amongst 110 undergraduate applicants to the McMaster University School of Medicine for the 2006-2007 admissions cycle. Both the audio-recorded response version and the typewritten response version of Casper demonstrated acceptable levels of generalizability with g-coefficients of 0.86 and 0.72, respectively (Dore et al., 2009). In an extension of the first study (*n*=167), the entire test (14 scenarios) demonstrated a strong g-coefficient of 0.82 (Dore et al., 2009). When further broken down, the video scenarios (8 scenarios) and the self-descriptive scenarios (6 scenarios) each demonstrated moderate generalizability of 0.75 and 0.69, respectively (Dore et al., 2009).

Generalizability relative to other metrics in the same area of assessment. Comparatively, evidence from MMI generalizability studies indicate that g-coefficients typically range from 0.65 to 0.78 (Burgos et al., 2020; Eva et al., 2004a; Eva et al., 2004b; Sebok et al., 2014) which are very similar to those produced by Casper. However, many MMI generalizability studies suggest that applicant ability accounts for much smaller proportions of the explained variance in scores relative to Casper. A generalizability study from the creators of the MMI, noted that 26.6% of the variance was explained by the applicants (Eva et al., 2004a), however more recent MMI studies indicate much lower levels of explained variance being accounted for by applicants, ranging from 8.7% (Burgos et al., 2020) to 16.3% (Sebok et al., 2014).

Reliability Conclusions

As evidenced by the aforementioned results, it is clear that Casper is a reliable measure of social intelligence and professionalism; a statement which is supported by continuous analyses of reliability across several sources of potential variance. Consistently high and uniform levels of internal-consistency provide evidence that all scenarios of the Casper test work together to measure a single construct. The low SEM values indicate that applicants' Casper scores are reflective of their true scores. High levels of IRR signify that Casper scores are consistent across raters. Results of test-retest reliability and parallel-forms reliability provide evidence that Casper scores remain consistent across time periods and across different variations of the test. Finally, large g-coefficients indicate that Casper is generalizable and that the main source of variance or difference in test scores is due to applicant ability.

In sum, the consistency of Casper scores across different scenarios, raters, time periods, and variations of the test provide support for Casper as a reliable measure of social intelligence and professionalism.

CHAPTER 2: VALIDITY

Chapter Outline

Validity refers to the extent to which a test measures what it is intended to measure (Price, 2017). Validity is not a single test or singular value, but rather a collection of evidence that can be used to support different aspects of validity. Each facet of validity is measured in unique ways (both qualitative and quantitative) and provides unique pieces of information, all of which contribute to the overall validity of the test (Price, 2017). In the case of Casper, validity is supported by the extent to which evidence is available on the test's ability to measure social intelligence and professionalism.

The following chapter is broken down into meaningful sections to discuss evidence as it relates to Casper's test structure, supporting documentation regarding Casper's ability to measure social intelligence and professionalism, ability to predict academic outcomes, as well as a discussion on Casper as an equitable measure across demographic groups.

General overviews of Casper's validity evidence are available below and are followed by subsections which provide more intricate details as well as comparisons to other admissions metrics when applicable.

Part I. This section of the validity chapter provides the reader with information on the test development process and the test structure. Here readers can find information on Casper's content validity, face validity, and structural validity.

- **Content Validity.** Casper content is the result of continuous feedback and iterations from several stakeholders both internal and external to Acuity Insights. Diverse groups of subject matter experts are consulted for each Casper scenario prior to test publication.
- **Face Validity.** Optional surveys are made available to every applicant who completes the Casper test. Survey responses indicate that overall, applicants perceive Casper as a useful tool in allowing them to demonstrate their social intelligence and professionalism skills. Additionally, applicants generally report positive testing experiences.
- **Structural Validity.** Results from a series of exploratory factor analyses (EFAs) and confirmatory factor analyses (CFAs) support Casper as having an overarching factor subtended by two highly correlated sub-factors. These findings mean that although the content of each scenario reflects a broad range of interpersonal and professional characteristics, the critical thinking and social interpretation process required to address each dilemma is similar across all scenarios. Further, these findings suggest that the typed-response

section and video-response section are probing unique means of communicating this construct.

Part II. This section houses information which supports Casper as a valid measure of social intelligence and professionalism. Information is provided on Casper as it relates to other measures of non-technical skills and to measures of technical skills and knowledge. Data is also presented on the relationship between Casper and several construct-irrelevant variables.

- **Convergent Validity.** Casper has demonstrated positive, often significant, correlations with relevant non-technical metrics that evaluate related constructs. In particular, Casper has evidenced significant positive correlations with MMIs, interview performance, a variety of emotional and social competencies, and non-technical subsections of several evaluations.
- **Discriminant Validity.** Opposite to convergent validity, Casper consistently demonstrates negligible, often negative, correlations with unsimilar measures. Casper has demonstrated minimal relationships with GPA (mean r= 0.11) as well as knowledge-based exams (mean r=0.17).
- **Construct-Irrelevant Variables.** Due to the nature of the Casper test, there are several construct-irrelevant variables that may impact scores; however, Casper has evidenced minimal relationships with several of these potential variables. Specifically, there is research to support that Casper scores are not impacted by spelling, grammar, reading level, or test preparation.

Part III. Evidence Showing that Casper is Predictive of Admissions and Academic Outcomes. This section provides information on Casper's ability to predict future outcomes and behaviours. In addition, results from specific programs are presented which showcase a wide variety of statistical analyses and results supportive of Casper's predictive validity, convergent validity, and discriminant validity.

- **Predictive Validity.** Across several studies, Casper has demonstrated an ability to predict for both short-term and long-term behaviours, even up to 6 years later. Specifically, Casper has evidenced a capacity for predicting performance on national licensure exams, scores from in-program OSCE exams, and in-program professional behaviour.
- **Program Specific Outcomes.** Results from a variety of programs across Canada, the United States, Australia, and the United Kingdom are discussed in detail.

Part IV. Research Examining Casper as a Measure that is Equitable Across Demographic Groups. At Acuity Insights, we strive to ensure that the Casper test is as fair and equitable as possible for all applicants. We spend considerable time developing the content of the tests, monitoring the performance of the test so that we can be aware of any situation in which the test may be unfairly disadvantaging certain groups of applicants, and experimenting with new approaches to the test to ensure we are offering the best and most equitable assessment to applicants.

- **Demographic Differences.** Casper is continuously evaluated for demographic group differences throughout each cycle. This involves assessing for group differences across applicant race, gender, socio-economic status, language, age, employment experience, ability status, and rurality. Generally speaking, Casper tends to produce smaller demographic group differences relative to other assessments that are often used in admissions processes.

- **Mitigating Content Bias.** We recognize that systematic inequalities that are rooted deep within our society may contribute to some discrepancies observed in Casper scores, similar to those found in GPA (Whitcomb et al., 2021) and standardized tests like the SAT (Smith & Reeves, 2020) and GRE (Mortaz Hejri et al., 2022). That being said, we are committed to ensuring that we, as an organization, do everything within our control to mitigate these biases as much as possible. First, the content of the Casper test undergoes rigorous evaluation prior to being provided to applicants. Second, we monitor the fairness of the items using measurement invariance and differential item functioning, both of which are used to ensure that the test is working equivalently across all groups of applicants. Finally, the team at Acuity Insights is continuously experimenting with new methods and formats for the test to ensure that we are providing the most fair and equitable assessment to applicants.

Part I. Evidence Related to Casper's Test Structure

Content Validity- *Information on Casper's content development process.*

Validity is an essential requirement for every assessment and implies that the assessment is aligned with targeted competencies or outcomes. Content validity, most pertinent to the Casper test, refers to the extent to which the items or scenarios of a test represent the construct that it is measuring. Haynes et al. (1995) stressed the importance of proper content validity procedures as it directly impacts all inferences that can be drawn from the test. For Casper, each component of the test is subjected to a rigorous and iterative process to ensure all content is valid, relative, and representative of the construct being measured.

How we Define Social Intelligence and Professionalism. As previously discussed, Casper provides a measure of social intelligence and professionalism. This construct is defined as *the ability to effectively reflect upon interpersonal and professional dilemmas and respond to a unique set of questions using critical reasoning and social interpretation.* Social intelligence and professionalism, in this context, are treated as a single construct because the behaviours and characteristics reflected in both are highly intertwined and interdependent. In order to provide a score for this construct, the Casper test probes for and assesses 10 prominent aspects of social intelligence and professionalism: empathy, communication, motivation, resilience, self-awareness, problem-solving, collaboration, ethics, equity, and professionalism.

These 10 aspects were carefully selected based on extensive review of the competencies outlined by professional colleges, associations, and regulatory bodies such as the Association of American Medical Colleges (AAMC), the Canadian Medical Education Directives for Specialists (CanMEDS), and the Australian Institute for Teaching and School Leadership (AITSL). While there are other aspects that may contribute to social intelligence and professionalism, these 10 were determined by Acuity Insights to be the most prominent across professional regulatory bodies. These aspects of social intelligence and professionalism are vital to the entire testing process. Specifically, they work to: (1) *inform* scenario development, (2) *direct* question development, (3) *balance* test blueprints, and (4) *guide* response ratings.

Scenario and Question Development. The 10 aspects of the social intelligence and professionalism construct (collaboration, communication, empathy, equity, ethics, motivation, problem solving, professionalism, resilience, and self-awareness) have been precisely defined and reviewed by several internal and external experts to ensure they are clear and all-encompassing. These definitions

provide a solid base for the development of test scenarios using the following procedures:

1. Content subject matter experts (SMEs) generate scenario proposals.
2. Internal teams, consisting of content experts and educationists at Acuity Insights, develop scenarios based on these proposals and send them to external expert reviewers.
3. Scenario reviewers are external SMEs who are unique from the scenario generators and may come from a variety of professional backgrounds including healthcare, psychometrics, psychology, human resources, etc. To ensure that the scenario content is culturally appropriate, these individuals must reside in the same country for which the content is being designed and reviewed.
4. Reviewers' feedback is incorporated into the scenarios and provided to the internal research team at Acuity Insights for a final review.
5. The scripts for each finalized scenario are written by an external production company and every single script is passed through the same rigorous review process by a group of content experts at Acuity Insights.
6. The scripts are then sent to the production team where the video-based scenarios are filmed with professional actors.
7. The internal team of experts at Acuity Insights then refine the newly produced scenario, add the associated questions, and bank the finalized scenario for future use.

Test Construction Process. The test construction process itself is uniform across all Casper tests, but as previously mentioned, the content of each test is unique. For each test, 9 video-based and 5 word-based scenarios are selected from a bank of scenarios. Each scenario in this bank is tagged with a primary and secondary aspect that the scenario is intended to probe for based on the item content. As an example, a scenario designed to probe for *collaboration* and *professionalism* would be tagged as having collaboration as the primary aspect and professionalism as the secondary aspect. Test builders are instructed to select one scenario from the bank for each of the 10 aspects to ensure each is incorporated into the test. The other four scenarios are chosen at random. It is important to note that these are *intended* themes of the scenario which means that applicants may approach the scenario in a unique way and may display behaviours that correspond to a different aspect. Casper's open-response format encourages unique responses and therefore applicants are not penalized if their response does not directly address the intended aspect.

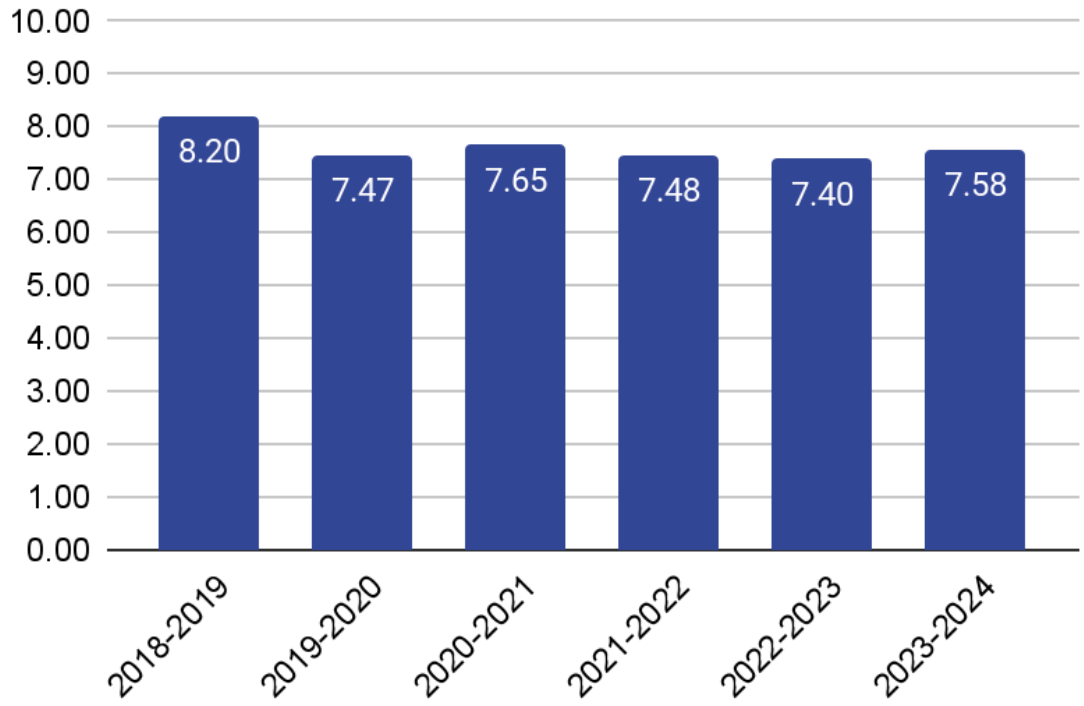
Face Validity Based on Perceptions of Applicants- *Information on How Casper is Perceived by Applicants.*

Face validity measures the extent to which the items or scenarios of a test appear to measure the intended outcome of said test. Face validity is an invaluable metric for assessing the perceived usefulness of a test. One appropriate way of gauging the face validity of the Casper test is through examining the perceptions of applicants. Using an optional feedback survey presented to applicants after the completion of their test, Acuity Insights collects several pieces of information regarding applicants' perceptions of the test.

The first question in the exit survey, rated on a scale from 1-10, asks applicants "overall how do you rate Casper?". This question was introduced during the 2018-2019 application cycle and, in receiving feedback from 591,572 applicants, has shown that applicants consistently rate Casper favorably with an average rating of 7.63 out of 10 across the most recent application cycles (Figure 4).

Figure 4

Average Applicant Rating For The Question "Overall How Do You Rate Casper? (1=Extremely Negatively, 10=Extremely Positively)



In addition to this question about applicants' general rating of Casper, applicants are also asked a series of more specific questions regarding their perceptions of and experience with the Casper test. While the exact wording (and inclusion) of these

questions may vary slightly over the years, the general sentiment remains the same. Table 3 details questions on the applicant survey that relate to applicants' perception of the Casper test as it relates to ability to demonstrate one's strengths (*n*=562,571) and effectiveness (*n*=550,934). Please note that unlike the first question in the survey which is on a scale from 1-10 (Figure 4), the following questions, as illustrated in Table 3, are on a scale of 1 to 7. Results from these optional surveys provide evidence that applicants, on average, perceive Casper to be effective in evaluating social intelligence and professionalism skills and a test which provides a means for showcasing their strengths in this particular area.

Table 3

Mean Scores For Applicant Survey Questions Rated On A 7-Point Scale

	2018-2019	2019-2020	2020-2021	2021-2022	2022-2023	2023-2024
How well do you think Casper allowed you to demonstrate your strengths relative to other applicants?	4.65	4.28	4.49	4.40	4.46	4.30
How effective do you think Casper is as a tool for evaluating one's personal and professional characteristics for the profession?	5.07	4.36	4.50	4.40	4.45	4.28

Note. A score of 1 corresponds to low levels of ability to demonstrate strengths or effectiveness, while a score of 7 corresponds to high levels of each.

Face Validity Across Languages. Similar results of face validity have been demonstrated in an independent study across the two official languages that Casper is offered in (English and French). At the University of Ottawa Medical School for the 2017-2018 application cycle, French applicants (*n*=315) reported slightly higher, albeit significant, face validity scores than English applicants (*n*=3,802; difference = 0.37, *p*<.001).

Face Validity Relative to Other Metrics. Based on evaluations of other measures of personal or soft skills, it appears that Casper may be a preferred choice. In a small study, applicants (*n*=77) and faculty members (*n*=17) at the Stanford University general surgery residency program were asked about their perceptions of Casper relative to traditional assessments of personal or soft skills (*e.g.*, letters of recommendation and personal statement letters; Shipper et al., 2017). Although applicants preferred traditional methods, the faculty members believed that Casper was more accurate (*p*=.002) than traditional evaluation metrics for these skills (Shipper et al., 2017).

Structural Validity- *Evidence that Casper is a two-dimensional correlated test.*

Structural (or factorial) validity is a component of construct validity, in that the structural validity of a test refers to the degree to which the scores from the test appropriately reflect the true dimensionality of the construct that the test is intended to measure (Cronbach & Meehl, 1955). There are several methods that can be used to evaluate the structural validity of a test; however, most research on the Casper test to date has focused on factor analysis.

Factor Analysis (FA). FA is a statistical technique that evaluates the inter-correlations of a set of items (*e.g.,* test scenarios) to form a parsimonious rendering of the test's structure (Cronbach & Meehl, 1955; Price, 2017). FA tells us what items of a test cluster together and the extent to which they belong together (Price, 2017). The underlying theory of FA is that test items are correlated with one another because of a common unobserved influence; this unobserved influence is referred to as the latent variable (Price, 2017). Latent variables cannot be directly measured or observed and thus must be inferred from other observable or measurable variables. For example, you cannot measure intelligence on a scale like you would weight, but you can observe variables that infer intelligence such as one's average math score, the number of words in their vocabulary, etc.

Exploratory Factor Analysis (EFA). As the name suggests, EFA is used early on in test construction to determine how a set of items relate to (or define) underlying constructs. EFA allows the structure within the data to reveal itself, the results of which are used to develop or refine a theory of the test's structure. For Casper, a series of EFAs have been conducted for several test instances across each application cycle.

For all EFAs conducted on Casper, a maximum likelihood extraction method is used. Current literature suggests that if data are relatively normally distributed, then this method allows "for the computation of a wide range of indexes of the goodness of fit of the model…[and] permits statistical significance testing of factor loadings" (Fabrigar et al., 1999). To determine how many factors should be retained, we currently rely on results from parallel analysis, which has been suggested by some to be one of the most accurate methods for determining factor retention (Hayton et al., 2004). Parallel analysis requires several random datasets to be generated (*i.e.,* a minimum of 50) that are equal to the original dataset in terms of the number of variables and cases; thus, making them 'parallel' to that of the original (Howard, 2016). Factors are retained if the magnitude of the eigenvalues produced in the original data are greater than the average of those produced by the randomly generated datasets (Howard, 2016). The underlying theory of parallel analysis is that the eigenvalues derived from random datasets can only be considered statistical artifacts, thus when the original dataset produces greater eigenvalues, they provide information beyond that which is considered a statistical artifact (Howard, 2016). Historically, parallel analysis results have shown that a one-factor structure is the best

fit for Casper. However, with the integration of the video-response format, a two-factor structure has emerged as the best fit.

Confirmatory Factor Analysis (CFA). Following theory development, researchers can conduct a CFA to confirm that the test's structure matches that which was proposed theoretically. For the most recent application cycle, CFA analyses were conducted on 22 unique Casper test instances across a variety of geographies and program types.

At a high-level, the process of conducting a CFA boils down to imposing a model onto a dataset to evaluate how well the model fits the data. Since results from EFAs suggest that Casper is a two-dimensional test, a series of two-factor models were imposed on the data: two-factor orthogonal model, two-factor oblique model, and a bifactor model. The degree to which these models fit the data was evaluated using the following fit indices: (i) comparative fit index (CFI), (ii) root-mean-square error of approximation (RMSEA), and (iii) standardized root-mean-square residual (SRMR).

Results indicate that a two-factor oblique model was the best fit. All fit statistics were supported as "good" fit (Hu & Bentler, 1999) (*i.e.,* CFI ≥ .95, RMSEA ≤ .05, SRMR ≤ .08) and are available in Table 4. Together, these results indicate that although the typed-response section and the video-response section act as two independent factors, these factors are highly correlated and thus share a strong relationship with one another. This suggests that the two response formats are working to measure the same construct through two unique modes of communication.

Table 4

CFA Fit Statistics For a Two-Factor Oblique Model

Country	Program	CFI	RMSEA	SRMR
US	Medicine	0.995	0.015	0.022
Canada	Health Sciences	0.974	0.033	0.024
Canada	Health Sciences	0.990	0.025	0.021
Australia	Health Sciences	0.991	0.017	0.022
Australia	Health Sciences	0.973	0.030	0.031
Australia	Health Sciences	0.981	0.025	0.030
US	Health Sciences	1.000	0.000	0.019
Canada	Allied Health (French)	0.986	0.026	0.028
Canada	Graduate Health Sciences	1.000	0.000	0.016
Canada	Teacher's Education	0.992	0.017	0.025
US	Medicine	1.000	0.000	0.016
Australia	Teacher's Education	0.984	0.028	0.037
Australia	Teacher's Education	0.956	0.041	0.041
Canada	Health Sciences	0.986	0.027	0.022
US	Medicine	0.997	0.012	0.018
US	Health Sciences	0.993	0.020	0.022
US	Medicine	0.981	0.032	0.021
Canada	Health Sciences	0.993	0.018	0.022
Canada	Allied Health (French)	0.974	0.025	0.023

Part I Summary

This section of the manual provided details on how the content of each Casper test is developed, information on applicants' perceptions of the Casper test, and evidence to support that Casper is measuring a single construct: social intelligence and professionalism with two unique communication formats. The meticulous process of content development ensures that input from stakeholders and SMEs is incorporated into every stage of scenario development and test construction. This scrupulous procedure which every item is required to pass through prior to being incorporated into a test is a key factor in the high reliability estimates that are evident in the Casper test. Overall, applicants report highly positive experiences with their Casper test and generally perceive the Casper test to be effective in evaluating social

intelligence and professionalism skills and a quality method for demonstrating their strengths in this particular area.

Part II. Evidence Supporting Casper as a Measure of Social Intelligence and Professionalism

Convergent Validity - *Evidence for Casper as an effective measure of non-technical soft skills.*

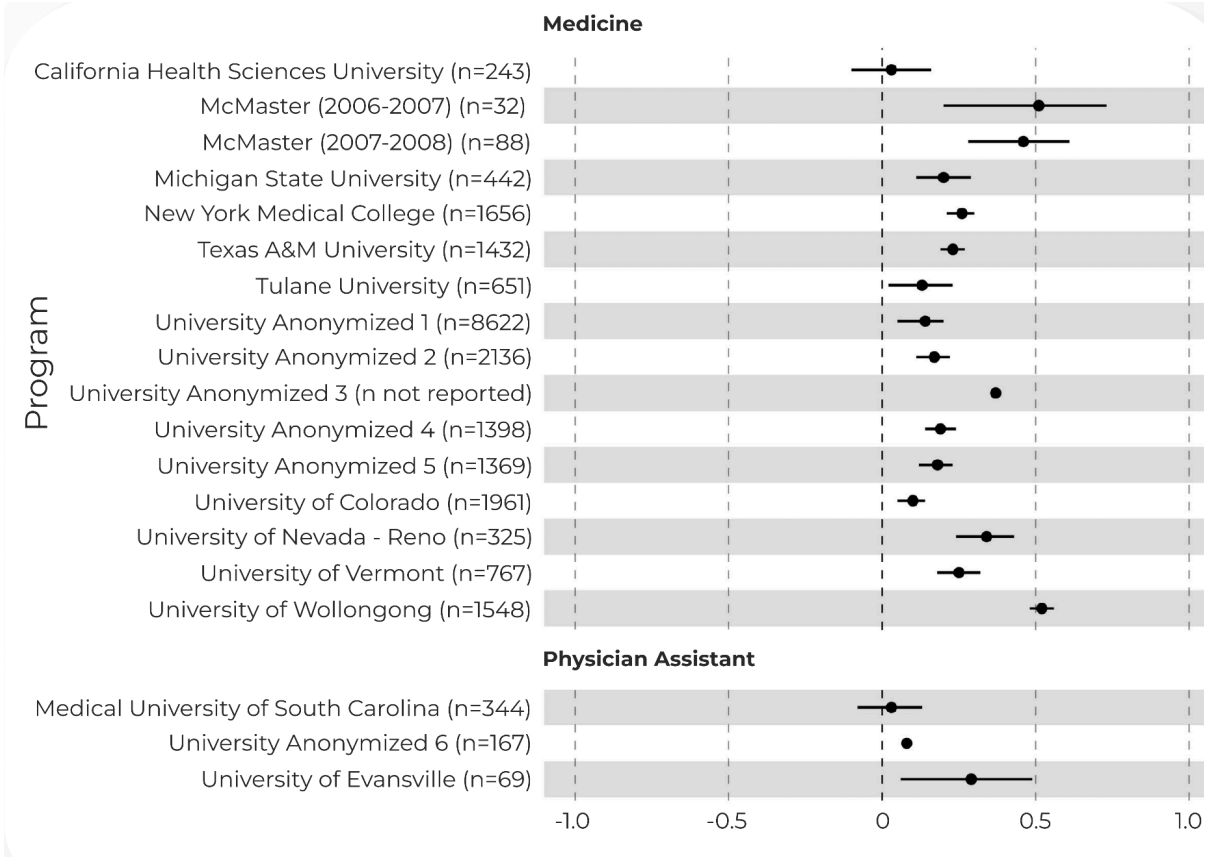
Convergent validity evaluates the extent to which two instruments that measure the same or similar construct are related (Carlson & Herdman, 2012). The relationship between similar measures is evaluated using correlation coefficients that range from r=0.00 to r=1.00 with higher values indicative of a stronger relationship. As a general guideline, in this context, a correlation coefficient of r=0.10 is considered a small relationship while correlations of r=0.30 and r=0.50 are the thresholds for moderate and large relationships, respectively (Cohen, 1992).

As Casper is a measure of social intelligence and professionalism, it is expected to demonstrate some level of association with similar measures such as the MMI or interview. Notably, the make-up of MMIs and interviews varies across programs which inevitably impacts their relationship with Casper.

As can be seen in Figure 5, Casper scores have demonstrated moderate correlations with MMI or interview from 18 unique programs and over 23,000 applicants from Canada, the United States, and Australia. Of the 18 programs in this figure, 83% (n=15) evidenced a significant positive correlation between Casper and interview or MMI performance. This figure displays correlations with total MMI scores, however significant correlations (using Spearman's r [r_s])have also been demonstrated with MMI subdomain scores of compassion (r_s=0.19, p<.001), sociability (r_s=0.12, p=.010), calm disposition (r_s=0.17, p<.001), and morality (r_s=0.11, p=.020).

Figure 5

Correlations Between Casper Scores And MMI or Interview Performance



A 2022 study of Casper's convergent validity observed a relationship between Casper and a broad range of emotional and social competencies (ESCs; Henning et al., 2023). For this study, a group of applicants who completed the Casper test as part of their admissions process voluntarily offered to also complete a self report measure of ESCs (the Multidimensional Inventory of Personal Intelligence (MIPI); Parker, 2022). Participants were grouped into two categories based on their Casper score: top-performers (top 15%; *n*=46) and bottom-performers (bottom 15%; *n*=46).

A mixed model ANOVA was conducted for each of the 14 scales/subscales, using the Casper groups as the between-subjects variable and scales on the MIPI as the within-subjects dependent variables. As evidenced in Table 5, the Casper top-performers (T) scored significantly higher than Casper bottom-performers (B) for 11 of the 14 scales. The magnitude of the differences between these two groups are considered to be moderate to large in size (Cohen, 1992; Cohen's d ranging from 0.44 to 0.84). For the other three subscales, the difference between groups was non-significant (NS). These results support a strong relationship between Casper and other measures of similar constructs.

Table 5

Group Differences On MIPI Between Top And Bottom Performing Casper Test Takers

Result	Cohen's *d*	MIPI Scale or Subscale	Brief Definition
T > B	0.73	Total Emotional Intelligence	A set of abilities, behaviours, and dispositions that allow a person to process emotional information, place that information in context, and use it to make effective decisions.
T > B	0.44	Emotional Understanding	EQ Subscale: Your ability to recognize and label your emotions as you feel them and recognize those feelings in other people.
T > B	0.81	Introspectiveness	EQ Subscale: Your ability to think about, prioritize, and appreciate both your feelings and the feelings of others.
NS	NA	Attentiveness	EQ Subscale: Your ability to maintain focus on other people and details in your environment even amid distraction.
T > B	0.75	Emotional Communication	EQ Subscale: Your ability to meaningfully describe your feelings to others and to converse about others' feelings.
T > B	0.68	Total Social Intelligence	A set of abilities, behaviours, and dispositions that help a person to navigate social situations, subject themselves to evaluation by others and make effective decisions when under social pressure.
T > B	0.84	Social Integration	SQ Subscale: Your ability to create meaningful and effective relationships with others across a variety of contexts.
NS	NA	Performance Readiness	SQ Subscale: Your ability to subject yourself to others' evaluation in order to achieve a specific outcome.
T > B	0.59	Social Agency	SQ Subscale: Your ability to make effective decisions and maintain a sense of personal control during social situations.
T > B	0.61	Total Motivational Intelligence	A set of abilities, behaviours, and dispositions that allow a person to set goals and objectives and pursue them with an appropriate level of intensity.
NS	NA	Motivational Self-Efficacy	MQ Subscale: Your ability to influence others to behave in ways that align with your goals.
T > B	0.66	Motivational Influence	MQ Subscale: Your ability to create long-term goals and encourage yourself to pursue them.

T > B	0.62	Perseverance	MQ Subscale: Your ability to expend energy and resources to pursue goals even through difficult times.
T > B	0.74	Total Personal Intelligence	Combined score for EQ, SQ, and MQ.

On several occasions, Casper scores have been particularly successful at identifying applicants who may potentially demonstrate concerning or problematic behaviour. At the University of Wollongong, 99% of applicants whose Casper score was -1 or below, also demonstrated responses or behaviors during their MMI that warranted a "red flag" from interviewers (Parker-Newlyn et al., 2019). This finding is similar to internal analyses which have indicated that applicants who receive "flags" during their Casper test due to concerning behaviour (*e.g.,* being rude to staff, suspected of cheating, etc.) tend to perform slightly, yet significantly, lower than applicants who are not flagged (effect size ranging from d=0.28 to d=0.44).

In addition to evidencing convergent validity with MMIs, Casper scores have also demonstrated positive relationships with measures of psychosocial and emotional abilities (r=0.10, p<.05; Yingling et al., 2018) and moral reasoning (r=0.09, p<.01; Yingling et al., 2018).

Discriminant Validity - *Evidence that Casper provides unique information separate from technical admissions metrics.*

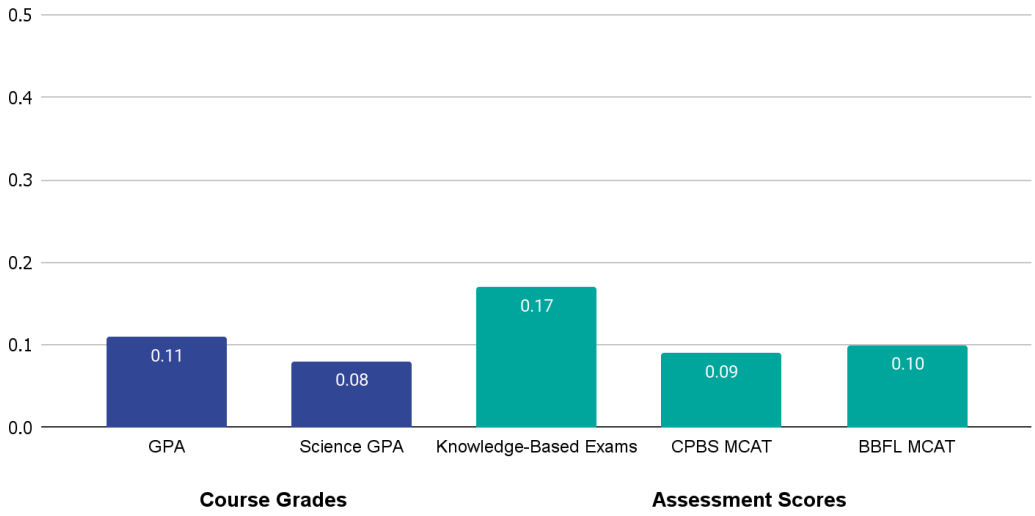
Discriminant validity assesses the extent to which two measures that should not be related, are in fact, not related (Campbell & Fiske, 1959). Since Casper is a measure of social intelligence and professionalism, it should not exhibit relationships with measures of technical abilities. Further, since Casper is used in the admissions process, evaluations of discriminant validity are focused on the relationship, or lack thereof, with measures that are also used in the admission process such as grade point average (GPA) and knowledge-based exams. Earliest analyses of Casper showed weak or negative correlations with measures of technical abilities, and these strong discriminant validity results have continued to persist over time.

In examining the relationship between Casper and GPA across a variety of programs, we were able to gather information from 142,972 applicants (Figure 6). On average, the correlation between Casper and overall GPA was small (r=0.11) and even smaller when examining Casper's relationship with GPA calculated using only science courses (r=0.08). Similarly, in examining data from 137,347 applicants, Casper demonstrated minimal relationship (mean r=0.17) to knowledge-based exams (MCAT, DAT, and GRE; Figure 6). Again, this relationship was reduced further when examining specific subsections of these knowledge-based exams that target Chemical and Physical Foundations of Biological Systems (CPBS; r=0.09) or Biological and Biochemical Foundations of Living Systems (BBFL; r=0.10). Recall that generally

in this context, correlations around *r*=0.10 are considered small and correlations of *r*=0.30 are considered moderate (Cohen, 1992). With these thresholds in mind, it is clear that Casper consistently demonstrates minimal correlations with measures of technical ability.

Figure 6

Correlations Between Casper Scores And GPA And Knowledge-Based Exam Scores



Discriminant validity relative to other metrics in the same area of assessment. These correlations are comparable to those observed between MMIs and technical assessments. MMIs have demonstrated similarly low correlations with GPA (*r*=0.06; Kulasegaram et al., 2010; *r*=0.006; Eva et al., 2012), GAMSAT (*r*= -0.12 to *r*=0.20; Roberts et al., 2008), and the MCAT (*r*=0.10; Kulasegaram et al., 2010).

Construct-Irrelevant Variables - *Evidence that Casper is not influenced by irrelevant variables.*

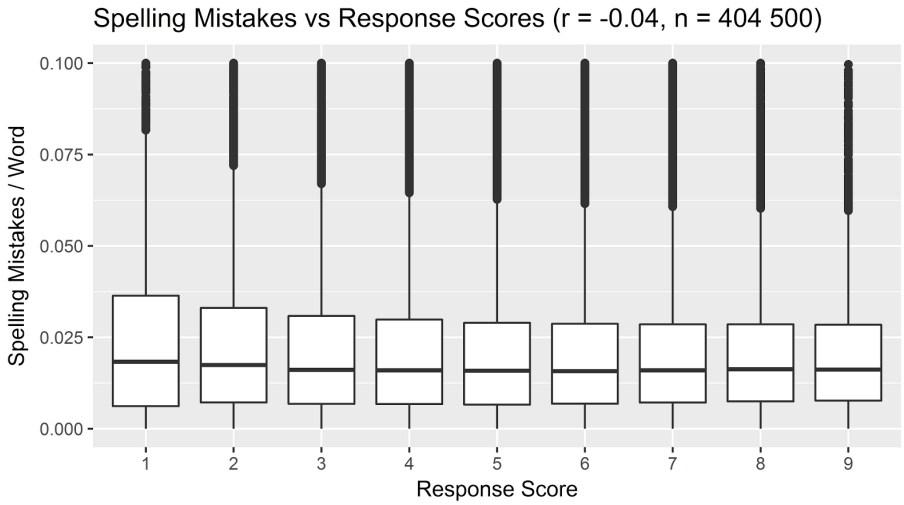
In addition to the construct-irrelevant variance that is discussed in the Demographic Differences section of this report, there are several other potential variables that may inappropriately impact Casper test scores. Since Casper is designed as a measure of social intelligence and professionalism, it is essential that construct-irrelevant factors such as typing speed, spelling ability, or reading level do not artificially inflate or deflate applicants' scores. Any concerns that a confounding variable may be having an undue influence on test scores is addressed by the research team at Acuity Insights and analyzed to determine if changes to the test are required.

Casper and Spelling. The time limit on the Casper test leaves little room for response revisions, however both applicants and raters are informed that spelling is not to be considered during the test scoring process. Data collected from just over 400,000 applicants (n=404,500) who applied to an assortment of health science

programs for the 2016 and 2017 application cycles verified that this is in fact true, as rate of spelling errors produced a negligible, near-zero correlation with Casper scores (r= -0.04). Additionally, as evidenced in Figure 7, it is clear that the rate of spelling errors remains constant across all response scores providing additional evidence that Casper scores are not impacted by spelling errors.

Figure 7

Relationship Between Casper Scores For Each Response Score And Rate Of Spelling Errors



Casper and Word Count. The open-ended response format of the Casper test allows for fluctuations in response length across applicants. Although longer responses could distinguish applicants who have written comprehensive responses from applicants with surface-level responses, there is a possibility that longer responses are instead the result of an applicant with fast typing abilities.

To gain an understanding of this relationship, the correlation between the number of words per response and Casper scores are being continually examined. Table 6 is a collection of data from 861,329 applicants and displays the average correlation coefficient for each application cycle which ranges between r=0.50 and r=0.68. Given these moderately strong correlations between response length and Casper scores, it is likely that longer answers may be more nuanced and reflect a more in depth understanding and resolution to the situational dilemma being presented. However, part of the observed difference could be due to construct-irrelevant factors such as applicant typing speed, english proficiency level, etc. Future research will aim to fully evaluate the contribution of the underlying factors that connect Casper scores to response length.

Table 6

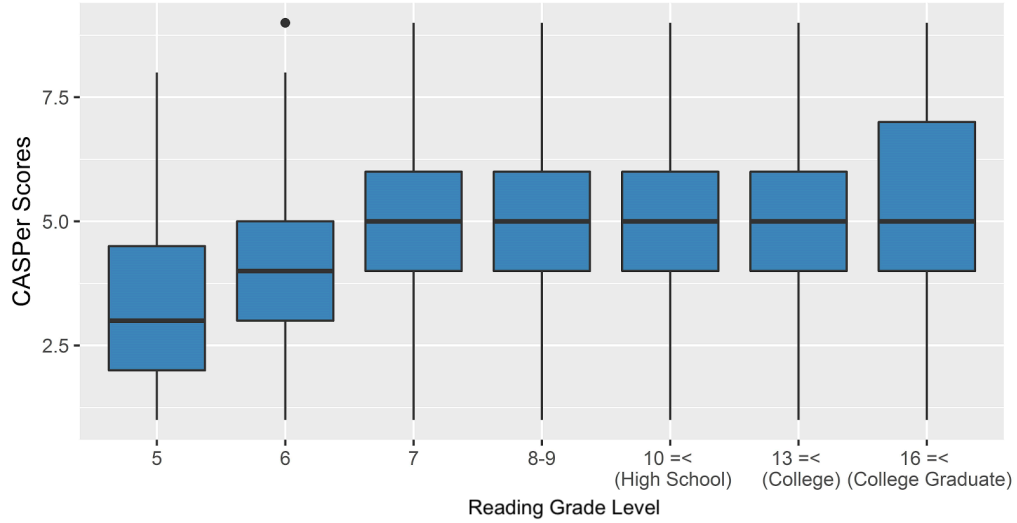
Average Correlation Between Word Count And Casper Scores.

Application Cycle	Number of Test Instances	Number of Applicants	Average Correlation	Minimum Correlation	Maximum Correlation
2016-2017	68	30,698	0.50	0.12	0.79
2017-2018	87	56,389	0.51	0.24	0.74
2018-2019	113	88,614	0.51	0.21	0.72
2019-2020	124	106,005	0.58	0.12	0.76
2020-2021	178	156,844	0.61	0.35	0.75
2021-2022	184	151,198	0.62	0.34	0.74
2022-2023	169	140,062	0.64	0.40	0.74
2023-2024	155	133,137	0.68	0.35	0.79

Casper and Reading Level. Although most of the scenario prompts on the Casper test are in video format, applicants are required to read the corresponding questions for the videos in addition to a few written scenarios. To ensure reading ability did not impact Casper scores, the Flesch-Kincaid Grade Level was calculated for 9,690 applicants to various health sciences programs in the 2016-2017 application cycle to attain an estimate of their literacy level. As can be seen in Figure 8, the correlation between Casper scores and reading level was small (r=0.10) indicating that there are no differences in Casper scores for literacy skills at and above a typical grade 7 level.

Figure 8

Relationship Between Casper Scores And Estimates Of Grade-Level Literacy Skills



Casper and Test Preparation & Practice. Differences in applicants' scores on their first and second attempt at writing Casper may also be attributed to practice effects. That is, test score differences are the product of the applicant becoming more familiar with the test format or test content, or from direct coaching on how to write the test. To examine the impact practice may have on Casper scores, two analyses of various preparatory methods have been examined.

In a post-test follow-up survey, applicants were asked about the methods they used to prepare for their Casper test. Multiple regression analyses of data collected from 233,627 applicants from 2020-2022 (Table 7) showed that all test preparation methods had a significant, yet small in magnitude impact on test performance, however the direction of this impact (positive or negative) varied for each method. Notably, applicants who took a 3rd party prep course performed significantly worse (B = -0.06, $p < .001$), while applicants who did not prepare at all performed the worst (B = -0.28, $p < .001$). The highest performing applicants were those who used the free resources available on the Acuity Insights website (B = 0.16, $p < .001$) and those who took the full practice test (B = 0.14, $p < .001$). Overall, model fit was low (adjusted R-squared = 0.03), suggesting that Casper test preparation will have some positive impact on test performance, but will not dramatically change an individual's score.

Table 7

Results From A Multiple Regression Analysis Of Applicant Preparatory Methods (N=233,627)

Preparation Method	Estimate	SE	*t*	*p*
(Intercept / Did not prepare for the Casper test)	-.28	.004	-66.5	< .001
Reviewed the applicant resources on the Take Casper website (e.g. FAQs, blogs, webinar)	.16	.005	35.6	< .001
Completed the three-section practice test on the Casper website	.04	.004	9.3	< .001
Completed the 12-section practice test in your Casper account	.14	.004	33.8	< .001
Participated in a third-party Casper test preparation course	-.06	.007	-8.1	< .001
Studied potential Casper questions based on the competencies	.12	.005	25.0	< .001
Rehearsed responses with technology	.13	.005	26.4	<.001
Rehearsed responses without technology	.07	.005	12.9	< .001

These results are remarkably similar to what was observed in a previous analysis in which multiple regression analyses of data collected from 16,642 applicants from 2019 showed that most preparatory methods had a small positive effect on Casper scores. Using a third-party Casper preparation course again had a negative effect on Casper scores (B = -0.10, p = .007) as did not preparing at all for the test (B = -0.34, p<.001). The adjusted R^2 value (0.04) of this model indicated that these preparatory methods account for only 4% of the variance in Casper test scores.

Taken together, these results suggest that while practicing for Casper will not have a dramatic impact on scores, there is a benefit to ensuring that applicants become familiar with the test prior to the official assessment. Familiarity with the test format and test process will help to avoid the applicants receiving lower scores simply due to the fact that they are being presented with a novel task for the first time.

Casper and AI-Supported Cheating. The launch of generative AI tools (*i.e.,* ChatGPT) comes with serious concerns regarding honesty in the testing process. Like other assessments, these concerns are also applicable to the Casper test. To determine if and how the use of ChatGPT may impact Casper scores, we partnered with researchers from Saint Mary's University and conducted a multi-phased study to explore the impact of ChatGPT on Casper scores.

In Phase 1, we analyzed a large dataset from 107,805 applicants, comprising approximately 1 million responses to identify if applicants' scores had changed since the release of ChatGPT. When examining the results, we found that the number of months since ChatGPT release was negatively associated with Casper performance ($b = -0.02$, $SE = 0.00$, $p < .001$). However, it is difficult to determine if the release of ChatGPT itself was responsible for this negligible claim or if there were other influencing factors. In order to gain more insights into ChatGPT's potential impact on Casper, we conducted an experimental study (Phase 2).

Building upon Phase 1 findings, Phase 2 (experimental) sought to deepen our understanding by simulating real-world conditions. A total of 138 participants were immersed in a mock Casper test in which they were randomly assigned to one of the three conditions: (1) no outside help (honest answers), (2) help using the internet (no AI), and (3) use of ChatGPT only. Results from phase 2 indicated that participants in the ChatGPT condition achieved only slightly higher scores (0.33 points) than those who did not use ChatGPT. Interestingly, participants in the ChatGPT condition reported that the restrictions embedded within the Casper test (*e.g.,* time limit, disabled copy and paste functionality, etc.) made it extremely difficult to use ChatGPT within the testing window, further deterring them from using it.

Overall, evidence shows that the availability of ChatGPT has a negligible impact on Casper scores. The Casper design and functionalities seem to effectively limit any potential impact ChatGPT may have on test integrity and response authenticity.

Casper Methodology to Avoid Gaming the Assessment. Casper's online platform provides applicants with the flexibility to take the test in any location with internet access and increases accessibility for applicants who may have difficulty travelling to physical testing locations. However, with increased-accessibility through online proctoring comes the increased possibility of applicants attempting to game the assessment. Three major considerations are in place to mitigate the chances of this occurring.

1. ***Unique Content and Testing Systems.*** Every Casper test is made up of unique sets of scenarios and questions. This process ensures that applicants do not have access to scenarios or associated questions prior to the assessment. The system which applicants use to take the Casper test is also fitted with several safeguards to prevent any attempts to manipulate the testing process. Specifically, applicants are not able to manipulate the videos or timers in any way (*e.g.,* rewind, pause, etc.). In the unlikely event that the system crashes during the test (either due to applicants' attempts to manipulate the system or due to external factors such as a power outage), the videos and timers will restart at 2 seconds prior to the time of the crash. Keyboard shortcuts that allow applicants to copy or paste responses are also disabled within the

Casper testing system to ensure that applicants are only providing real-time answers as opposed to pre-constructed responses.

2. ***Time Limit.*** The time limit for each scenario reduces any opportunity an applicant may have to utilize external resources when crafting a response. The time limit also reduces the opportunity for applicants to consult with other individuals during the assessment; a notion which has been supported by internal analysis. To examine if Casper scores could be influenced through a collaborative approach to response writing, 52 participants were recruited to either write the test independently (*n*=18) or with a partner (*n*=34; Dore et al., 2016). Results indicated that there was no significant difference in mean Casper scores between those who wrote the test independently (5.9) and those who wrote in pairs (6.2). In a follow-up survey, 71% of participants noted that they would prefer to write the test independently. Together, these results suggest that with the time limit, there is virtually no qualitative or quantitative support for having assistance during the writing process.

3. ***Proctoring.*** Finally, applicants are required to write their test in front of their computer's webcam and all tests are proctored via a combination of artificial intelligence and human proctors. These practices ensure that the test-taker matches their official ID and that no external materials are being used during the test writing process. Proctors have the ability to temporarily or permanently stop an applicant's test if suspicious behaviour is detected.

Part II Summary

Casper's correlations with the MMI, interview scores, and other non-technical measures supports Casper as a measure of its intended construct. At the same time, the correlations are not so high as to indicate that Casper is redundant with these other metrics. Instead, these mid-range correlations suggest that Casper is providing unique information that may not be attained through traditional admissions metrics. Additionally, Casper consistently displays minimal correlations with measures of technical abilities across several programs and multiple countries. These findings indicate that Casper is not measuring the same underlying construct(s) as technical metrics such as MCAT and GPA. Finally, there is evidence to support the notion that Casper scores are not impacted by spelling, grammar, reading level, test preparation, or the use of generative AI tools used while taking the test. Taken together, this set of evidence indicates that Casper is an effective measure of non-technical soft skills and is not influenced by numerous irrelevant variables.

Part III. Evidence Showing that Casper is Predictive of Admissions and Academic Outcomes

Predictive Validity - *Evidence that Casper is able to predict future outcomes.*

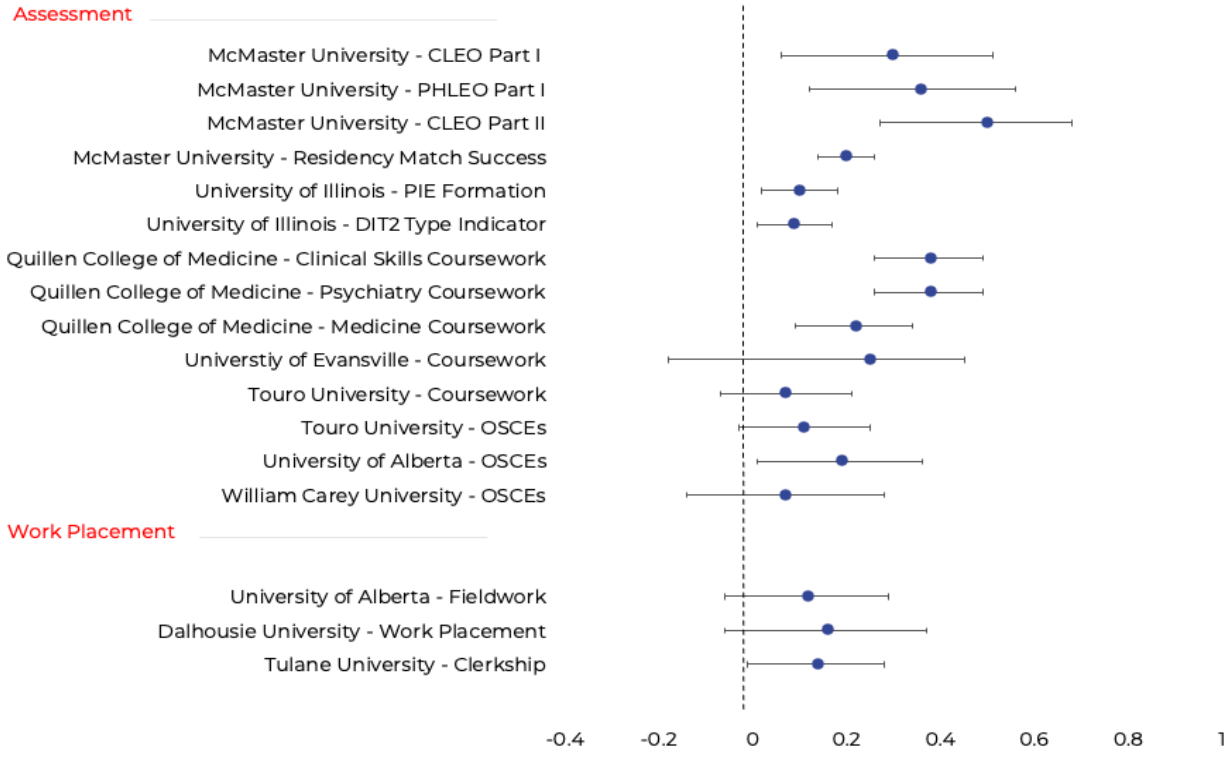
Predictive validity refers to the extent to which scores from a particular measure evidence a relationship with a future outcome (Frey, 2018). The predictive validity of a test can be measured using correlation coefficients or by comparing scores of individuals who exhibited a certain behaviour to those who did not. Casper has demonstrated an ability to predict future personal and professional characteristics and behaviours across several metrics.

Figure 9 provides a visual overview of correlations between Casper scores and in-program performance measures. Highlighting some of the findings, Casper produced moderately strong correlations with scores from two non-technical subsections of the Medical Council of Canada Qualifying Examination (MCCQE): 1) cultural-communication, legal, ethical, and organizational aspects of the practice of medicine Part I (CLEO; r=0.30), CLEO Part II (r=0.50), and 2) population health and ethical, legal, and organizational aspects of medicine (PHELO; r=0.36; Dore et al., 2017). Part I of the MCCQE is administered at the end of medical school and Part II is administered during the second year of residency, demonstrating Casper's ability to predict national licensing exam scores 3-6 years later.

This figure also hosts evidence of Casper's relationship with scores from an in-program objective structured clinical exam (OSCE) from the University of Alberta (r=0.19). There is also evidence pointing to associations between Casper and work-placement indicators of professionalism at Dalhousie University Rowe School of Business. During this pilot, we saw a significant relationship between Casper and work placement Employee Evaluation scores (r=0.16).

Figure 9

Correlations Between Casper Scores And In-Program Performance Measures



Predictive validity relative to other metrics in the same area of assessment. The strength of the relationship observed between Casper scores and future non-technical exam scores is equivalent to the strength of the relationship often observed between technical-knowledge admissions metrics and future technical-knowledge exam scores. The MCAT for example has demonstrated moderate correlations with national licensing exam scores (ranging from *r*=0.31 to *r*=0.60; Association of American Medical Colleges, 2020; Gauer et al., 2016) and clerkship exam scores (*r*=0.52; Association of American Medical Colleges, 2020). Similarly, GPA has demonstrated moderately strong correlations with national licensing exam performance (*r*=0.49) and clerkship exam scores (*r*=0.46; Association of American Medical Colleges, 2020).

Program Specific Outcomes- *Evidence that Casper relates to and is predictive of academic outcomes across various programs and countries.*

The research team at Acuity Insights also partners with academic programs to conduct collaborative research analyses to evaluate Casper within various institutions. These partnerships provide rich data on how Casper relates to other admissions metrics *and* in-program measures of success.

This section of the chapter provides detailed information on how Casper has benefited specific programs throughout Canada, the United States, Australia, and the United Kingdom. The analyses covered in this section vary greatly as each academic institution has unique research questions that are most salient for them. Thus, this section is not only disaggregated by country, but also by individual program. When possible, specific University and program names are disclosed, however, some are anonymized as research is currently being conducted. The anonymous schools are described in terms of their size, program type, and geographical location.

CANADA

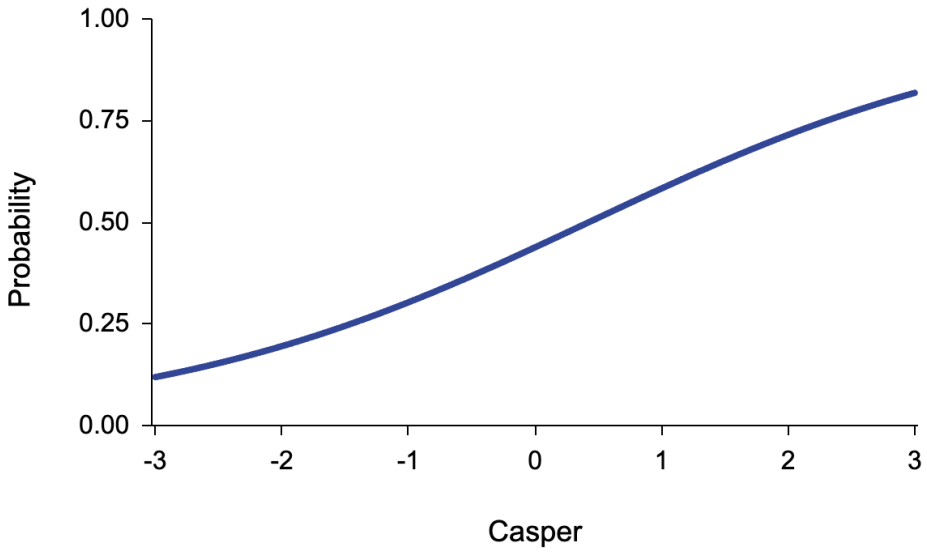
Dalhousie University - *Casper's relationship with employee evaluation scores in eight-month corporate residency (Validation in Business Education).*

The Rowe School of Business at Dalhousie University is an innovative school committed to developing strong business leaders who value quality and integrity (Dalhousie University, 2023). Dalhousie University recognizes that the future of business is a diverse workforce made up of people who care about societal impact and that this requires shifting how business students are selected; looking beyond academic achievement to values, ethics, social intelligence, and professionalism.

With this goal in mind, the Rowe School of Business piloted the use of Casper within their Corporate Residency MBA program. This 2021 pilot study found that higher Casper scores were associated with students being more likely to demonstrate high Employee Evaluation scores during an eight-month corporate residency (Figure 10). More specifically, for every 1-unit increase in Casper score, the odds of receiving a high Employee Evaluation score increased by 1.79 times (OR=1.79, 95%CI[1.10,3.09]).

Figure 10

Casper's Relationship With In-Program Work Placement Performance



Applicant Acceptability. In addition to this validity evidence, Casper also proved to have high levels of acceptability among students who were admitted to the MBA program. A large majority (81%) of students felt that it was important that the program values getting an objective measure of their personal characteristics and 84% felt that Casper allowed them to demonstrate their strengths moderately to extremely well. Additionally, on average, applicants felt that the Casper test was no more or less difficult relative to other exams and that, if required, it would not make them any more or less likely to apply to a specific program.

University of Alberta - *Casper's relations with OSCE scores and GPA*.

The University of Alberta houses a competitive Occupational Therapy (OT) program which admits approximately 120 students each year (University of Alberta, 2021). As an occupation centered on patient interactions, the University of Alberta has chosen to use Casper in their admissions process as a measure of applicants' social intelligence and professionalism.

A 2019 study of the University of Alberta's OT program showed that Casper scores had a stronger relationship with the Objective Structured Clinical Exams (OSCEs; r=0.21, p<.05) compared to GPA (r=0.14, p<.05), evidencing particularly strong relationships with non-technical sub-sections. Specifically, Casper scores were significantly (p<.05) correlated with measures of communication (r=0.21), performance management (r=0.19), and professional interactions and responsibility (r=0.18).

On average, when Casper was used in conjunction with GPA, the ability to predict students' performance on competency-based fieldwork evaluations improved by 4% over and above the predictive ability when only GPA was used.

University of Saskatchewan - *Casper's relations with remediation issues.*

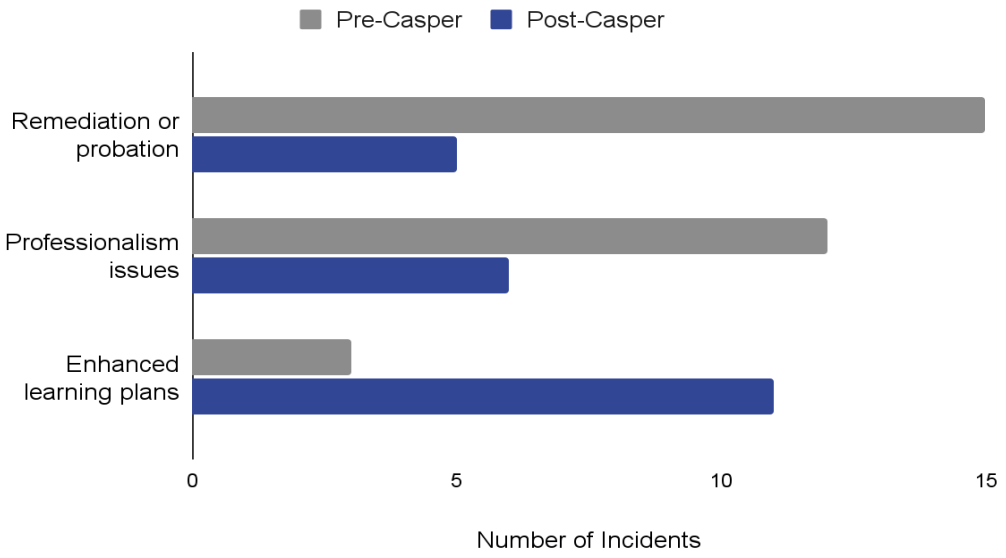
A retrospective study (Figure 11) compared the number of residents who required *formal* learning interventions (*i.e.,* remediation or probation) between residents who were selected prior to implementation of Casper and those who were selected after. The program noted a 67% reduction in the number of residents who required *formal* learning interventions (from 15 to 5). The number of professionalism issues also decreased from 12 to 6 following the introduction of Casper. The third outcome examined was the number of residents requiring *informal* learning interventions. These informal learning interventions such as enhanced learning plans allow the residents to continue the program with additional focused effort in areas that need to be addressed. These informal interventions increased from 3 to 11 upon incorporation of the Casper test.

The reduction of *formal* learning interventions (*i.e.,* remediation or probation) from the pre- to post-Casper group was associated with a 96% reduction in costs (*e.g.,* salary for additional training, preceptor remunerations, additional assessments to tailor interventions, logistics [vacations, leaves, travel], resident resource office

support). This improvement was equivalent to approximately $120,000 worth of savings for the university.

Figure 11

Casper's Relations With Remediation Issues



University of Ottawa - *Casper's Relations with OSCE and Clerkship scores.*

The University of Ottawa is ranked as one of the top medical programs in Canada and offers a unique undergraduate medical education program that provides applicants the option to obtain their education in English or French (University of Ottawa, 2021).

For the University of Ottawa's 2020 cohort, Casper scores demonstrated a significant relationship with Professional and Skill Development scores from the OSCE (β =0.53, p=.02) and accounted for 8% of the total variance in clerkship scores ($R^2_{adjusted}$=0.08, p=.001). When other variables were taken into account, Casper scores did not predict clerkship scores beyond what can be explained by GPA and demographic factors.

McMaster University School of Medicine - *Casper's Relations with residency match.*

McMaster University's School of Medicine emphasizes a patient-centred approach and social responsibility in their education program (McMaster University, 2021). A retrospective study of five graduating cohorts from the 2014-2018 application cycles (n=1,021), employed a logistic regression analysis to examine the relationship between Casper scores and the Canadian Resident Matching Service (CaRMS). Casper scores proved to be a significant predictor of residency match with a

corresponding odds ratio of 2.011 (95%CI [1.077, 3.753], *p*=.028) which indicates that for every one unit increase in Casper scores, the odds of being matched to a residency program through CaRMS doubles (Burgess et al., 2020).

UNITED STATES

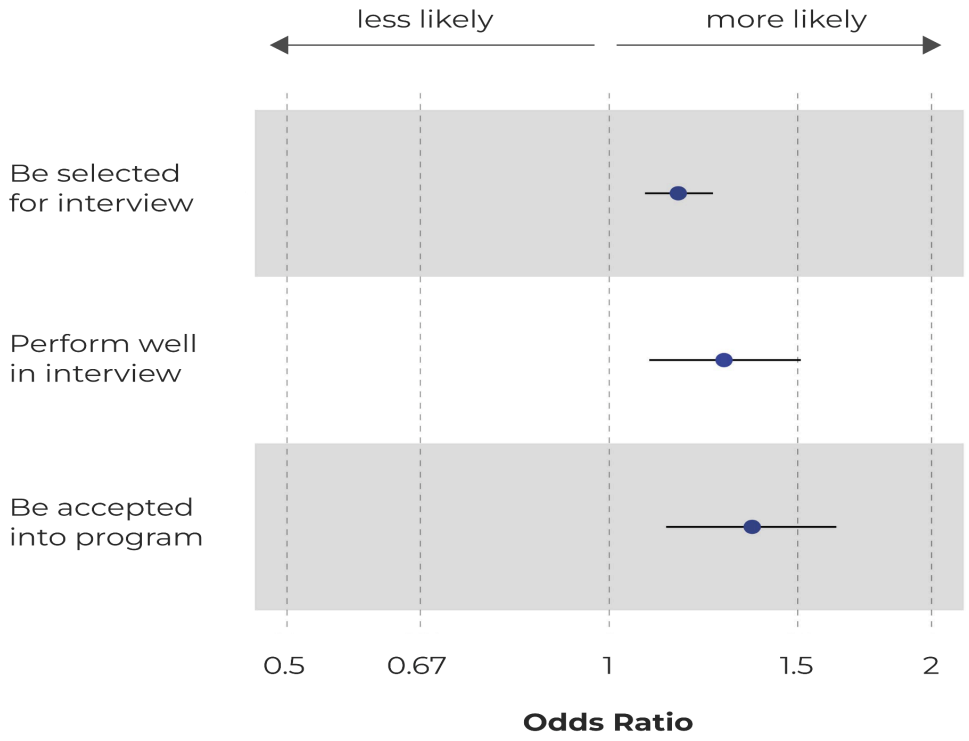
Boston University Chobanian & Avedisian School of Medicine - *Casper's relations with interview selection, interview scores, and program acceptance.*

Boston University Chobanian & Avedisian School of Medicine places a distinct focus on recruiting and developing students, trainees, and faculty who are devoted to serving disadvantaged and underrepresented populations (Boston University, 2023). This focus is achieved, in part, through their use of a comprehensive and holistic applicant review process (Boston University, 2023). Prior to officially incorporating Casper as part of their holistic admissions process, Boston University partnered with Acuity Insights to assess how Casper related to their admissions process.

Casper in Relation to Interview Selection, Interview Performance, and Program Acceptance. A sample of 7,076 applicants from the 2019-2020 admissions cycle were examined. Results indicated that for every 1-unit increase in Casper score, applicants' odds of being selected for interview increased by 16%, their odds of performing well in the interview (*i.e.,* receiving a score of 4 or greater out of 5) increased by 28%, and their odds of being offered acceptance into the program increased by 36% (see Figure 12).

Figure 12

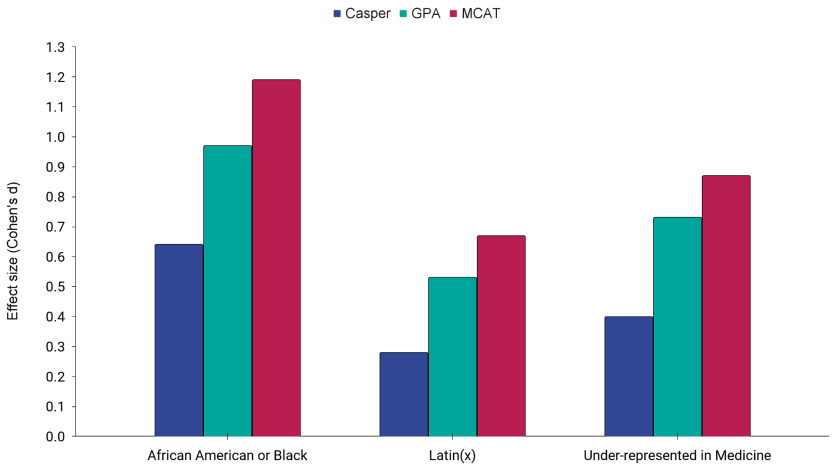
Casper's Relation To Interview Selection, Interview Performance, And Program Acceptance



Demographic Differences. Group differences produced by admissions metrics were also examined in a larger sample of 17,132 applicants who applied in the 2019-2020 and 2020-2021 cycles. As evidenced in Figure 13, when comparing African American or Black applicants and Hispanic applicants to White applicants, Casper had the smallest group differences relative to both GPA and MCAT. The same pattern was observed when comparing under-represented in medicine (URM) applicants and non-URM applicants.

Figure 13

Demographic Group Differences Across Admissions Metrics For Boston University



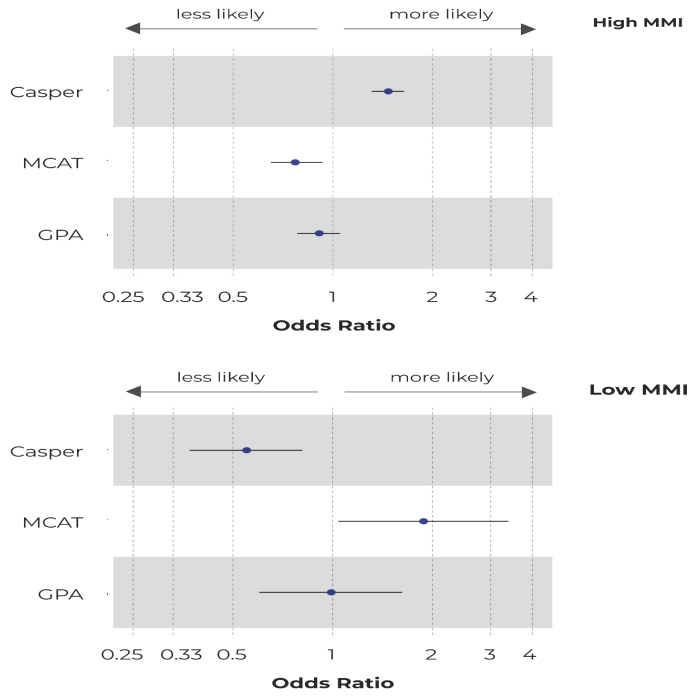
University of Illinois College of Medicine - *Casper's relations with MMI performance.*

The University of Illinois College of Medicine is one of the largest medical schools in the United States, housing a student body of more than 1,300 people (University of Illinois College of Medicine, 2023). With a mission focused, in part, on social responsibility, the program wanted to know if Casper could help them determine who to invite to interview and how it related to their current screening processes.

Casper in Relation to MMI Performance. A sample of 12,601 applicants from the 2020-2021 and 2021-2022 application cycle were examined. Results (see Figure 14) revealed that for every 1-unit increase in Casper scores, applicants were 1.47 times *more* likely to score high in their MMI (*i.e.,* score equal to or greater than 4) and 0.55 times less likely to score low in their MMI (*i.e.,* score less than 3). Interestingly, MCAT showed the opposite relationship such that higher MCAT predicted *lower* MMI performance.

Figure 14

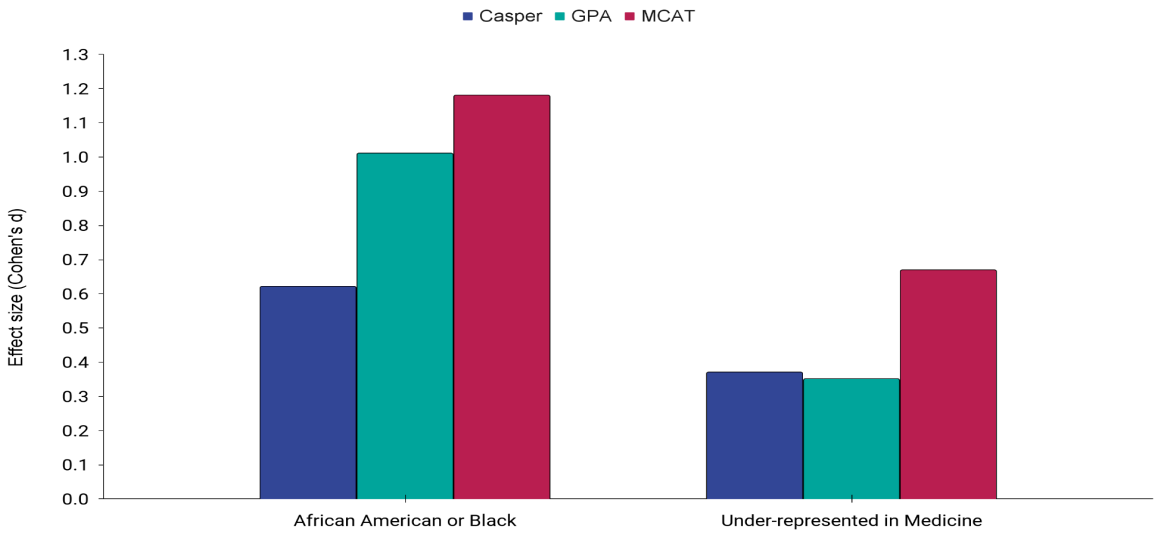
Casper's Relation To MMI Performance



Demographic Differences. Group differences produced by the admissions metrics in applicants were also examined. In the comparison of African American or Black applicants to White applicants, Casper had the smallest group differences (*d*=0.62) relative to both GPA (*d*=1.01) and MCAT (*d*=1.18; see Figure 15 below). In the comparison between URM and non-URM applicants, Casper (*d*=0.37) and GPA (*d*=0.35) showed notably smaller group differences relative to MCAT(*d*=0.67).

Figure 15

Demographic Group Differences Across Admissions Metrics



Texas A&M University College of Medicine - *Casper's relations with interview scores, assessment scores, and program acceptance*.

The College of Medicine at Texas A&M University was founded specifically to provide care to underserved populations across the state (Texas A&M University, 2021). With such a client-centred mission and approach to medical education, it may be of little surprise that the University chose to adopt Casper into their admissions process.

In collaboration with Texas A&M University, the Research Team at Acuity Insights evaluated how well various admissions metrics (Casper, GPA, and MCAT) could predict interview scores, Standardized Patient Exercise (SPE) scores, and acceptance rates. Demographic differences were also assessed across the three admissions metrics.

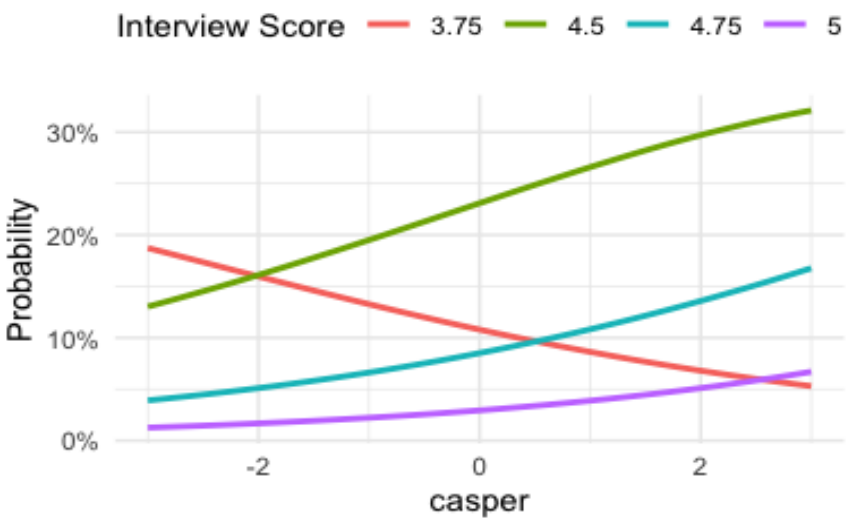
Casper in relation to program interview scores. General linear regression was employed to predict program interview scores using three admissions metrics: Casper, GPA, and MCAT subsections (*n*=1,323). A one point increase in Casper score led to a 0.05 increase in interview score. This increase in interview score was larger than what resulted from a one point increase in two subsections of the MCAT. Specifically in the Critical Analysis and Reasoning Skills (CARS) subsection (0.03 point increase in interview score) and in the Psychological, Social, and Biological Foundations of Behavior (PSBB) subsection (0.04 point increase in interview score). These increases in interview score were, however, all smaller than produced by a one unit increase in GPA which led to an increase of 0.37 points in interview score.

When examining two regression models, one with only GPA and MCAT subsections and one with GPA, MCAT subsections, *and* Casper scores, the inclusion of Casper improved the predictive ability of the model by 2% resulting in a model which accounted for 16% of the variance in interview scores.

Ordinal logistic regression models were used to predict interview scores with Casper, GPA, and MCAT. As can be seen in Figure 16, as Casper scores increased, the probability of receiving a high interview score (4.5 and above) increased while the probability of receiving a lower interview score (3.75) decreased. Specifically, a one-unit increase in Casper scores (0 to 1) increased applicants' probability of receiving an interview score of 5 (the highest score) by 3% and increased applicants' probability of receiving an interview score of 4.75 by 3% as well. Alternatively, changes in GPA from 3.6 to 3.9 increased applicants' probability of receiving an interview score of 5 by only 1% and increased applicants' probability of receiving an interview score of 4.75 by 3%.

Figure 16

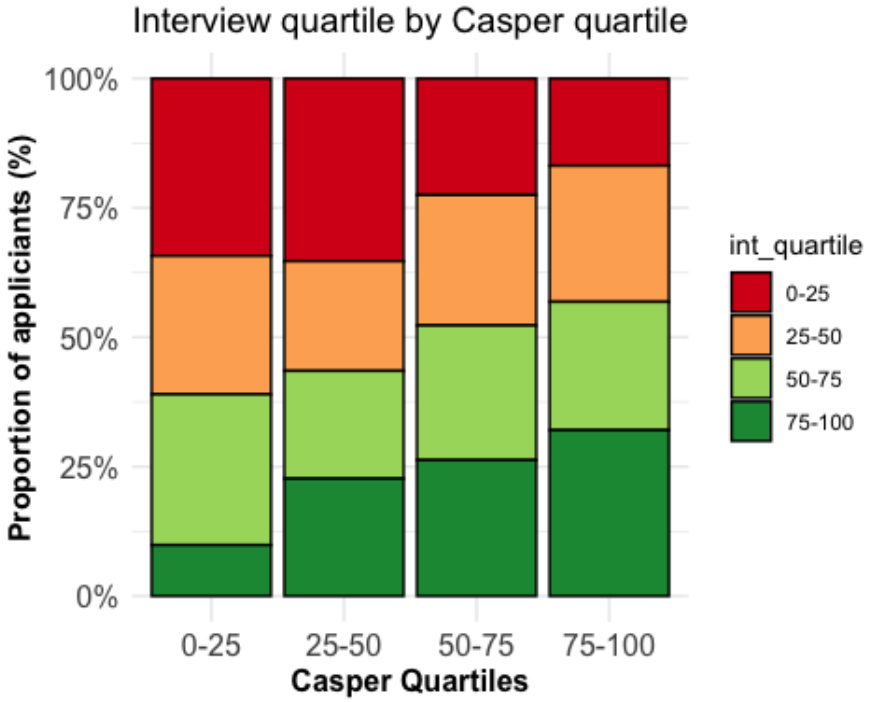
Probability Of Receiving Various Interview Scores As A Function Of Casper Scores



The relationship between Casper scores and interview scores was also examined using quartiles. As evident in Figure 17, as Casper quartiles increase, so too does the proportion of applicants who scored in the top quartile of interview scores. Simultaneously, as Casper quartiles increase, the proportion of applicants who scored in the bottom quartile of interview scores decreased.

Figure 17

The Relationship Between Casper Quartile Scores And Interview Quartile Scores

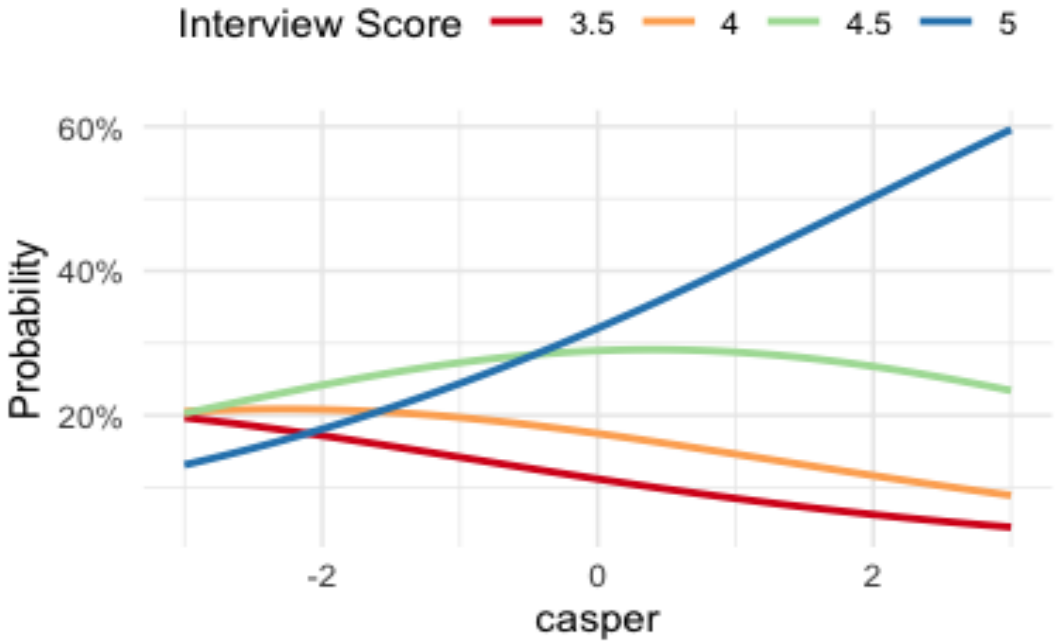


Casper in relation to program Standardized Patient Exercise (SPE) scores. General linear regression was also employed to predict program Standardized Patient Exercise (SPE) scores using Casper, GPA, and MCAT subsections (*n*=1,323). A one point increase in Casper scores led to a 0.11 point increase in SPE scores. Alternatively, a one-unit increase in the Chemical and Physical Foundations of Biological Systems (CPBS) subsection scores of the MCAT and a one-unit increase in GPA resulted in a *decrease* of 0.03 points and a *decrease* of 0.19 in SPE scores, respectively.

Ordinal logistic regression analysis was also performed to predict Standardized Patient Exercise (SPE) scores using GPA, MCAT, and Casper. Of the three, Casper was the only measure that proved to be a significant predictor of SPE score. A one-unit increase in Casper scores (0 to 1) increased applicants' probability of receiving a score of 5 (highest score) by 10% (Figure 18).

Figure 18

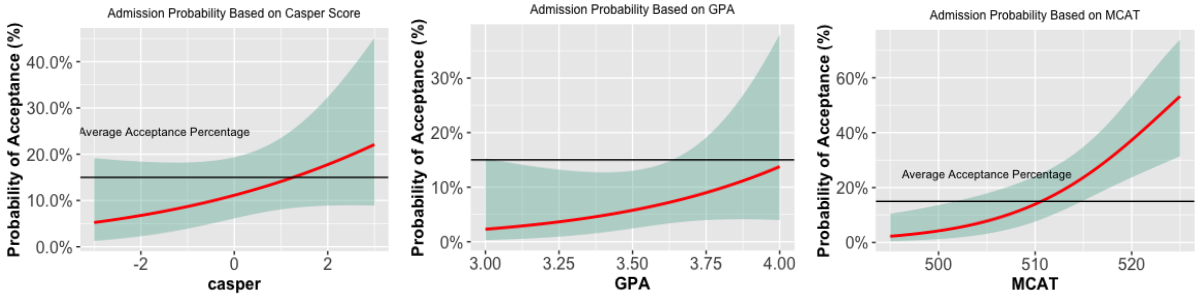
Probability Of Receiving Various SPE Scores as a Function Of Casper Scores



Casper in relation to program acceptance. A series of logistic regression analyses were used to assess how well Casper scores, MCAT scores, and GPA could predict applicant acceptance (Figure 19) using data from 3 application cycles between 2018 and 2020 (*n*=6,936). Results indicated that for every 1 standard deviation increase in Casper scores (*e.g.,* 0 to 1), applicants' odds of being accepted increased by 30% (OR= 1.31, 95%CI=[1.22, 1.42]). A 1 standard deviation increase in GPA (*e.g.,* 3.6 to 3.9) doubled applicants' odds of program acceptance (OR=6.81, 95%CI=[4.83, 9.71]). Finally, a 1 standard deviation increase in MCAT (*e.g.,* 500 to 507) increased applicants' odds of acceptance by approximately 90% (OR=1.14, 95%CI=[1.12, 1.15]).

Figure 19

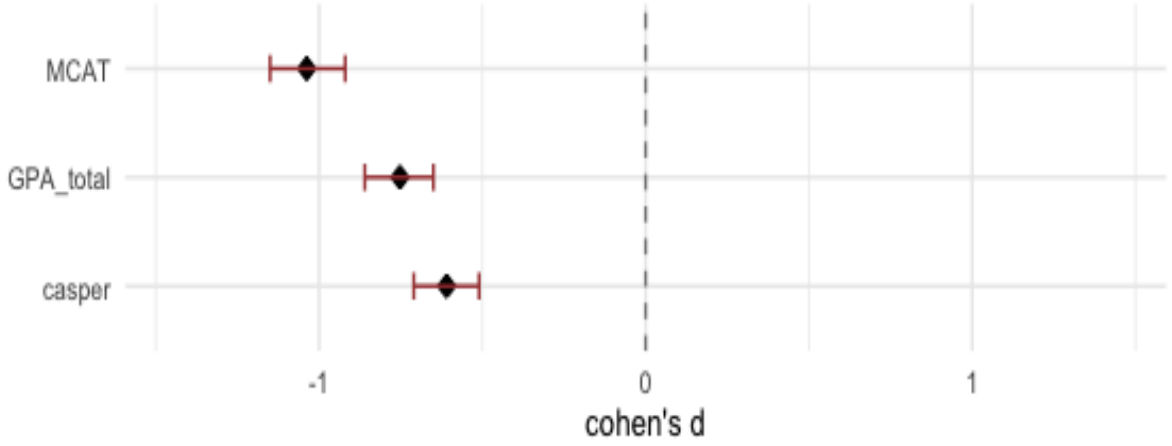
Probability Of Program Acceptance As A Function Of Casper, GPA, And MCAT Scores



Casper in relation to Demographic Group Differences Across Applicant Race. Across the three application cycles (2017-2020; *n*=6,936), Black, African, Caribbean, or African American applicants tended to produce lower scores on admissions metrics relative to White or European applicants, but the size of this difference varied by metric. As can be seen in Figure 20, Casper produced the smallest group differences (*d*=0.61) relative to the group differences observed in GPA (*d*=0.75) and MCAT scores (*d*=1.04). Casper also produced the smallest group differences when comparing White or European applicants to (1) Hispanic, Latino, or Spanish origin applicants (Casper: *d*=0.23; GPA:*d*=0.35; MCAT: *d*=0.57) and to (2) Asian applicants (Casper: *d*=0.09; GPA: *d*=0.10; MCAT: *d*=0.13).

Figure 20

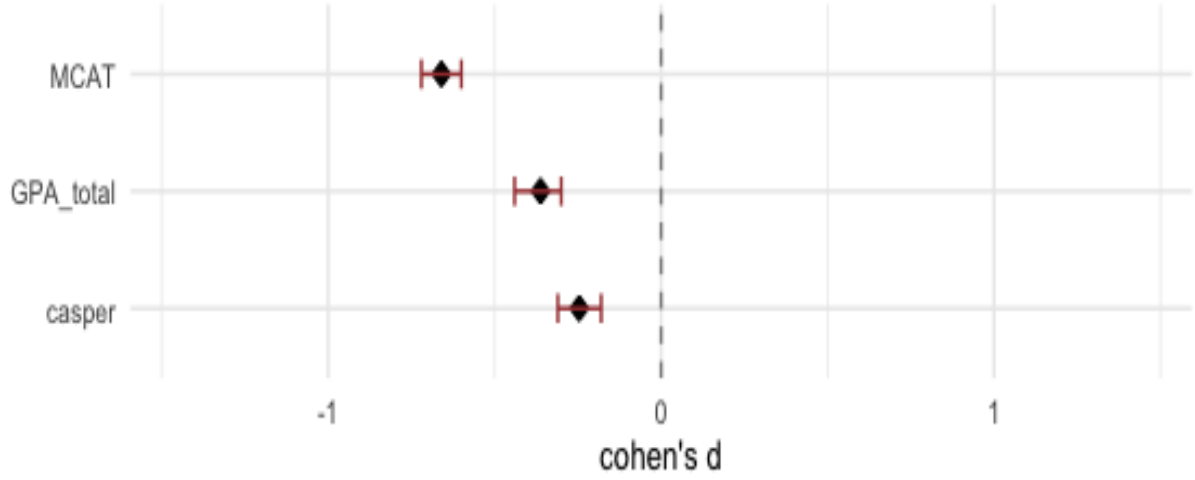
Comparison Of Black, African, Caribbean, Or African American And White Or European Applicants Across Admissions Metrics At Texas A&M University



Casper in relation to Demographic Group Differences Across Applicant SES. Applicants with low SES status or who self-declared as disadvantaged tended to produce lower scores than applicants without these statuses. While this pattern remained consistent across the three admissions metrics, it is clear in Figure 21 that Casper still produced the smallest group differences (*d*=0.25) compared to GPA (*d*=0.36) and MCAT (*d*=0.66).

Figure 21

Comparison Of Low SES And Non-Low SES Applicants Across Admissions Metrics At Texas A&M University



Hofstra University - ***Casper as a Screening tool prior to interview.***

Hofstra University, Long Island's largest private university is home to over 10,000 students with over 400 of them studying medicine. Hofstra is committed to ensuring a diverse body of students and faculty make up the academic population (Hofstra University, 2021). To examine the usefulness of Casper in their admissions process relative to their MMI, Hofstra conducted a simulation study. Although this program did not require applicants to include Casper scores as part of their admission package at the time, more than 90% of their applicants had taken Casper in order to apply to other programs. Since applicants are permitted to take the Casper test only once a year for all programs, this allowed the team at Acuity Insights to assess what would have happened if Casper were used in the admissions process.

In the simulation study, Casper scores produced a moderate correlation with the University's MMI scores. In addition, researchers found that if the University used low Casper scores (*e.g.,* -1.5 or less) as a metric for screening out applicants, there would have been a reduction of 210 applicants at the interview stage, 25 of whom received below-acceptable MMI scores in the interview.

Although these are results of a simulation study, it appears that the incorporation of Casper into the admissions process for Hofstra University would have decreased the applicants brought to the interview process, some of whom subsequently did not pass the interview. Consequently, the use of Casper scores early in the process would have reduced resources required to interview so many applicants.

Indiana University School of Dentistry - *Casper reliability, discriminant validity, and comparison between groups.*

The Indiana University School of Dentistry trains 80% of dentists who practice in the state of Indiana and places great emphasis on patient-centred care providing oral health care to more than 19,000 patients every year, many of whom may not have the opportunity otherwise (The Trustees of Indiana University, 2021).

Within this study, Casper evidenced excellent levels of test reliability (α=0.85). Strong evidence for discriminant validity was also noted as Casper scores demonstrated minimal correlations with GPA (r=0.08, p<.05) and scores on the Dental Aptitude Test (r=0.16, p<.001).

Taken together, it is clear that Casper is a reliable measure for the Indiana University School of Dentistry and provides information on applicants that could not be collected from the technical skill metrics typically examined.

Tulane University School of Medicine - *Casper relations with clerkship honours designation and incomplete grades.*

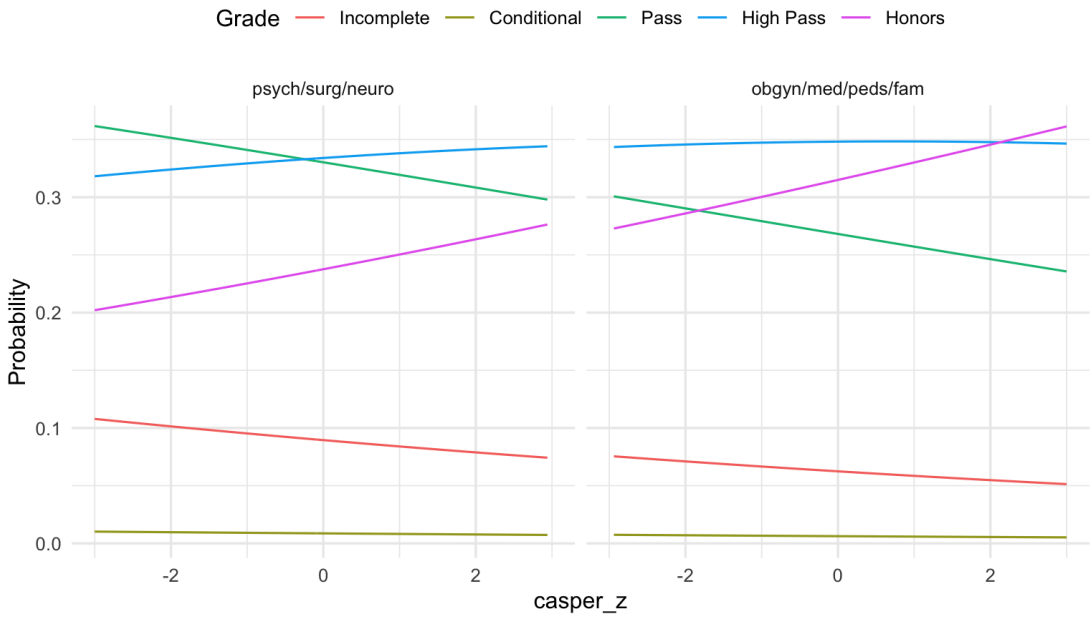
Tulane University School of Medicine is considered one of the most recognized centers for medical education in the United States and is the nation's 15th oldest medical school (Tulane University, 2021). With an emphasis on education, patient care, and research, it is important that applicants accepted into the program succeed in the program; one way to evaluate this is to examine the relationship between admissions metrics (*i.e.,* Casper scores) and clerkship grades.

Based on data from 92 students, increases in Casper scores were associated with increased probability of a student receiving an Honors grade in their clerkship and a decreased probability of receiving an incomplete or simple pass grade (Figure 22). This is true across various clerkships including psychiatry, surgery, neurology, obstetrics and gynecology, medicine, pediatrics, and family medicine.

Although applicants' GPA had the largest effect on the odds of receiving a favorable clerkship grade (OR=3.575, 95%CI[0.729, 18.748]), Casper evidenced a similar effect to that produced by the MCAT. Importantly, as Casper scores increased, so too did the odds of receiving a favorable clerkship grade (OR=1.072, CI95%[0.737, 1.564]) which was opposite to that of the MCAT which indicated that increases in MCAT scores reduced the odds of receiving a favourable clerkship grade (OR=0.984, CI95% [0.937, 1.033]).

Figure 22

Relationship Between Casper Scores And Grade Probability

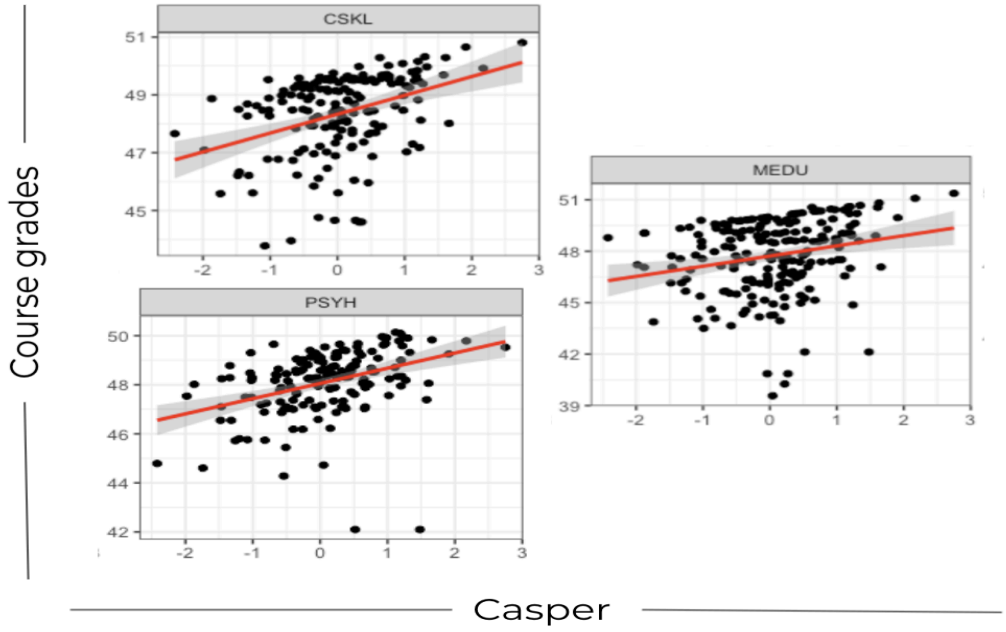


Anonymous Program: Mid-Sized United States Doctor of Medicine Program - *Casper relations with class performance and information on demographic group differences.*

In an analysis of a mid-sized United States Medical program (*n*=144), Casper proved to be significantly associated with grades from three courses: clinical skills (CSKL), psychiatry (PSYH), and medicine (MEDU; Figure 23). Using a linear regression model, the explained variance increased significantly when Casper was added to the model which originally only contained information on demographics and MCAT scores. Specifically, with the inclusion of Casper scores, the adjusted R^2 value increased from 0.13 to 0.27. While the full regression model (demographics, MCAT, and Casper) explained a significant portion of variance for 7 subjects, Casper was only a significant predictor of the three aforementioned subjects and not for the other four (anatomy, biochemistry, genetics, and physiology). This latter result speaks to the discriminant validity of Casper scores as they do not evidence a relationship with technical knowledge courses.

Figure 23

Relationship Between Casper Scores And Course Grades



Note. CSKL=clinical skills; PSYH=Psychiatry; MEDU=Medicine

Demographic Group Differences. In terms of demographic group differences, similar patterns were observed in both Casper and MCAT scores, however the standardized mean difference between groups was smaller and non-significant for Casper scores. On average, White or European applicants produced higher Casper scores (0.10) relative to applicants of any other race (-0.06; d=0.32, *non-significant*). White or European applicants also, on average, produced higher scores relative to applicants of any other race, on the MCAT (d=0.37, p<.05).

Evansville Physician Assistant Program - *Casper relations with technical and non-technical admissions metrics and in-program course grades.*

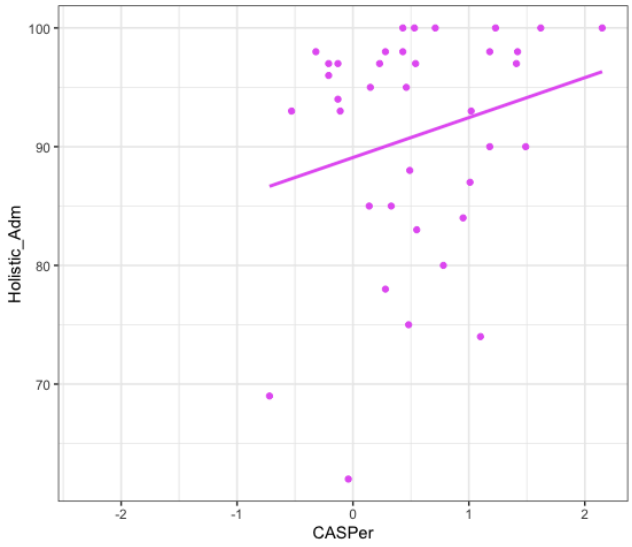
The University of Evansville Physician Assistant Program receives a large number of applicants (approximately 400) relative to the number of spots available (approximately 40; University of Evansville, 2021). With this limited number of spots each year, it is vital that the best applicants are selected during the admissions process. The University of Evansville stresses the importance of incorporating soft skills (*e.g.,* effective communication skills) into the definition of what a successful applicant looks like, which is why they chose to incorporate Casper scores into their admissions process (University of Evansville, 2021). In collaboration with the Research Team at Acuity Insights, the University of Evansville assessed the relationship between Casper scores and other (i) admissions metrics and (ii) in-program course grades.

Casper in relation to other admissions metrics. Casper evidenced near-zero correlations with several technical-skill admissions metrics that are used in the Evansville selection process. In a sample of 80 applicants, Casper evidenced non-significant minimal correlations with GRE total scores (r=0.06) as well as the math (r=0.04) and verbal (r=0.02) subsections of the GRE. Similarly, Casper evidenced a non-significant minimal correlation with science GPA (r=0.08, n=117). Casper also demonstrated a negative correlation (r=-0.11) with scores from the Physician Assistant National Certifying Exam (PANCE) which is used to assess technical knowledge and skills. In a regression analysis which examined the amount of variance in PANCE scores that could be accounted for by Casper, it was clear that there was no predictive relationship between PANCE and Casper, as Casper scores could only account for 1% of the variance in PANCE scores. These results support the strong discriminant validity of the Casper test.

Alternatively, Casper evidenced stronger relationships with other assessments of non-technical skills such as the MMI (r=0.18, n=80) which speaks to the convergent validity of the test. Although this relationship was non-significant, it is clear that Casper has a much stronger relationship with MMI scores relative to GPA and GRE. Casper evidenced an even larger correlation with the program's Holistic Review scores (r=0.23, n=37; see Figure 24 for a visualization of the relationship).

Figure 24

Relationship Between Casper Scores And Holistic Admissions Scores



Casper in relation to in-program course grades. In addition to understanding how Casper relates to admissions metrics, it is equally important to know if Casper is related to applicants' future behaviour by evaluating the relationship between Casper and in-program grades. As expected, results were in line with previous discriminant validity findings in that Casper evidenced minimal

relationships with course grades that reflected technical knowledge such as physiology (*r*=-0.03), pharmacology (*r*=0.12), behavioural health (*i.e.,* psychiatric conditions; *r*=0.07), anatomy (*r*=-0.01), medical literature (*r*=0.04), and therapeutic interventions (*r*=0.09). Similarly, courses which reflect students' ability to perform various medical skills did not evidence strong correlations with Casper including diagnostics (*r*=0.07), medical imaging (*r*=0.00), electrocardiography reading (*r*=-0.04), clinical skills (*r*=0.02), as well as taking patient history and conducting a physical exam (part I: *r*=0.12; part II: *r*=-0.02), all of which support the discriminant validity of Casper.

Anonymous Program: Small United States Doctor of Osteopathic Medicine Program - *Casper relations with critical analysis and reasoning skills as well as interview scores.*

A small Doctor of Osteopathic Medicine Program in the United States examined the relationship between Casper scores and various other admissions metrics for a sample of 12,870 applicants during the 2018-2019 and 2019-2020 application cycles.

Casper scores produced significant correlations with the Critical Analysis and Reasoning Skills section of the MCAT exam (*r*=0.24, *p*<.001). Casper scores also had a significant, yet small, correlation with the interview scores provided by members of the University (*r*=0.11, *p*<.001), indicating that Casper is providing information that may not be collected through traditional interview practices.

Anonymous Program: Mid-Sized United States Doctor of Medicine Program - *Casper relations with MMI competencies.*

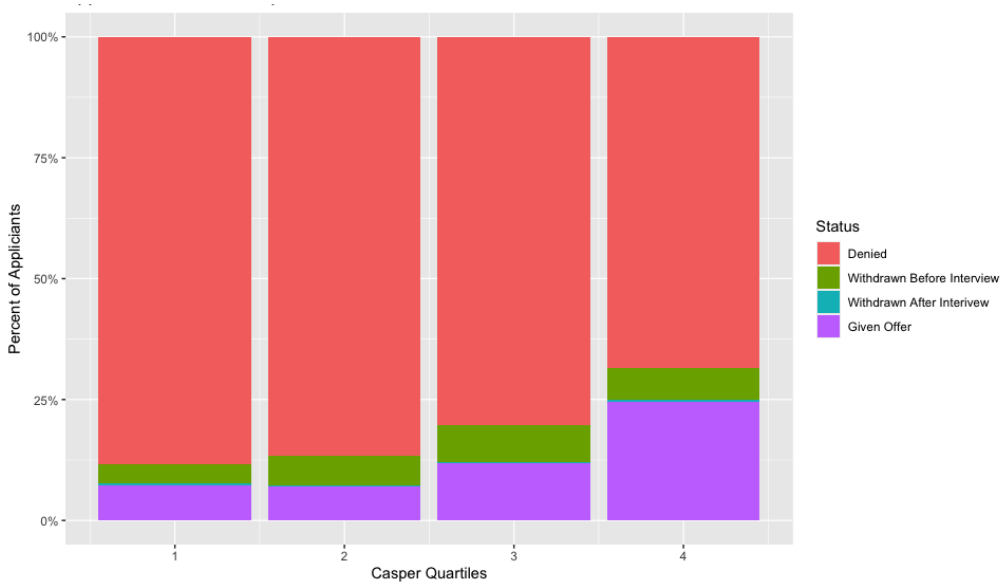
Results from a mid-sized medical education program in the United States showed that Casper scores had much stronger correlations with MMI competencies (ranging from *r*=0.21 to 0.33) than with technical-knowledge measures such as the MCAT (*r*=0.11) and GPA (*r*=0.04). More specifically, Casper demonstrated correlations with the following MMI competencies: conflict resolution (*r*=0.28), cultural competency (*r*=0.21), ethics and morality (*r*=0.33), collaboration (*r*=0.23), problem solving (*r*=0.24), and self-appraisal (*r*=0.26). Notably, GPA and MCAT had no significant correlation with any of the MMI competencies.

Anonymous Program: Mid-Sized Southern United States Physician Assistant Program - *Casper relations with interview performance and acceptance.*

Results from a mid-sized physician assistant program in the United States indicated that high Casper quartiles were associated with a greater percentage of applicants receiving an offer (see figure 25). Additionally, Casper evidenced a statistically significant association with interview performance (*b*=0.78, *p*<.05), although the Casper score did not add explained variance when technical-knowledge measures were included in the model.

Figure 25

Relationship Between Casper Quartile Scores And Offers Of Admission



AUSTRALIA

Monash University - *Evidence that Casper is able to select applicants who will succeed in their academic placements.*

Monash University, the largest University in Australia and home to over 86,000 students in the 2019 academic year (Monash University, 2012) noted significant decreases in *notifications of concern* (any concern regarding the relationship between a student and placement school) amongst pre-service teachers after Casper was introduced. Data from Monash's 2017-2019 application cycles (n=590) showed that when applicants were selected without considering Casper scores, 9.8% of the cohort received a notification of concern. This decreased significantly to 0.8% (p<.001) for the applicants in the subsequent cohort who were selected with the inclusion of Casper scores. Further, the few students in the latter cohort who attained a notification of concern (n=4), had lower mean Casper scores (z= -0.20) than those who did not receive a notification of concern (z= 0.51).

For Monash University's Teacher Education program, the incorporation of Casper not only reduced the resources required to handle notifications of concern, but also aided in preserving relationships with partner schools.

University of Wollongong - *Evidence that Casper is similar to the other assessments of soft skills and dissimilar from assessments of technical knowledge.*

The University of Wollongong's (UOW) Graduate Medical program is dedicated to developing top-tier medical professionals who are dedicated to serving rural communities through a patient-centered approach (University of Wollongong, 2021).

Producing doctors with non-technical skills that are equal to their technical skills requires selecting applicants who have excellent social intelligence and professionalism, which is exactly what Casper has aided with since it was first piloted at UOW in 2019.

Applicants applying for the 2019 academic year (n=1,548) were asked to complete the Casper test as part of the admissions process. Casper scores demonstrated a clear relationship with applicants' performance on the MMI (r=0.38, p<.001 before correcting for disattenuation; r=0.52, p<.001 after correcting for disattenuation), with a particularly impressive ability to identify potentially problematic applicants. Ninety-nine percent of applicants who received a Casper score of less than -1 also received a "red flag" from at least one interviewer during their subsequent MMI (Parker-Newlyn et al., 2019).

It was also clear, from small correlations, that Casper scores provided unique applicant information that could not be gathered from the traditional admissions portfolio scores (r=0.19, p<.001) or from technical measures such as the GAMSAT (r=0.23, p<.001) and GPA (r <0.15, p<.001). A particularly potent benefit of Casper scores is that they demonstrated absolutely no correlation with rural upbringing or rural education, and negligible correlations with age and gender (r <.15, p<.001).

Casper's relationship with MMI scores, divergence from technical measures, and lack of bias, paired with the fact that 86% of applicants reported being satisfied, very satisfied, or extremely satisfied with their test experience, were all driving factors in the University of Wollongong's decision to include Casper in their admissions process.

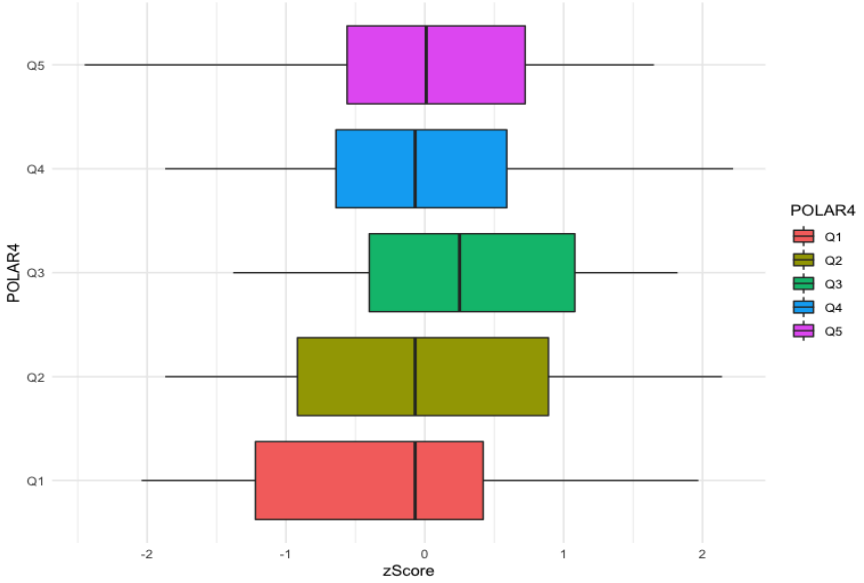
UNITED KINGDOM

Kent University - *Evidence that Casper scores do not vary by applicant SES.*

Established in 2020, the Kent and Medway Medical school aims to attract students from a variety of backgrounds and locations in order to develop a diverse physician workforce (Kent and Medway Medical School, 2021). To determine if Casper would aid in this mission, a research study was conducted to determine if Casper scores varied by scores on the Participation of Local Areas (POLAR) system. Using the POLAR system, the geography of the United Kingdom is classified into five groups based on the proportion of young adults (18-19 years old) who enter post-secondary education. This is meant to act as a proxy of socio-economic status to identify applicants from disadvantaged areas. As can be seen in Figure 26, results provide evidence that Casper scores do not vary as a function of geographic location of applicants. This indicates that applicants' socioeconomic status does not impact their Casper score.

Figure 26

Relationship Between Casper Scores And Geographical Location Across The United Kingdom



Part III Summary

Overall, it is clear that Casper has a significant ability to predict a range of future non-technical exam scores and in-program behaviour. The level of predictive ability matches that of technical-knowledge measures used in the admissions process. Casper has consistently evidenced an ability to predict applicant performance on licensure exams, performance on in-program measures of success (*e.g.,* OSCE exams, clerkship grades, etc.), interview scores, and professional behaviour. Simultaneously, across the select partner schools presented here, Casper has also continuously demonstrated minimal or non-existent relationships with dissimilar measures (*e.g.,* GPA, technical skill exams, etc.) and lower demographic group differences relative to other assessments. As evidenced by the partner research presented in this section, Casper provides substantial value to the admissions process for a variety of programs across Canada, the United States, Australia, and the United Kingdom.

Part IV. Research Examining Casper as a Measure that is Equitable Across Demographic Groups

Demographic Differences - *Information on how Casper performs across applicants.*

With *fairness* being a cornerstone of the Casper test, detecting and mitigating demographic group differences is paramount to achieving this goal. In working toward this goal, several measures have been taken in developing Casper as a fair and equitable assessment. This chapter begins by summarizing sources of potential variance and sharing how Acuity Insights is working to address these. Following this, the demographic group differences observed in the Casper test are summarized and discussed. The chapter is closed with a discussion on the steps Acuity Insights is taking to identify and mitigate test-level bias within Casper.

Sources of Potential Variance. Within the Casper test, there are four main sources of potential bias: systematic, test format, raters, and test content. Below, the methods we are using to address these sources of potential bias are outlined.

- ***Systemic Issues.*** Although unwavering in the pursuit to vastly reduce demographic group differences, the team at Acuity Insights recognizes that long-standing social inequalities rooted within the structure of society contribute to demographic differences at a larger scale (Williams, 1983), as reflected across different admissions metrics more broadly (Whitcomb et al., 2021; Smith & Reeves, 2020; Mortaz Hejri et al., 2022). Although these systemic issues must be addressed at a much higher, societal level, Acuity Insights is dedicated to ensuring everything that can be done at the test-level to address construct-irrelevant variance in group differences is addressed. If one specific test shows group differences while other tests in the same area of assessment do not, then there is a good chance the test may be biased toward a specific group. If, on the other hand, group differences reappear across tests in an area of assessment, then the group differences may be partly or largely due to systemic issues. It is incumbent upon test publishers to find ways to mitigate or reduce such systemic factors, while recognizing the importance of changes needed at a broader societal level.
- **Test Format:** The format of the Casper test was developed with several considerations for equity. First, the test uses behavioural tendency questions (what would you do) opposed to knowledge-based questions (what should you do). Behavioural tendency questions have shown to produce lower demographic group differences relative to knowledge-based questions (Whetzel et al., 2008). Second, the Casper test uses an open-ended (*i.e.,*

constructed response) format which means that applicants are not forced to select a response from a list, but rather have the flexibility and opportunity to demonstrate what they would do based on their lived experiences and personal values. Open-ended responses have also shown to produce lower demographic differences relative to close-ended responses (Lievens & Peeters, 2008). As a final, and notable piece, the team at Acuity Insights has worked for several years to determine if and how a video-response component could be incorporated into the test to enhance equity. This new response format has officially been incorporated into the test and the research used to make this decision can be found in the **Experimenting to Further Improve Equity** section below.

- **Raters:** Several considerations are in place for raters, who are responsible for scoring every Casper response. A diverse group of raters, each holding unique perspectives, are recruited for each geographical location. Raters may only rate Casper tests from the same geographical location in which they are living to ensure that they are scoring in accordance with cultural norms. Further, prior to scoring, raters receive implicit bias training to assist them in identifying any potential biases they may possess and learn how to combat these. To further reduce the impact of bias during the scoring process, each question of an applicants' Casper test is scored by a unique rater to dilute any implicit bias that may still be present despite the numerous safe-guards in place.

- **Test Content:** As previously described in this document, the content of each Casper test is meticulously crafted and evaluated by numerous stakeholders both internal and external to Acuity Insights to assess for any potential bias in the scenario or question set. In addition to this initial process to mitigate test bias, the research team has conducted analyses to assess for potential bias in test content (measurement invariance and differential item functioning) which is described in more detail later in the document.

Group Differences. Group differences for Casper and other tests are generally assessed via standardized mean difference values and regression analyses. 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). These analyses are conducted each application cycle for all demographic group comparisons. Complementary to these statistics, regression analyses were conducted for the most recent application cycle (2023-2024) to provide information on the effect of each demographic variable while controlling for all others. A regression model for each geographic region is provided in Appendix 2 along with a measure of effect size (η_p^2) for each demographic variable which explains the size of the effect each variable had on z-scores. These effect sizes are

often interpreted such that values of 0.01, 0.06, and 0.14 are the thresholds for small, moderate, and large effect sizes, respectively (Cohen, 1988).

The subsections below provide an overview of the demographic group differences observed in Casper. Where data is available, direct comparisons between the magnitude of group differences for Casper and various admissions measures (*e.g.,* MMI and technical-knowledge measures) are also drawn. For each group comparison, group-level mean Casper scores and standardized mean difference scores are presented. To gather this information, applicants are provided a post-test survey (specific for each country) which asks them to share demographic characteristics if they feel comfortable to do so. Applicants are informed that the survey is optional and anonymous prior to starting. They are also informed that if they choose to participate in the survey, they are not required to answer all questions; instead, they are asked to answer only questions they feel comfortable providing a response to. As we continually work to upgrade the survey to ensure it is safe, inclusive, and fair for all applicants, the questions and response options have changed over the years. For this reason, readers may notice that certain questions have information for all application cycles while others start or stop at certain time points.

For ease of reading, the tables within this section have been truncated to include only the three most recent application cycles. Full tables with information for all application cycles are available in Appendix 1. This section covers detailed information on group differences in performance across 9 demographic characteristics:

- Gender
- Socio-Economic Status
- Age
- Rurality
- Language
- Employment
- Ability Status
- Domestic or International Status
- Race and Ethnicity

Gender

The first question in the post-test survey asks applicants which response option most accurately describes them, with the option to self-describe as well. While we collect information on applicants with all genders and gender identities, small sample sizes prevent us from conducting group difference analyses for all groups at this time. Therefore, for the time being, the comparison of performance based on gender is restricted to female and male applicants.

Typically, female applicants tend to produce scores slightly higher than their male counterparts (Table 8). These differences in test performance are consistently

negligible to small in size with an average Cohen's *d* value of *d*=0.14 across all geographies and languages from the three most recent application cycles. Regression analyses (Appendix 2) indicated that female gender identity had a small positive effect on Casper scores across all geographies (β= 0.11 to 0.25, *p*<.001), but the effect size was either negligible (η_p^2 = 0.00) or small (η_p^2 = 0.01).

Table 8

Mean Scores of Female Applicants Compared to Mean Scores of Male Applicants

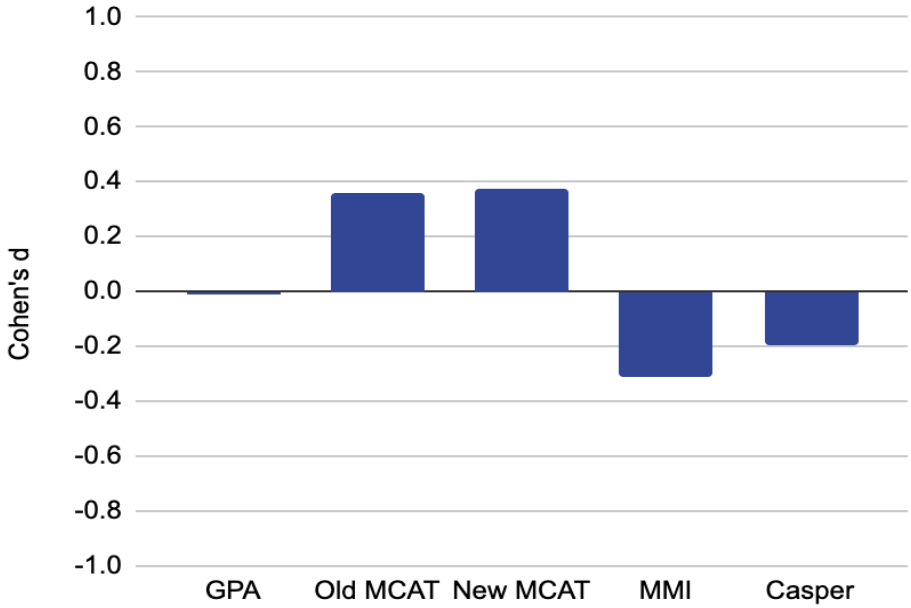
Application Year	Country	Female	Male	Cohen's *d* [95%CI]	*p*
2021-2022	Canada (English)	0.03 (*n*=27,606)	-0.05 (*n*=9,743)	0.08 [0.05, 0.10]	<.001
	Canada (French)	0.07 (*n*=5,437)	-0.10 (*n*=2,006)	0.18 [0.13, 0.23]	<.001
	United States	0.07 (*n*=46,456)	-0.12 (*n*=24,899)	0.19 [0.18, 0.21]	<.001
	Australia	0.04 (*n*=8,265)	-0.03 (*n*=3,242)	0.06 [0.02, 0.11]	.002
2022-2023	Canada (English)	0.03 (*n*=22,177)	-0.03 (*n*=7,804)	0.06 [0.03, 0.08]	<.001
	Canada (French)	0.08 (*n*=4,606)	-0.07 (*n*=1,680)	0.15 [0.10, 0.21]	<.001
	United States	0.05 (*n*=41,109)	-0.11 (*n*=20,965)	0.16 [0.15, 0.18]	<.001
	Australia	0.04 (*n*=6,491)	-0.05 (*n*=2,298)	0.09 [0.05, 0.14]	<.001
2023-2024	Canada (English)	0.04 (*n*=19,809)	-0.05 (*n*=6,764)	0.10 [0.07, 0.13]	<.001
	Canada (French)	0.06 (*n*=4,596)	-0.08 (*n*=1,487)	0.13 [0.08, 0.19]	<.001
	United States	0.08 (*n*=34,663)	-0.15 (*n*=16,630)	0.23 [0.21, 0.24]	<.001
	Australia	0.08 (*n*=6,698)	-0.10 (*n*=3,089)	0.19 [0.15, 0.23]	<.001

Casper Demographic Differences Relative to Other Admissions Metrics. In an analysis of 9,096 applicants to the New York Medical College School of Medicine

(NYMCSM), direct comparisons were made between Casper, MMI, and technical-knowledge measures regarding male and female differences (Juster et al., 2019). As detailed in Figure 27, Casper demonstrated lower group differences between male and female applicants compared to the MMI, modern MCAT, and the former version of the MCAT.

Figure 27

Comparison Of Male And Female Demographic Differences Across Admissions Metrics (Juster et al., 2019)



Note. Negative values indicate that female applicants produced higher scores than males for this particular study.

Socio-Economic Status

To assess group differences between applicants of varying socio-economic statuses (SES), two proxies of SES are examined: parental income and highest level of parental education (see Table 9 for examples of the question asked to applicants). For ease of interpretation, both variables are dichotomized. For parental income, applicants are split into two groups: applicants whose annual household income is equal to or greater than $100,000 and applicants whose annual household income is less than $100,000. Likewise, for parental education, applicants are split into two groups: applicants whose parents have a Bachelor's degree or higher and applicants whose parents have less than a Bachelor's degree.

Table 9

SES Proxies and Questions Asked

SES Proxy	Question in Post-Test Survey
Parental Income	Regardless of your dependency status, please indicate the gross income for one or more of your parents' for last year (a rough estimate is sufficient).
Parental Education	What is the highest degree or level of school either of your parents have completed? (If they are currently enrolled in school, please indicate the highest degree they have received).

Parental Income. As can be seen in Table 10, applicants with an annual household income equal to or above $100,000 consistently produce higher Casper scores relative to applicants with annual household incomes below $100,000. Across the three most recent application cycles, the average magnitude of group differences was small (*d*=0.28; range: *d*=0.12 - 0.33). Across all geographies and languages, results from each regression analysis (Appendix 2) suggest that, when holding all other demographic variables constant (*e.g.,* race, gender, age, etc.), applicants with annual household incomes equal to or greater than $100,000 tend to produce higher Casper scores relative to applicants with household incomes below $100,000 (β= 0.10 to 0.16). Although these results are statistically significant (*p*<.05; which is expected with large sample sizes), the effect of this variable on z-scores is small in each regression model (η_p^2 = 0.01 - 0.02). Thus, parental income when examined in relation to other demographic factors, does not impact Casper scores. The effect is likely mediated by other demographic factors.

Table 10

Mean Scores Of Household Incomes Of $100,000 Or More Compared To Mean Scores Of Household Incomes Below $100,000

Application Year	Country	Income Equal to or Above $100,000	Income Below $100,000	Cohen's *d* [95%CI]	*p*
2021-2022	Canada (English)	0.22 (*n*=13,169)	-0.10 (*n*=16,127)	0.33 [0.31, 0.36]	<.001
	Canada (French)	0.20 (*n*=3,353)	-0.09 (*n*=2,541)	0.30 [0.25, 0.35]	<.001
	United States	0.16 (*n*=29,334)	-0.16 (*n*=27,510)	0.33 [0.31, 0.34]	<.001
	Australia	0.20 (*n*=4,158)	-0.06 (*n*=4,546)	0.27 [0.23, 0.31]	<.001
2022-2023	Canada (English)	0.23 (*n*=6,706)	-0.04 (*n*=11,281)	0.28 [0.24, 0.31]	<.001
	Canada (French)	0.23 (*n*=2,069)	-0.04 (*n*=1,793)	0.29 [0.23, 0.35]	<.001
	United States	0.17 (*n*=16,945)	-0.14 (*n*=21,616)	0.31 [0.29, 0.33]	<.001
	Australia	0.16 (*n*=1,586)	0.05 (*n*=2,934)	0.12 [0.06, 0.18]	<.001
2023-2024	Canada (English)	0.24 (*n*=6,454)	-0.05 (*n*=10,306)	0.31 [0.27, 0.34]	<.001
	Canada (French)	0.22 (*n*=2,016)	-0.06 (*n*=1,797)	0.31 [0.24, 0.37]	<.001
	United States	0.16 (*n*=15,411)	-0.14 (*n*=18,256)	0.31 [0.29, 0.33]	<.001
	Australia	0.19 (*n*=2,085)	-0.03 (*n*=3,287)	0.22 [0.17, 0.28]	<.001

Parental Education. Across all geographies and languages, applicants whose parents possessed at least a Bachelor's degree as their highest level of education tend to produce higher Casper scores relative to applicants whose parents did not possess a Bachelor's degree (Table 11). The size of this difference however is small with an average of *d*=0.19 (range: *d*=0.09 - 0.31) across the three most recent application cycles. In examining the regression analyses, each model produced a

positive regression coefficient (Appendix 2) which indicates that applicants whose parents possess a Bachelor's degree or more tend to produce higher Casper scores even when all other demographic variables were controlled for (β= 0.07 to 0.17, p<.05). However, the effect of this demographic variable was negligible (η_p^2 = 0.00) or small (η_p^2 = 0.01) across all models.

Table 11

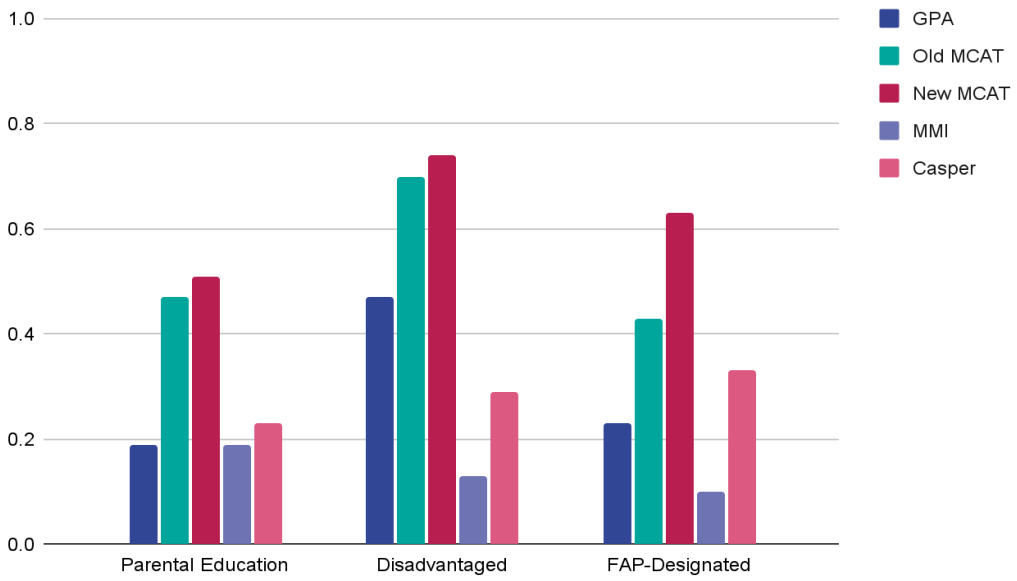
Mean Scores Of Applicants Whose Parents Possess A Bachelor's Degree Or Higher Compared To Mean Scores Of Applicants Whose Parents Do Not Possess A Bachelor's Degree

Application Year	Country	Parents with Bachelor's Degree or More	Parents with Less Than Bachelor's Degree	Cohen's *d* [95%CI]	*p*
2021-2022	Canada (English)	0.09 (*n*=20,731)	-0.06 (*n*=12,199)	0.15 [0.13, 0.17]	<.001
	Canada (French)	0.10 (*n*=4,770)	-0.08 (*n*=1,801)	0.18 [0.12, 0.23]	<.001
	United States	0.08 (*n*=45,713)	-0.20 (*n*=16,346)	0.29 [0.27, 0.31]	<.001
	Australia	0.10 (*n*=5,336)	0.01 (*n*=4,703)	0.09 [0.05, 0.13]	<.001
2022-2023	Canada (English)	0.09 (*n*=15,923)	-0.07 (*n*=9,138)	0.17 [0.14, 0.19]	<.001
	Canada (French)	0.11 (*n*=3,818)	-0.05 (*n*=1,471)	0.17 [0.11, 0.23]	<.001
	United States	0.08 (*n*=36,463)	-0.22 (*n*=12,587)	0.30 [0.28, 0.32]	<.001
	Australia	0.09 (*n*=3,932)	0.00 (*n*=3,252)	0.09 [0.05, 0.14]	<.001
2023-2024	Canada (English)	0.10 (*n*=15,419)	-0.08 (*n*=8,207)	0.18 [0.16, 0.21]	<.001
	Canada (French)	0.10 (*n*=3,795)	-0.14 (*n*=1,458)	0.24 [0.18,0.31]	<.001
	United States	0.09 (*n*=32,263)	-0.22 (*n*=11,080)	0.31 [0.29, 0.33]	<.001
	Australia	0.10 (*n*=5,199)	-0.06 (*n*=3,162)	0.16 [0.12, 0.21]	<.001

Casper Demographic Differences Relative to Other Admissions Metrics. In the 2019 study of 9,096 applicants to NYMCSM, three proxies of SES were examined: parental education (dichotomized in the same fashion as described above), self-declared disadvantaged status, and applicants who qualified for a fee assistance program (FAP; Juster et al., 2019). As depicted in Figure 28, it is evident that Casper produces lower group differences relative to both versions of the MCAT across all three SES variables. Casper group differences were similar to that of the MMI when using parental education as a proxy, but larger than MMI using the other two proxy methods. In relation to GPA, Casper evidenced similar group differences when using parental education as the proxy, smaller group differences when examining self-declared disadvantaged status as the proxy, and larger group differences when using qualification for FAP as the proxy.

Figure 28

Comparison Of SES Demographic Differences Across Admissions Metrics (Juster et al., 2019)



Age

Generally speaking, younger applicants tend to produce higher Casper scores than older applicants. Age of applicants has been dichotomized so that standardized mean differences can be calculated and assessed. As evidenced in Table 12, it is clear that across geographies and languages, applicants under the age of 28 produced higher Casper scores than applicants over the age of 28. On average for the three most recent applicant cycles, the size of this group difference is moderate (d=0.51, range: d=0.14 - 0.72). Results from regression analyses (Appendix 2) indicate that

across US, Canadian English, and Canadian French tests, older applicants (over the age of 28) tend to produce lower Casper scores relative to younger applicants, even when holding all other demographic variables constant (β= -0.21 to -0.33, p<.05), but the effect size is considered small across all models (η_p^2 = 0.02 - 0.03). Age was not statistically significant in the Australian regression model.

Table 12

Mean Scores of Applicants Under the Age of 28 Compared to Mean Scores of Applicants Over the Age of 28

Application Year	Country	Age 28 and Under	Over Age 28	Cohen's *d* [95%CI]	*p*
2021-2022	Canada (English)	0.11 (*n*=29,230)	-0.50 (*n*=4,336)	0.63 [0.60, 0.66]	<.001
	Canada (French)	0.12 (*n*=6,047)	-0.56 (*n*=614)	0.72 [0.63, 0.80]	<.001
	United States	0.05 (*n*=57,995)	-0.42 (*n*=5,200)	0.48 [0.45, 0.50]	<.001
	Australia	0.11 (*n*=8,817)	-0.21 (*n*=1,785)	0.33 [0.28, 0.38]	<.001
2022-2023	Canada (English)	0.10 (*n*=23,930)	-0.44 (*n*=3,345)	0.56 [0.52, 0.60]	<.001
	Canada (French)	0.12 (*n*=5,193)	-0.46 (*n*=530)	0.62 [0.53, 0.71]	<.001
	United States	0.05 (*n*=50,370)	-0.45 (*n*=4,625)	0.51 [0.48, 0.54]	<.001
	Australia	0.08 (*n*=6,844)	-0.16 (*n*=1,242)	0.26 [0.19, 0.32]	<.001
2023-2024	Canada (English)	0.09 (*n*=22,771)	-0.50 (*n*=2,660)	0.62 [0.58, 0.66]	<.001
	Canada (French)	0.09 (*n*=5,055)	-0.54 (*n*=534)	0.66 [0.57, 0.75]	<.001
	United States	0.06 (*n*=43,850)	-0.46 (*n*=3,859)	0.53 [0.50, 0.57]	<.001
	Australia	0.06 (*n*=7,766)	-0.08 (*n*=1,439)	0.14 [0.09, 0.20]	<.001

Rurality

Negligible-sized differences are observed between applicants who live in a rural or remote community (*i.e.,* a community with a population less than 10,000 people) and applicants who live in more urban communities (Table 13). Across the three most recent application cycles, the difference between these groups is consistently negligible in size with an average difference of d=0.08 (range: d=0.00-0.13). In examining the regression models, rurality has shown to have a negative (β= -0.06 to -0.19, p<.05), albeit, negligible-sized effect (η_p^2 = 0.00) on Casper scores when all other variables are held constant (Appendix 2).

Table 13

Mean Scores of Rural Applicants' Compared To Mean Scores of Non-Rural Applicants

Application Year	Country	Non-Rural	Rural	Cohen's *d* [95%CI]	*p*
2021-2022	Canada (English)	0.05 (*n*=27,611)	-0.06 (*n*=5,169)	0.11 [0.08, 0.14]	<.001
	Canada (French)	0.05 (*n*=5,392)	0.03 (*n*=989)	0.03 [-0.04, 0.09]	.430
	United States	0.03 (*n*=50,719)	-0.08 (*n*=10,486)	0.11 [0.09, 0.13]	<.001
	Australia	0.06 (*n*=8,313)	0.01 (*n*=1,383)	0.05 [0.00, 0.11]	.060
2022-2023	Canada (English)	0.05 (*n*=20,885)	-0.05 (*n*=3,995)	0.10 [0.07, 0.13]	<.001
	Canada (French)	0.06 (*n*=4,305)	0.06 (*n*=832)	0.00 [-0.07, 0.08]	.972
	United States	0.03 (*n*=40,013)	-0.10 (*n*=8,289)	0.13 [0.11, 0.15]	<.001
	Australia	0.06 (*n*=5,849)	-0.02 (*n*=1,067)	0.08 [0.02, 0.15]	.013
2023-2024	Canada (English)	0.05 (*n*=20,070)	-0.04 (*n*=3,432)	0.09 [0.06, 0.13]	<.001
	Canada (French)	0.04 (*n*=4,277)	-0.01 (*n*=829)	0.05 [-0.02, 0.13]	.122
	United States	0.03 (*n*=35,582)	-0.07 (*n*=7,048)	0.10 [0.07, 0.12]	<.001
	Australia	0.05 (*n*=7,061)	-0.06 (*n*=1,038)	0.11 [0.04, 0.17]	.001

Note. Rural refers to applicants who were raised in communities that had a population of less than 10,000 people.

Language

In order to assess group differences in performance across applicants with varying language backgrounds, applicants are asked two questions on the optional exit survey: (1) which language they primarily speak at home and (2) how they would rate their English (or French for French tests) proficiency level.

Language at Home. In order to calculate standardized mean difference values, this question has been segmented into two groups: applicants who primarily speak English at home and applicants who primarily speak another language at home (Table 14). For French tests, the same split is made, but into applicants who primarily speak French at home and applicants who primarily speak another language at home (Table 15). While applicants who primarily speak English at home tend to produce higher Casper scores relative to those who primarily speak another language at home, the magnitude of these group differences is unique across countries. In Canada and the United States, these differences tend to be small with an average of d=0.31 (range: d=0.29 - 0.34) for the three most recent application cycles. In Australia, differences are more pronounced with an average of d=0.65 (range: d=0.56-0.78) for the three most recent application cycles. Across all regression models (Appendix 2), speaking primarily another language at home had a negative effect on Casper scores across all geographies (β= -0.08 to -0.22, p<.001), but the effect size was small (η_p^2 = 0.00 - 0.02).

Table 14

Mean Scores Of Applicants Who Primarily Speak English At Home Compared To Mean Scores Of Applicants Who Primarily Speak Another Language At Home

Application Year	Country	Primarily Speak English at Home	Primarily Speak Another Language at Home	Cohen's *d* [95%CI]	*p*
2021-2022	Canada (English)	0.10 (*n*=26,804)	-0.23 (*n*=7,531)	0.34 [0.31, 0.36]	<.001
	United States	0.06 (*n*=52,285)	-0.25 (*n*=11,045)	0.31 [0.29, 0.33]	<.001
	Australia	0.18 (*n*=8,527)	-0.55 (*n*=1,721)	0.78 [0.72, 0.83]	<.001
2022-2023	Canada (English)	0.09 (*n*=20,067)	-0.20 (*n*=5,919)	0.31 [0.28, 0.34]	<.001
	United States	0.06 (*n*=40,465)	-0.23 (*n*=9,520)	0.29 [0.27, 0.31]	<.001
	Australia	0.14 (*n*=6,052)	-0.40 (*n*=1,225)	0.56 [0.50, 0.62]	<.001
2023-2024	Canada (English)	0.10 (*n*=18,886)	-0.18 (*n*=5,514)	0.29 [0.26, 0.32]	<.001
	United States	0.07 (*n*=35,874)	-0.22 (*n*=8,279)	0.29 [0.27, 0.32]	<.001
	Australia	0.16 (*n*=6,747)	-0.42 (*n*=1,717)	0.60 [0.54, 0.65]	<.001

For French tests, applicants who primarily spoke French at home produced higher Casper scores compared to those who primarily spoke another language at home, albeit, these differences were small (mean *d*=0.27; range: *d*=0.16-0.37). Regression analyses (Appendix 2) indicate that primarily speaking another language at home had a slight, yet negligible, positive impact on Casper scores when all other variables were held constant (β= 0.12, *p*=.046, η_p^2 = 0.00).

Table 15

Mean Scores Of Applicants Who Primarily Speak French At Home Compared To Mean Scores Of Applicants Who Primarily Speak Another Language At Home

Application Year	Country	Primarily Speak French at Home	Primarily Speak Another Language at Home	Cohen's *d* [95%CI]	*p*
2021-2022	Canada (French)	0.10 (*n*=5,552)	-0.26 (*n*=1,208)	0.37 [0.30, 0.43]	<.001
2022-2023	Canada (French)	0.11 (*n*=4,406)	-0.17 (*n*=1,017)	0.28 [0.21, 0.35]	<.001
2023-2024	Canada (French)	0.05 (*n*=4,365)	-0.11 (*n*=1,009)	0.16 [0.09, 0.23]	<.001

Language Proficiency. Similar to results presented above, applicants who reported that they were native for functionally native English speakers evidenced higher Casper scores than those who were non-native English speakers (Table 16). The magnitude of these group differences is consistent across countries with an average of *d*=0.63 (range: *d*=0.59-0.66) over the three most recent application cycles. Regression models (Appendix 2) indicated that being a non-native English speaker had a negative effect on Casper scores across all geographies (β= -0.38 to -0.42, *p*<.001), but the effect size for each was small (η_p^2 = 0.02 to 0.04).

Table 16

Mean Scores of Native English Speaking Applicants Compared to Mean Scores of Non-Native English Speaking Applicants

Application Year	Country	Native English Speakers	Non-Native English Speakers	Cohen's *d* [95%CI]	*p*
2021-2022	Canada (English)	0.20 (*n*=24,129)	-0.40 (*n*=9,207)	0.64 [0.61, 0.66]	<.001
	United States	0.07 (*n*=56,167)	-0.57 (*n*=6,029)	0.66 [0.63, 0.69]	<.001
	Australia	0.24 (*n*=6,784)	-0.36 (*n*=3,047)	0.63 [0.59, 0.68]	<.001
2022-2023	Canada (English)	0.20 (*n*=18,086)	-0.39 (*n*=6,956)	0.62 [0.59, 0.65]	<.001
	United States	0.08 (*n*=43,214)	-0.55 (*n*=5,712)	0.64 [0.62, 0.67]	<.001
	Australia	0.22 (*n*=4,841)	-0.34 (*n*=2,070)	0.59 [0.54, 0.65]	<.001
2023-2024	Canada (English)	0.19 (*n*=17,865)	-0.41 (*n*=5,845)	0.63 [0.60, 0.66]	<.001
	United States	0.08 (*n*=38,564)	-0.55 (*n*=4,680)	0.65 [0.62, 0.68]	<.001
	Australia	0.20 (*n*=5,980)	-0.40 (*n*=2,193)	0.64 [0.59, 0.69]	<.001

Similar results are evident when examining French proficiency as well (Table 17). Native or functionally native French speakers tend to produce higher scores relative to those who are non-native French speakers (mean *d*=0.47; range: *d*=0.41-0.57). The regression model (Appendix 2) indicated that being a non-native French speaker had a negative effect on Casper scores (β= -0.25, *p*<.001), but the effect size was small (η_p^2 = 0.01).

Table 17

Mean Scores of Native French Speaking Applicants Compared to Mean Scores of Non-Native French Speaking Applicants

Application Year	Country	Native French Speakers	Non-Native French Speakers	Cohen's *d* [95%CI]	*p*
2021-2022	Canada (French)	0.12 (*n*=5,568)	-0.42 (*n*=1,124)	0.57 [0.50, 0.63]	<.001
2022-2023	Canada (French)	0.14 (*n*=4,352)	-0.27 (*n*=993)	0.42 [0.35, 0.49]	<.001
2023-2024	Canada (French)	0.09 (*n*=4,419)	-0.31 (*n*=904)	0.41 [0.34, 0.49]	<.001

Employment

To determine if there are differences in performance between applicants of varying employment histories and experience, applicants are asked about their work experience in two ways: (1) "How many years of work experience do you have, *relevant* to the program you're applying to?" (2) "How many years of work experience do you have? This can be experience in any field of work." In pursuit of continually enhancing the quality of our post-test survey, this question has been the subject of change over several years. Up to and including the 2020-2021 application cycle, there was only one, more generic version of these questions which asked applicants "how much work experience do you have." This was adapted in the 2021-2022 application cycle to ask applicants about their work experience that was *relevant* to the program they were applying to. Finally, in the 2022-2023 application cycle, we determined that these questions were probing for unique applicant experiences, so we divided them into the two questions presented at the start of this section. Both of these questions have been parsed into two groups of applicants: (1) those with 10 years or less of work experience and (2) those with more than 10 years of work experience.

Program-Relevant Work Experience. Generally, there tend to be a minimal number of applicants who report to have over 10 years of work experience relevant to the program to which they are applying. This observation is perhaps not surprising given that a large majority of test takers are under the age of 30. Nonetheless, we wanted to understand if program-relevant experience provided an unintended benefit to these applicants on the Casper test. Interestingly, applicants who reported having more than ten years of program-relevant experience produced lower scores than those with less experience (Table 18). In Canada and the United states, these group differences tend to be moderate to large with an average of *d*=0.75 (range: 0.57

- 1.04). In Australia, group differences are less pronounced with difference values ranging from *d*=0.19 to *d*=0.34 for the three most recent application cycles. In examining the US and Canadian (English and French) regression models (Appendix 2), having more than 10 years of program-relevant work experience had a negative effect on Casper scores (β= -0.17 to -0.32, *p*<..05), but the effect size for each was negligible (η_p^2 = 0.00). In the Australian model, non-significant regression coefficients indicate that having more than 10 years of work experience did not have an effect on Casper scores.

Table 18

Mean Scores Of Applicants With 10 Years or Less of Program Relevant Work Experience Compared To Mean Scores Of Applicants With Over 10 Years

Application Year	Country	Applicants with 10 Years or Less Work Experience	Applicants with Over 10 Years Work Experience	Cohen's *d* [95%CI]	*p*
2021-2022	Canada (English)	0.04 (*n*=31,638)	-0.67 (*n*=836)	0.73 [0.66, 0.80]	<.001
	Canada (French)	0.06 (*n*=6,261)	-0.95 (*n*=123)	1.04 [0.86, 1.22]	<.001
	United States	0.01 (*n*=59,226)	-0.55 (*n*=983)	0.57 [0.51, 0.63]	<.001
	Australia	0.07 (*n*=9,139)	-0.22 (*n*=329)	0.29 [0.18, 0.40]	<.001
2022-2023	Canada (English)	0.04 (*n*=23,856)	-0.66 (*n*=528)	0.72 [0.63, 0.81]	<.001
	Canada (French)	0.08 (*n*=4,988)	-0.66 (*n*=118)	0.77 [0.59, 0.95]	<.001
	United States	0.01 (*n*=46,076)	-0.69 (*n*=699)	0.71 [0.64, 0.79]	<.001
	Australia	0.06 (*n*=6,702)	-0.28 (*n*=169)	0.34 [0.19, 0.50]	<.001
2023-2024	Canada (English)	0.04 (*n*=22,816)	-0.65 (*n*=392)	0.71 [0.61, 0.81]	<.001
	Canada (French)	0.05 (*n*=5,022)	-0.78 (*n*=110)	0.86 [0.67, 1.05]	<.001
	United States	0.02 (*n*=41,364)	-0.62 (*n*=666)	0.65 [0.58, 0.73]	<.001
	Australia	0.04 (*n*=7,881)	-0.15 (*n*=171)	0.19 [0.04, 0.35]	.024

General Work Experience. Similar to results regarding applicants' program-relevant work experience, applicants with 10 years or less of general work experience tend to produce higher Casper scores than those with more than 10 years of work experience (Table 19). Across Canada and the United States, group differences tended to be small in size (average *d*=0.37; range: *d*=0.33-0.44). While in

Australia, group differences were negligible with difference values of *d*=0.12 and *d*=-0.06 for the 2022-2023 and 2023-2024 application cycles, respectively.

Table 19

Mean Scores Of Applicants With 10 Years or Less of General Work Experience Compared To Mean Scores Of Applicants With Over 10 Years

Application Year	Country	Applicants with 10 Years or Less Work Experience	Applicants with Over 10 Years Work Experience	Cohen's *d* [95%CI]	*p*
2022-2023	Canada (English)	0.06 (*n*=22,255)	-0.26 (*n*=2,406)	0.33 [0.28, 0.37]	<.001
	Canada (French)	0.09 (*n*=4,780)	-0.34 (*n*=366)	0.44 [0.34, 0.55]	<.001
	United States	0.02 (*n*=44,259)	-0.30 (*n*=3,027)	0.33 [0.29, 0.37]	<.001
	Australia	0.06 (*n*=6,087)	-0.05 (*n*=850)	0.12 [0.05, 0.19]	.002
2023-2024	Canada (English)	0.06 (*n*=21,403)	-0.26 (*n*=1,961)	0.33 [0.28, 0.37]	<.001
	Canada (French)	0.06 (*n*=4,781)	-0.37 (*n*=384)	0.44 [0.33, 0.54]	<.001
	United States	0.04 (*n*=39,604)	-0.31 (*n*=2,733)	0.35 [.031, 0.39]	<.001
	Australia	0.03 (*n*=7,063)	0.09 (*n*=1,057)	-0.06 [-0.13, 0.00]	.060

Domestic or International Status

Group differences between applicants who identify as domestic students and international students are also examined, with results unique for each geography and language (Table 20). In the United States, group differences tend to be small in size (mean *d*=0.34; range: *d*=0.32 to 0.36), in Canada (English) and Australia, group differences tend to be moderate in size (mean *d*=0.62; range: *d*=0.43 to 0.77), and in French tests the differences are more pronounced (mean *d*=1.18; range: *d*=1.06 to 1.36). Regression models across Canada and Australia (Appendix 2) indicate that being an international student had a negative effect on Casper scores (β= -0.21 to -0.80, *p*<.05), but the effect size was negligible or small for each model (η_p^2 = 0.00 to 0.02). The

regression model for the United States indicated that being an international student had a negligible, yet positive, effect on Casper scores (β= 0.08, $p < .05$, $\eta_p^2 = 0.00$).

Table 20

Mean Scores of Domestic Applicants Compared to Mean Scores of International Applicants

Application Year	Country	Domestic Applicants	International Applicants	Cohen's *d* [95%CI]	*p*
2021-2022	Canada (English)	0.07 (*n*=33,553)	-0.68 (*n*=2,062)	0.77 [0.73, 0.82]	<.001
	Canada (French)	0.10 (*n*=6,555)	-0.91 (*n*=405)	1.06 [0.96, 1.16]	<.001
	United States	0.02 (*n*=64,369)	-0.33 (*n*=2,593)	0.35 [0.31, 0.39]	<.001
	Australia	0.12 (*n*=9,245)	-0.37 (*n*=1,627)	0.51 [0.46, 0.56]	<.001
2022-2023	Canada (English)	0.07 (*n*=27,327)	-0.63 (*n*=1,830)	0.71 [0.67, 0.76]	<.001
	Canada (French)	0.11 (*n*=5,680)	-0.94 (*n*=330)	1.11 [1.00, 1.22]	<.001
	United States	0.02 (*n*=55,734)	-0.33 (*n*=3,212)	0.36 [0.32, 0.39]	<.001
	Australia	0.10 (*n*=7,216)	-0.32 (*n*=1,162)	0.43 [0.37, 0.49]	<.001
2023-2024	Canada (English)	0.06 (*n*=25,706)	-0.64 (*n*=1,384)	0.72 [0.67, 0.78]	<.001
	Canada (French)	0.09 (*n*=5,598)	-1.19 (*n*=322)	1.36 [1.25, 1.48]	<.001
	United States	0.02 (*n*=48,896)	-0.29 (*n*=2,116)	0.32 [0.28, 0.37]	<.001
	Australia	0.09 (*n*=8,522)	-0.46 (*n*=1,197)	0.57 [0.51, 0.63]	<.001

Ability Status

Ensuring that the Casper test is safe and fair for applicants of all ability status is very important to us. To assess group differences between applicants who identify as a person with a disability and applicants who do not identify as a person with a disability, starting in the 2021-2022 application cycle, we began to ask applicants (in the optional post-test survey) to describe their ability status. In order to examine standardized mean differences, the results of this question have been dichotomized (Table 21). Across all geographies and languages, applicants who identify as a person with a disability tend to produce higher Casper scores relative to applicants who do not identify as a person with a disability. However, group differences have been consistently negligible in size (mean d=0.09; range: d= 0.03 to 0.15). In examining the regression models (Appendix 2), no models evidenced a statistically significant regression coefficient.

Table 21

Mean Scores of Applicants Who Identify as a Person with a Disability Compared to Mean Scores of Applicants Who Do Not Identify as a Person with a Disability

Application Year	Country	Applicants Who Identify as Person with a Disability	Applicants Who Do Not Identify as Person with a Disability	Cohen's *d* [95%CI]	*p*
2021-2022	Canada (English)	0.12 (*n*=3,050)	0.02 (*n*=30,605)	-0.10 [-0.13, -0.06]	<.001
	Canada (French)	0.12 (*n*=529)	0.03 (*n*=6,169)	-0.08 [-0.17, 0.00]	.071
	United States	0.08 (*n*=4,949)	0.01 (*n*=58,079)	-0.07 [-0.10, -0.05]	<.001
	Australia	0.19 (*n*=672)	0.05 (*n*=9,183)	-0.15 [-0.22, -0.07]	<.001
2022-2023	Canada (English)	0.13 (*n*=6,807)	0.01 (*n*=13,723)	-0.12 [-0.15, -0.09]	<.001
	Canada (French)	0.10 (*n*=1,224)	0.07 (*n*=3,087)	-0.03 [-0.10, 0.03]	.302
	United States	0.06 (n=11,053)	0.00 (*n*=30,116)	-0.06 [-0.08, -0.04]	<.001
	Australia	0.17 (*n*=1,770)	0.03 (*n*=4,117)	-0.14 [-0.20, -0.09]	<.001
2023-2024	Canada (English)	0.10 (*n*=6,807)	0.03 (*n*=13,191)	-0.08 [-0.10, -0.05]	<.001
	Canada (French)	0.06 (*n*=1,346)	0.03 (*n*=3,070)	-0.03 [-0.09, 0.04]	.391
	United States	0.07 (*n*=10,703)	0.01 (*n*=27,020)	-0.06 [-0.08, -0.03]	<.001
	Australia	0.16 (*n*=2,107)	0.04 (*n*=4,986)	-0.12 [-0.17, -0.07]	<.001

Race and Ethnicity

Despite a few exceptions, it is apparent that White or European applicants tend to produce higher Casper scores relative to applicants from other racial and ethnic backgrounds. While information is collected on a variety of racial identities, occasionally the sample size is too small to conduct and/or report statistical analyses which is why some comparisons are absent throughout this section.

Black, African, Caribbean, or African American v. White or European Applicants. Typically, the largest group differences when examining race tend to be observed between Black, African, Caribbean, or African American applicants and White or European applicants (Table 22). Across the three most recent application cycles, group differences tend to be moderate to large in size (mean d=0.79; range: d= -0.51 to d=-1.06). The regression models (Appendix 2) indicate that when holding all other demographic variables constant, there is a negative effect on Casper scores (β= -0.45 to -0.48, p<.01), but that the effect is negligible to small in size (η_p^2 = 0.00 to 0.03).

Table 22

Black, African, Caribbean, Or African American Applicants' Mean Scores Compared To White Or European Applicants' Mean Scores

Application Year	Country	White or European Applicants	Black, African, Caribbean, or African American Applicants	Cohen's *d* [95%CI]	*p*
2021-2022	Canada (English)	0.15 (*n*=18,190)	-0.58 (*n*=2,344)	-0.80 [-0.84, -0.76]	<.001
	Canada (French)	0.16 (*n*=4,830)	-0.74 (*n*=465)	-0.98 [-1.08, -0.89]	<.001
	United States	0.12 (*n*=35,896)	-0.56 (*n*=5,676)	-0.73 [-0.76, -0.70]	<.001
	Australia	0.24 (*n*=6,872)	-0.42 (*n*=149)	-0.74 [-0.90, -0.58]	<.001
2022-2023	Canada (English)	0.14 (*n*=14,682)	-0.57 (*n*=2,114)	-0.77 [-0.82, -0.73]	<.001
	Canada (French)	0.18 (*n*=3,946)	-0.79 (*n*=460)	-1.06 [-1.16, -0.96]	<.001
	United States	0.11 (*n*=30,436)	-0.59 (*n*=5,010)	-0.75 [-0.78, -0.72]	<.001
	Australia	0.16 (*n*=5,398)	-0.30 (*n*=79)	-0.51 [-0.73, -0.29]	<.001
2023-2024	Canada (English)	0.15 (*n*=13,110)	-0.58 (*n*=1,839)	-0.78 [-0.83, -0.73]	<.001
	Canada (French)	0.12 (*n*=3,889)	-0.78 (*n*=484)	-0.98 [-1.08, -0.88]	<.001
	United States	0.12 (*n*=25,844)	-0.48 (*n*=4,097)	-0.63 [-0.66, -0.60]	<.001
	Australia	0.20 (*n*=5,565)	-0.44 (*n*=133)	-0.69 [-0.87, -0.52]	<.001

Casper Demographic Differences Relative to Other Admissions Metrics. In the 2019 study comparing demographic differences of Casper and other admissions metrics at NYMCSM (*n*=9,096; Juster et al., 2019), analyses showed that Casper evidenced smaller demographic differences relative to three technical knowledge assessments, but not to the MMI (Figure 29). That fact that group differences are

evident for all of the metrics suggests that systemic issues may be contributing to these results. Additionally, in a more recent study across 7 physical therapy (PT) programs in the United States (*n*=3,747) from the 2021-2022 application cycle, Casper evidenced the smallest group differences relative to four other metrics: written GRE scores, total GPA, verbal GRE scores, and quantitative GRE scores (Figure 30). Although further improvements are needed, Casper seems to be least affected relative to the technical knowledge assessments and may serve to mitigate these group differences in the admissions process.

Figure 29

Comparison of Black, African, or African American And White Or European Applicants Across Admissions Metrics (Juster Et Al., 2019)

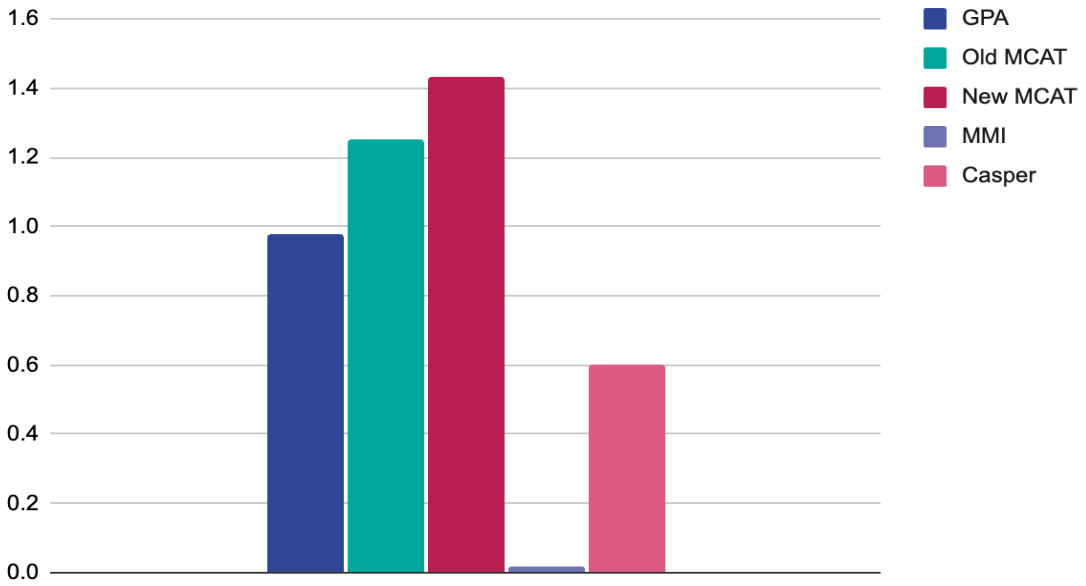
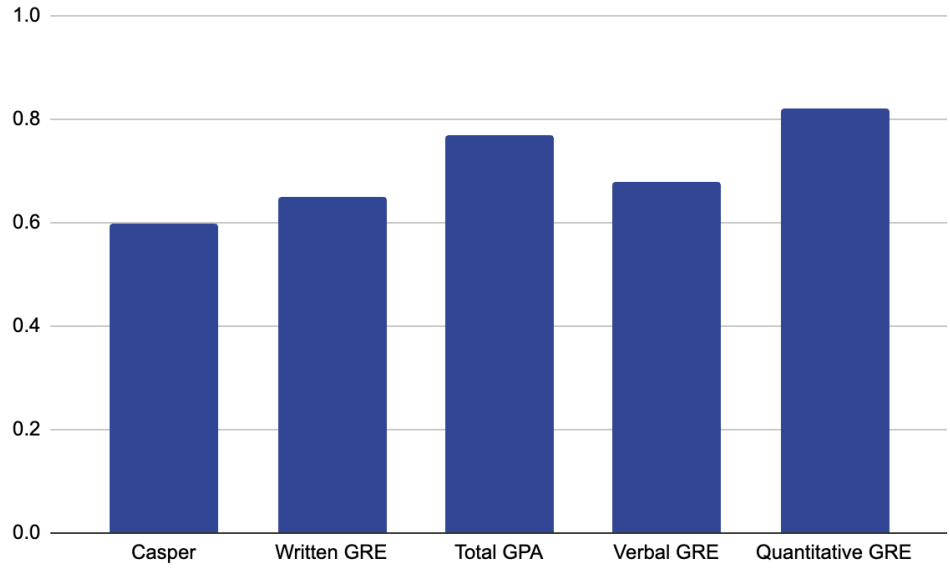


Figure 30

Comparison of Black, African, Or African American And White Or European Applicants Across 7 Physical Therapy Programs



Hispanic, Latinx, or Spanish origin v. White or European Applicants. As can be seen in Table 23, Hispanic, Latinx, or Spanish origin applicants tend to produce lower Casper scores relative to White or European applicants with a mean difference of *d*=0.40 (range from *d*=0.21 to *d*=0.55). When examining results from the regression analyses (Appendix 2), a negative effect was observed for Hispanic, Latinx, or Spanish origin applicants in the United States (β= -0.21, η_p^2 = 0.01, *p*<.001), but the the effect size is considered small. For the other three models (Australia, Canadian English, and Canadian French), regression coefficients were non-significant. This finding contradicts the results from the Cohen's *d* values in Table 23 which suggests that another factor, other than race, is likely influencing the difference in scores between these two groups, at least in Australian and Canadian tests.

Table 23

Hispanic, Latinx, Or Spanish Origin Applicants' Mean Scores Compared To White Or European Applicants' Mean Scores

Application Year	Country	White or European Applicants	Hispanic, Latinx, or Spanish origin Applicants	Cohen's *d* [95%CI]	*p*
2021-2022	Canada (English)	0.15 (*n*=18,190)	-0.25 (*n*=645)	-0.44 [-0.52, -0.37]	<.001
	Canada (French)	0.16 (*n*=4,830)	-0.34 (*n*=158)	-0.55 [-0.71, -0.40]	<.001
	United States	0.12 (*n*=35,896)	-0.27 (*n*=7,478)	-0.42 [-0.44, -0.39]	<.001
	Australia	0.24 (*n*=6,872)	-0.11 (*n*=91)	-0.39 [-0.59, -0.18]	.002
2022-2023	Canada (English)	0.14 (*n*=14,682)	-0.18 (*n*=569)	-0.35 [-0.44, -0.27]	<.001
	Canada (French)	0.18 (*n*=3,946)	-0.25 (*n*=148)	-0.49 [-0.65, -0.32]	<.001
	United States	0.11 (*n*=30,436)	-0.27 (*n*=6,785)	-0.40 [-0.43, -0.37]	<.001
	Australia	0.16 (*n*=5,398)	-0.30 (*n*=92)	-0.50 [-0.70, -0.29]	<.001
2023-2024	Canada (English)	0.15 (*n*=13,110)	-0.14 (*n*=503)	-0.31 [-0.40, -0.22]	<.001
	Canada (French)	0.12 (*n*=3,889)	-0.06 (*n*=121)	-0.21 [-0.39, -0.03]	.057
	United States	0.12 (*n*=25,844)	-0.25 (*n*=6,067)	-0.39 [-0.42, -0.36]	<.001
	Australia	0.20 (*n*=5,565)	-0.13 (*n*=108)	-0.36 [-0.55, -0.17]	.001

Casper Demographic Differences Relative to Other Admissions Metrics. The aforementioned NYMCSM study (Juster et al., 2019) also examined demographic differences between Hispanic or Latinx applicants and White or European applicants across several metrics. As can be seen in Figure 31, the Casper test produced the lowest demographic group differences relative to the MMI, GPA, and both versions of the MCAT (*n*=9,096). Additionally, in a 2021-2022 study across 7 physical therapy (PT)

programs in the United States (*n*=3,747), Casper evidenced the smallest group differences relative to four other metrics: written GRE scores, total GPA, verbal GRE scores, and quantitative GRE scores (Figure 32).

Figure 31

Comparison Of Hispanic Or Latinx And White Or European Applicants Across Admissions Metrics (Juster Et Al., 2019)

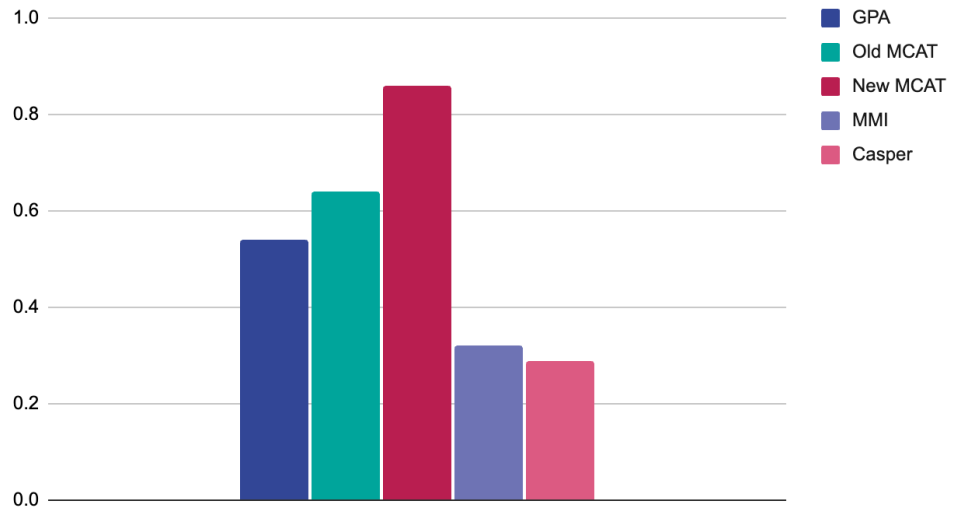
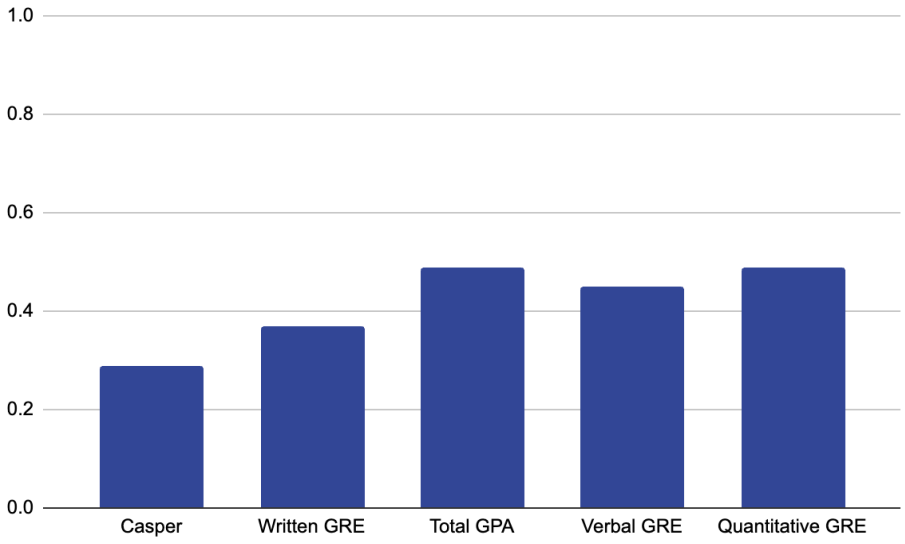


Figure 32

Comparison of Hispanic, Latinx, or Spanish Origin Applicants and White Or European Applicants Across 7 Physical Therapy Programs



Asian v. White or European Applicants. Across the three most recent application cycles, there is some variability in the group differences between Asian applicants and White or European applicants. In the United States, group differences are near-zero (mean *d*=0.03; range: *d*=0.01 to 0.06 (absolute)), in Canada (English and French), group differences tend to be negligible or small (mean *d*=0.15; range: *d*= 0.02 to 0.30 (absolute)), and in Australia, group differences tend to be moderate to large in size (mean *d*=0.47; range: *d*=0.36 to 0.61 (absolute)). Results from the regression analyses (Appendix 2) show that the regression coefficient produced from the Australian data was negative (***β***= -0.15, η_p^2 = 0.03 *p*<.001), but that the size of this effect was small. In the other three regression models, non-significant results were observed which suggests that the group differences observed in Table 24 are likely driven by demographic variables other than race of the applicant, at least in United States and Canadian data (English and French).

Table 24

Asian Applicants' Mean Scores Compared To White Or European Applicants' Mean Scores

Application Year	Country	White or European Applicants	Asian Applicants	Cohen's *d* [95%CI]	*p*
2021-2022	Canada (English)	0.15 (*n*=18,190)	0.00 (*n*=10,791)	-0.16 [-0.18, -0.14]	<.001
	Canada (French)	0.16 (*n*=4,830)	-0.11 (*n*=521)	-0.30 [-0.39, -0.21]	<.001
	United States	0.12 (*n*=35,896)	0.10 (*n*=14,919)	-0.02 [-0.04, -0.00]	.028
	Australia	0.24 (*n*=6,872)	-0.34 (*n*=2,830)	-0.61 [-0.65, -0.56]	<.001
2022-2023	Canada (English)	0.14 (*n*=14,682)	0.04 (*n*=8,752)	-0.10 [-0.13, -0.08]	<.001
	Canada (French)	0.18 (*n*=3,946)	0.05 (*n*=431)	-0.14 [-0.24, -0.04]	.006
	United States	0.11 (*n*=30,436)	0.11 (*n*=13,754)	-0.01 [-0.03, 0.01]	.573
	Australia	0.16 (*n*=5,398)	-0.18 (*n*=2,130)	-0.36 [-0.41, -0.31]	<.001
2023-2024	Canada (English)	0.15 (*n*=13,110)	0.00 (*n*=8,371)	-0.16 [-0.18, -0.13]	<.001
	Canada (French)	0.12 (*n*=3,889)	0.11 (*n*=388)	-0.02 [-0.12, 0.09]	.727
	United States	0.12 (*n*=25,844)	0.06 (*n*=11,762)	-0.06 [-0.08, -0.04]	<.001
	Australia	0.20 (*n*=5,565)	-0.22 (*n*=3,062)	-0.43 [-0.47, -0.38]	<.001

Indigenous v. White or European Applicants. Across geographies and languages, Indigenous applicants tend to evidenced lower Casper scores relative to White or European applicants (mean *d*=0.32; range: *d*=0.18 to *d*=0.45; Table 25). Results from the regression analyses (Appendix 2) indicate that across the United States and Canada (English) this demographic variable had a negative effect on Casper scores, but this effect was negligible for each model (***β***= -0.20 to -0.21, η_p^2 =

0.00, p<.05). In the Canadian French and Australian regression models, non-significant results were observed.

Table 25

Indigenous Applicants' Mean Scores Compared To White Or European Applicants' Mean Scores

Application Year	Country	White or European Applicants	Indigenous Applicants	Cohen's *d* [95%CI]	*p*
2021-2022	Canada (English)	0.15 (*n*=18,190)	-0.16 (*n*=835)	-0.34 [-0.41, -0.28]	<.001
	Canada (French)	0.16 (*n*=4,830)	-0.09 (*n*=30)	-0.28 [-0.64, 0.08]	.087
	United States	0.12 (*n*=35,896)	-0.30 (*n*=289)	-0.45 [-0.56, -0.33]	<.001
	Australia	0.24 (*n*=6,872)	-0.09 (*n*=85)	-0.37 [-0.58, -0.15]	.001
2022-2023	Canada (English)	0.14 (*n*=14,682)	-0.21 (*n*=713)	-0.38 [-0.46, -0.31]	<.001
	Canada (French)	0.18 (*n*=3,946)	-0.10 (*n*=34)	-0.32 [-0.66, 0.02]	.094
	United States	0.11 (*n*=30,436)	-0.17 (*n*=253)	-0.30 [-0.42, -0.18]	<.001
	Australia	0.16 (*n*=5,398)	-0.05 (*n*=83)	-0.23 [-0.45, -0.01]	.058
2023-2024	Canada (English)	0.15 (*n*=13,110)	-0.19 (*n*=564)	-0.37 [-0.46, -0.29]	<.001
	Canada (French)	0.12 (*n*=3,889)	-0.03 (*n*=50)	-0.18 [-0.45, 0.10]	.271
	United States	0.12 (*n*=25,844)	-0.20 (*n*=214)	-0.34 [-0.48, -0.21]	<.001
	Australia	0.20 (*n*=5,565)	-0.10 (*n*=101)	-0.33 [-0.53, -0.13]	.002

Note. In the Canadian context, Indigenous refers to Inuit, Métis, or Indigenous applicants, in the United States context, Indigenous refers to Indigenous American applicants, and in the Australian context, Indigenous refers to Torres Strait Islander and Māori applicants.

Middle Eastern or Northern African v. White or European Applicants. To date, the differences in Casper scores between Middle Eastern or Northern African applicants and White or European applicants has only been explored starting in the 2020-2021 application cycle. Based on the data available in Table 26, it is evident that Middle Eastern or Northern African applicants tend to produce lower Casper scores relative to White or European applicants, but the size of these differences are often classified as small (mean *d*=0.26; range: *d*=0.15 to 0.42). Across all regression models non-significant results were observed.

Table 26

Middle Eastern Or Northern African Applicants's Mean Scores Compared To White Or European Applicants' Mean Scores

Application Year	Country	White or European Applicants	Middle Eastern or Northern African Applicants	Cohen's *d* [95%CI]	*p*
2021-2022	Canada (English)	0.15 (*n*=18,190)	-0.16 (*n*=2,361)	-0.33 [-0.38, -0.29]	<.001
	Canada (French)	0.16 (*n*=4,830)	-0.01 (*n*=1,020)	-0.18 [-0.25, -0.11]	<.001
	United States	0.12 (*n*=35,896)	-0.11 (*n*=3,083)	-0.24 [-0.28, -0.20]	<.001
	Australia	0.24 (*n*=6,872)	-0.05 (*n*=433)	-0.32 [-0.41, -0.22]	<.001
2022-2023	Canada (English)	0.14 (*n*=14,682)	-0.16 (*n*=2,026)	-0.33 [-0.38, -0.28]	<.001
	Canada (French)	0.18 (*n*=3,946)	0.03 (*n*=952)	-0.17 [-0.24, -0.10]	<.001
	United States	0.11 (*n*=30,436)	-0.04 (*n*=2,875)	-0.16 [-0.20, -0.12]	<.001
	Australia	0.16 (*n*=5,398)	-0.23 (*n*=289)	-0.42 [-0.54, -0.30]	<.001
2023-2024	Canada (English)	0.15 (*n*=13,110)	-0.09 (*n*=2,036)	-0.25 [-0.30, -0.21]	<.001
	Canada (French)	0.12 (*n*=3,889)	-0.01 (*n*=924)	-0.15 [-0.22, -0.07]	<.001
	United States	0.12 (*n*=25,844)	-0.05 (*n*=2,427)	-0.18 [-0.22, -0.13]	<.001
	Australia	0.20 (*n*=5,565)	-0.13 (*n*=382)	-0.35 [-0.46, -0.25]	<.001

Native Hawaiian or Other Pacific Islander v. White or European Applicants. The difference in Casper scores between Native Hawaiian or Other Pacific Islander applicants and White or European applicants to date, has only been explored since the 2020-2021 application cycle. Due to low sample size, we are often unable to calculate this specific group difference. However, when possible, results tend to show negligible differences across Canada and the United States (mean

d=0.13; range: *d*=0.04 to 0.18 (absolute)) and moderate in Australia (mean *d*=0.55). In all countries, the regression coefficient for each was non-significant (Appendix 2).

Table 27

Native Hawaiian Or Other Pacific Islander Applicants' Mean Scores Compared To White Or European Applicants' Mean Scores

Application Year	Country	White or European Applicants	Native Hawaiian or Other Pacific Islander Applicants	Cohen's *d* [95%CI]	*p*
2021-2022	Canada (English)	0.15 (*n*=18,190)	-0.01 (*n*=22)	-0.18 [-0.60, 0.24]	.283
	Canada (French)	NA	NA	NA	NA
	United States	0.12 (*n*=35,896)	0.08 (*n*=178)	-0.04 [-0.19, 0.11]	.592
	Australia	0.24 (*n*=6,872)	-0.30 (*n*=61)	-0.61 [-0.86, -0.36]	<.001
2022-2023	Canada (English)	NA	NA	NA	NA
	Canada (French)	NA	NA	NA	NA
	United States	0.11 (*n*=30,436)	0.00 (*n*=122)	-0.12 [-0.30, 0.06]	.170
	Australia	0.16 (*n*=5,398)	-0.28 (*n*=58)	-0.48 [-0.74, -0.22]	.003
2023-2024	Canada (English)	NA	NA	NA	NA
	Canada (French)	NA	NA	NA	NA
	United States	0.12 (*n*=25,844)	-0.03 (*n*=107)	-0.16 [-0.35, 0.03]	.106
	Australia	NA	NA	NA	NA

Demographic Differences Summary

Demographic differences are consistently examined and reported across all geographies and languages in which the Casper test is administered. The

demographic differences observed for several variables often produce moderate effect sizes, as reflected in Cohen's d values. It is important to note that when demographic variables are examined in a multivariate context (regression analysis; Appendix 2), the effect sizes are uniformly negligible or small in magnitude as reflected in the partial eta squared statistic. These results indicate the complexity of how these variables interact together, and that the interaction or intersectionality between them is required to understand how these variables affect test scores like Casper.

Mitigating Test Bias - *Information on the steps Acuity Insights is taking to mitigate test-level bias within Casper.*

As evidenced in the aforementioned section on demographic differences, Casper tends to produce lower group differences in performance relative to other measures typically used in the admissions process. However, we do recognize that demographic differences are present within the Casper test and we are working diligently to reduce these as much as possible. This section provides information on how we are working to identify and combat group differences in performance.

Measurement Invariance. Measurement invariance (MI) is a statistical property of a test which, if established, indicates that the assessment measures the same construct(s) in the same manner across subgroups of applicants (Chen, 2008). At a high-level, MI is assessed by imposing a model onto each subgroup of applicants and evaluating the fit statistics (*e.g.,* CFI, RMSEA, and SRMR). This procedure is done in a stepwise fashion (4-step procedure) where, with each step, the model becomes more restrictive (*i.e.,* making it more difficult to evidence MI).

MI has been assessed across gender and race on two large historic (Casper with 12 typed-response items) test instances (n=2,650 and 2,332) of United States medical school applicants (Watters & Sitarenios, 2021). The model fit and difference statistics required for testing MI are displayed in Appendices 3 and 4 for gender (Test 1 and Test 2, respectively) and Appendices 5 and 6 for race (Test 1 and Test 2, respectively). Similar to baseline model fit, all increasingly restrictive multi-group CFA models achieved good fit (*i.e.,* CFI ≥ .95, RMSEA ≤ .05, SRMR ≤ .08). Furthermore, the difference in model fit did not change beyond the invariance threshold for any set of model comparisons (*i.e.,* change CFI ≤ -.01,RMSEA ≤ .015, and SRMR ≤ .03 for metric or ≤ .01 for scalar and residual levels). These results indicate that the same construct (social intelligence and professionalism) is being assessed in the same way across applicants from varying racial and gender identities.

Differential Item Functioning. Differential item functioning (DIF) is a method for measuring MI at an item-level. An item (*e.g.,* one of the scenarios of the Casper test) demonstrates DIF if it produces discrepancies in scores between groups of applicants (usually between a majority and minority group) who have the same ability level (Teresi & Fleishman, 2007). DIF allows researchers to identify items that

may be biased toward a certain subgroup of applicants. For the Casper test, items are evaluated for DIF across applicant race and gender.

To date, DIF has been assessed in over 2,000 items from 168 individual Casper test instances that occurred between 2018 and 2021 (see table 28 below). Using demographic information provided by applicants in a voluntary post-test survey, we compared the performance on these items along gender (self-identified as male vs. female) and racial lines (White vs. Black, African, Caribbean or African American/Hispanic, Latinx, or Spanish origin/Indigenous/Asian).

Ordinal logistic regression-based approaches were used to determine DIF. Based on Jodoin and Gierl's (2001) effect size criteria, which categorizes the magnitude of DIF into negligible, moderate and large, none of the items we tested exceeded a negligible amount of DIF. However, we flagged any items which exhibited even a negligible amount of DIF in at least two different test sittings. For those items with negligible DIF, the Acuity Insights team conducted a qualitative review to assess if any obvious signs of bias are present in the scenario or the wording of the questions.

As can be seen in Figure 33, the percentage of items that evidence DIF has been uniformly low across application cycles (less than 10%), and has continued to decrease in a linear fashion. This means that overall, Casper test items are fair across all groups of applicants.

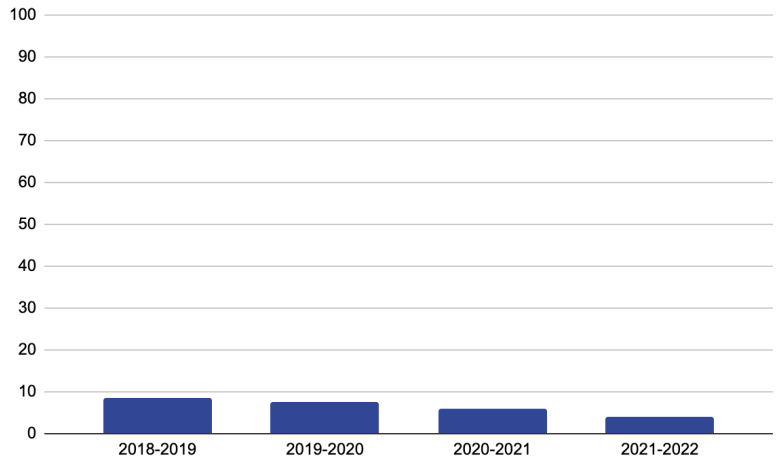
Table 28

Items Flagged By Dif Analysis Between 2018 – 2021

Reference Group	Focal Group	Magnitude of DIF	Number of items flagged	Percentage of items flagged
Male	Female	Not Significant	519	2.63%
		Negligible	14	
		Moderate	0	
White	Black or African American	Not Significant	1,992	0.00%
		Negligible	0	
		Moderate	0	
White	Hispanic, Latinx, or Spanish origin	Not Significant	519	0.76%
		Negligible	4	
		Moderate	0	
White	Asian	Not Significant	519	0.00%
		Negligible	0	
		Moderate	0	

Figure 33

Percentage Of Items That Evidenced DIF Across Each Application Cycle



Experimenting to Further Improve Equity- *How* Acuity Insights *made the decision to include a video response format*.

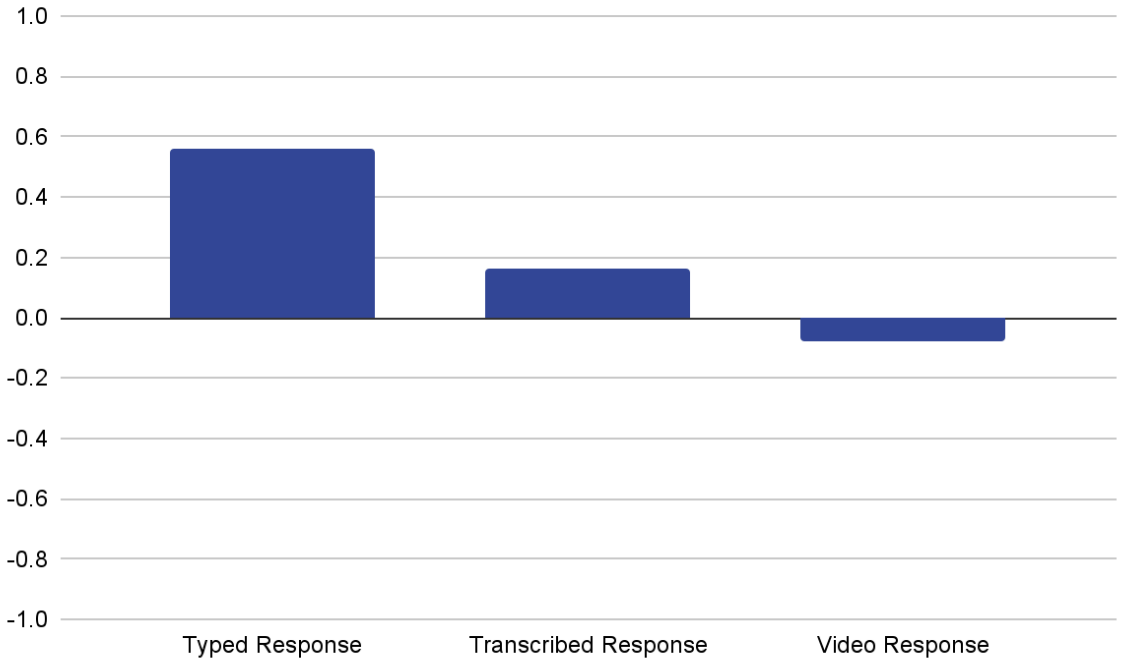
In pursuit of our goal to continually enhance the equity of the Casper test, it is not enough that we monitor the test in its current form, but that we continuously experiment with new approaches. This mindset is what drove the team at Acuity Insights to explore how a video-response component may enhance the equity of the test. Our interest in incorporating this new response format was sparked when several analyses indicated that the use of a video-response format resulted in reductions in the magnitude of demographic group differences. As an organization responsible for developing and administering such a high-stakes assessment, it was imperative that we ensured these results were generalizable and that the psychometric properties of the test met appropriate statistical thresholds before incorporating the video-response format permanently.

We are excited to share that based on several years of research, we decided to officially incorporate the video-response section into the Casper test starting in the 2023-2024 application cycle. Below, readers will find a brief history of our findings to date which support this decision.

2019 Study. In the 2019-2020 application cycle, two small pilot studies were conducted to determine if demographic differences in Casper scores would decrease if the response format changed from typed to video. This pilot study also examined how demographic differences would change if a transcribed response format was used. Applicants to United States Health Science programs voluntarily opted to take a 2-item video-response assessment after they completed their 12-item typed Casper test. The first pilot study compared scores between Black or African American applicants (n=157) and White applicants (n=217) and results (Figure 34) showed a dramatic reduction in Cohen's d effect size when response format changed from typed (d=0.56) to video (d=-0.08) or transcribed (d=0.16).

Figure 34

2019 Pilot Study 1: Demographic Group Differences Between Black Or African American Applicants And White Applicants Across Three Response Types

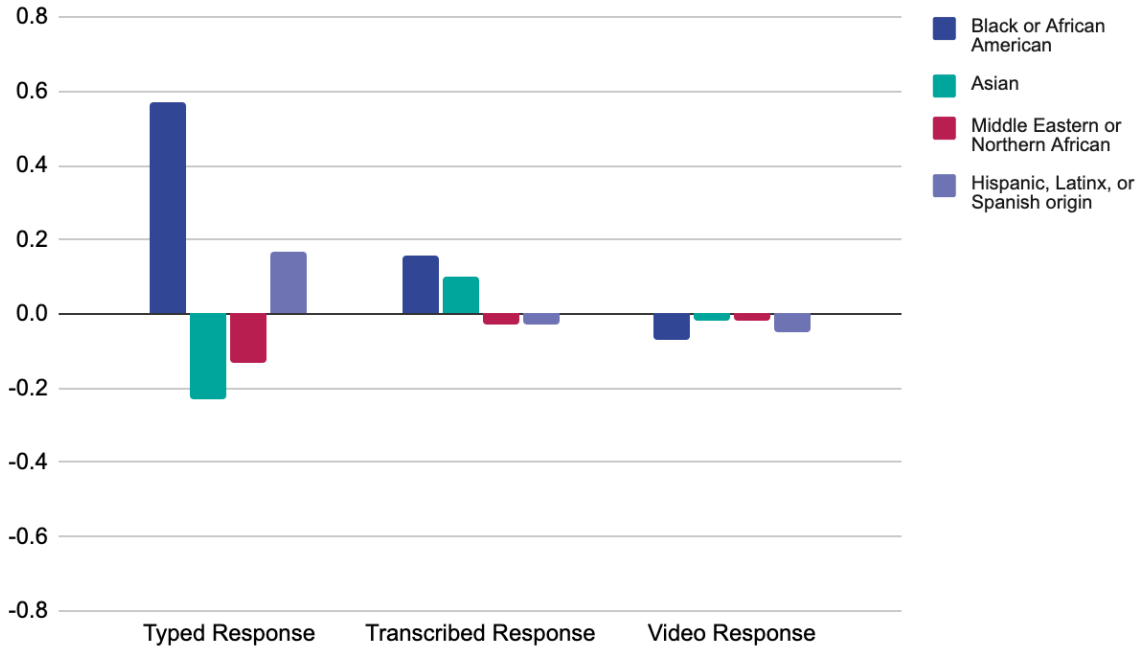


Note. Negative values indicate that the reference group (White applicants) produced lower Casper scores than the non-reference group.

The second pilot study examined demographic differences between White applicants (*n*=705) and (i) Asian applicants (*n*=304), (ii) Black or African American applicants (*n*=171), (iii) Hispanic, Latinx, or Spanish applicants (*n*=153), and (iv) Middle Eastern or Northern African applicants (*n*=47). Again, a dramatic reduction in Cohen's *d* values were observed when the response format was changed from typed to video or transcribed (see Figure 35). Namely, when altering the response format from typed to video, a reduction was observed between White applicants and (i) Asian applicants from *d*=-0.23 to *d*=-0.02), (ii) Black or African American applicants from *d*=0.57 to *d*=-0.07, (iii) Hispanic, Latinx, or Spanish applicants from *d*=0.17 to *d*=-0.05, and (iv) Middle Eastern or Northern African applicants from *d*=-0.13 to *d*=-0.02. These promising results subsequently led to a larger study in the 2021-2022 application cycle.

Figure 35

2019 Pilot Study 2: Demographic Group Differences Across Three Response Types



Note. Negative values indicate that the reference group (White applicants) produced lower Casper scores than the non-reference group.

2021 Study. The positive results evidenced in the 2019 pilot necessitated a larger-scale study to determine if the audio-visual responses were something that should be incorporated into the Casper test in the future. The 2021 study was designed not only to be larger in terms of the applicants permitted to take the optional video-response assessment, but also to address several of the limitations identified in the 2019 pilot. Specifically, the 2021 study used the entire pool of raters (the 2019 study only used top-performing raters), a large variety of Casper scenarios (the 2019 study only used 2 scenarios), and the 2-item video-response optional assessment was made available to every applicant who took the Casper test during the 2021-2022 application cycle (the 2019 study only allowed applicants to United States Health Science programs to take the video-response test).

Results from the 2021 study showed that when altering the response format from typed to video, demographic differences were dramatically reduced yet again (Figure 36). Demographic differences in scores were assessed across 7 variables: race, gender, English proficiency level, household income, community size, disability status, and military status. Overall, changing from a typed-response to a video-response format demonstrated reductions in the magnitude of demographic differences for a majority of group comparisons (Table 29). Using the video-response format, virtually all of these differences were reduced and the majority would be classified as negligible based on Cohen's d classification guidelines (Cohen, 1988).

Table 29

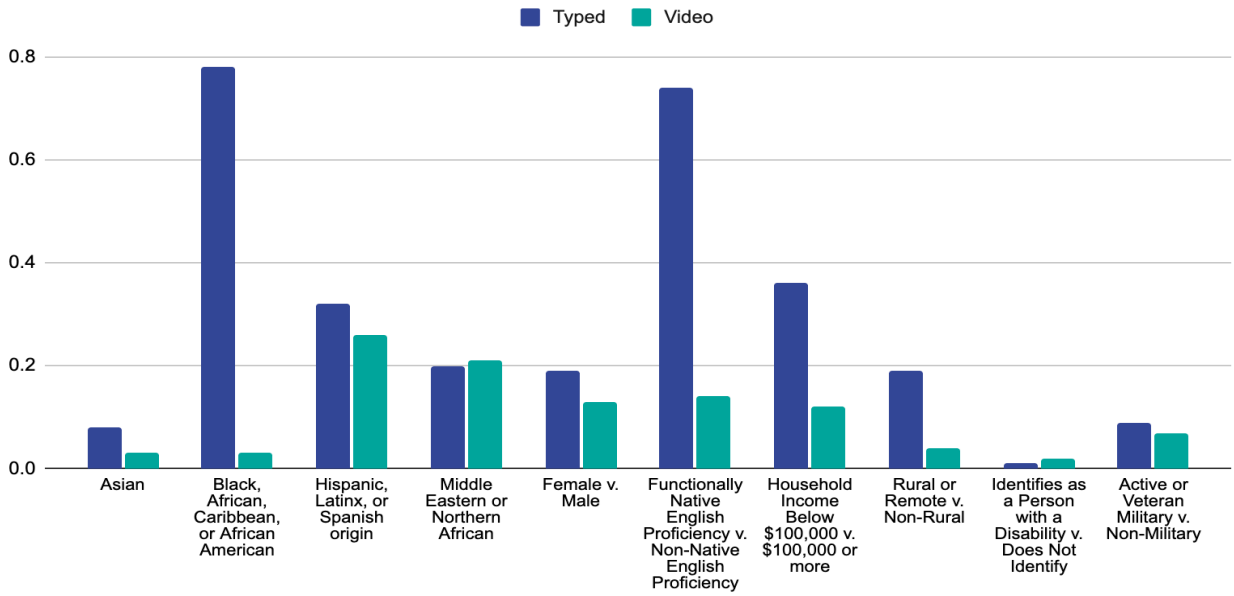
Demographic Differences Of Typed Responses And Video Responses In 2021 Study

Groups Compared	**Cohen's *d* (Typed)**	**Cohen's *d* (AVR)**
Asian v. White	0.08	0.03
Black, African, Caribbean, or African American v. White	0.78	0.03
Hispanic, Latinx, or Spanish origin v. White	0.32	0.26
Middle Eastern or Northern African v. White	0.20	0.21
Female v. Male	0.19	0.13
Functionally Native English Proficiency v. Non-Native English Proficiency	0.74	0.14
Household Income Below $100,000 v. $100,000 or more	0.36	0.12
Rural or Remote v. Non-Rural	0.19	0.04
Identifies as a Person with a Disability v. Does Not Identify	0.01	0.02
Active or Veteran Military v. Non-Military	0.09	0.07

Note. Guidelines for interpretation of Cohen's d: Negligible |d| < .2, Small |d| ≈ .2, Moderate |d| ≈ .5, Large |d| ≈ .8 (Cohen, 1988)

Figure 36

2021 Study: Demographic Group Differences Across Response Types



Note. Each racial group in the visualization is being compared White or European applicants.

2022 Study. Despite the incredibly promising findings from the 2019 and 2021 studies, the team at Acuity Insights wanted to ensure that the high-quality psychometric properties of the Casper test would be maintained with the inclusion of a new video-response section. Further, the team wanted to ensure that the promising results evidenced in the previous years were not a result of selection bias. That is, previous data reflected only those who *volunteered* to take the video-response assessment and may have done so because they were comfortable with sharing their thoughts and opinions orally. To ensure that the results would generalize, the Casper test for the 2022-2023 application cycle was changed such that applicants were required to complete a 15-item test. This test consisted of 9 typed-response scenarios and 6 video-response scenarios. We recognized that it was unfair and unethical to provide scores for such high-stakes decisions prior to building strong psychometric properties of the new test structure, thus the decision was made to provide programs with only the scores from the typed-response section (which had well established psychometric properties). That being said, select partner programs were provided with the information on the video-response section to aid in building validity evidence for this new test structure. The 2022 study was by far the largest and most comprehensive study to date. Data was collected from 18,685 applicants across 16 unique test instances from the United States, Canada, Australia, and New Zealand. Below, each piece of evidence to support the decision to officially incorporate the video-response component is detailed.

Reduced Test Sections. Notably, while examining the impact of the video-response section, the Acuity Insights team was simultaneously examining if the test length could be reduced. Our Research team conducted a series of in-depth

analyses and determined that a 14 scenario test (8 typed-response scenarios and 6 video-response scenarios) was the most appropriate structure. This structure ensures that excellent test reliability is maintained (reliability > 0.80) while simultaneously reducing the total test time for applicants. Based on this, the following data reflects a 14-item test to provide insight into how the test would function with the video-response section *and* reduced number of scenarios.

Score Calculation. With the inclusion of the video-response format, comes a new Casper score that now incorporates ratings from both typed-response and video-response scenarios. The Research team examined a variety of options for calculating each applicant's score now that there are two response formats. The team determined that the total raw score would be calculated by averaging the scores from all 14 scenarios. An applicant's raw score (as depicted below) will then be standardized within their unique test instance and each applicant will subsequently receive a z-score. The score provided, therefore, is exactly the same metric as it was in years past, but with the simple addition of applicants' scores from the video-response format.

Casper Raw Score *= ∑scenario scores / number of scenarios*

This total raw score calculation was chosen because it demonstrated an ability to maintain high reliability thresholds (reliability > 0.80) while simultaneously demonstrating a reduction in magnitude of demographic differences in group performance (relative to scores calculated using only a typed-response format). All of the analyses presented below reflect the new 14-scenario structure (8 typed-response sections and 6 video-response sections) and the new total score computation.

Descriptive Statistics. In general, applicants' mean and median Casper raw scores were similar across all test instances examined (see Table 30). Across all test instances, an average score of 5.31 and a median score of 5.36 was observed. In addition to the measures of central tendency, the skew and kurtosis of each test instance were also examined to determine if the scores formed a normal distribution. For context, a perfectly normal distribution will have a skew and kurtosis of 0, although it is extremely rare to see a perfectly normal distribution in applied research. Typical rules of thumb throughout the empirical literature suggest that absolute values of skewness and kurtosis greater than 1 and 3, respectively, indicate non-normality. Using these thresholds, it is clear that all tests examined demonstrated a normal distribution of scores.

Table 30

Descriptive Statistics Of The Total Score Across Test Instances

Program	n	Mean	Median	SD	Skew	Kurtosis
NZ Vet Sci 1	346	5.45	5.50	0.86	-0.51	0.27
AUS Med 2	339	5.70	5.79	0.76	-0.32	-0.28
US Med	1572	5.23	5.21	0.93	-0.11	-0.11
US HS2	1741	5.20	5.29	0.94	-0.29	-0.07
US Med	2143	5.22	5.29	0.90	-0.22	-0.26
US HS2	1244	5.11	5.14	0.99	-0.14	0.13
US Med	2312	5.34	5.36	0.89	-0.20	-0.08
US HS2	724	5.16	5.21	0.97	-0.16	-0.13
US Med	1440	5.31	5.36	0.92	-0.20	0.02
US HS2	647	5.17	5.14	0.97	-0.12	-0.34
US Med	2072	5.40	5.50	0.89	-0.32	0.19
US HS2	642	5.05	5.08	0.91	-0.26	-0.10
US Med	2010	5.26	5.31	0.86	-0.14	0.09
US HS2	684	5.14	5.15	0.91	-0.17	-0.04
CA HS2	143	5.58	5.79	1.15	-0.80	0.13
CA HS2	626	5.61	5.71	1.10	-0.54	0.01

Internal Consistency Reliability. The internal consistency reliability of the Casper test was evaluated using two different reliability estimates: coefficient alpha and omega (see Table 33). In general, internal consistency reliability measures the degree to which items on a test are consistent with each other in measuring the same underlying construct. This is achieved through assessment of the inter-relatedness of test items. Several methods of reporting internal consistency reliability exist, in which the underlying assumptions required of the data varies across methods. Although coefficient alpha is by far the most commonly used reliability estimate (Cronbach, 1951), it has limitations based predominantly on the restrictiveness of these underlying assumptions that in turn can lead to under or over estimation of reliability (Dunn et al., 2014). As such, researchers suggest reporting reliability estimates that are based on less strict assumptions of the data

alongside coefficient alpha, where alpha is reported due to its widespread familiarity and an additional estimate is reported to ensure the accuracy of the reliability estimates (Dunn et al., 2014). Although multiple estimates of reliability exist that do not require such strict underlying assumptions of the data in comparison to alpha, the most widely used estimate is coefficient omega (McDonald, 1999) and as such, is the reliability estimate that was chosen to complement alpha. As seen in Table 31, alpha and omega estimates were markedly similar, and test instances consistently demonstrated high levels of reliability that most often either met or exceeded a threshold of 0.80. More specifically, we observed an average coefficient alpha of 0.83 and an average omega of 0.85 across all test instances.

Table 31

Reliability Estimates Across Test Instances

Program	Coefficient Alpha	Omega Total
NZ Vet Sci 1	0.84	0.85
AUS Med 2	0.78	0.79
US Med	0.83	0.85
US HS2	0.82	0.83
US Med	0.81	0.83
US HS2	0.85	0.86
US Med	0.81	0.83
US HS2	0.85	0.87
US Med	0.83	0.85
US HS2	0.84	0.86
US Med	0.82	0.84
US HS2	0.82	0.83
US Med	0.79	0.82
US HS2	0.81	0.83
CA HS2	0.90	0.91
CA HS2	0.88	0.89
	Mean (Median)	**Mean (Median)**
	0.83 (0.83)	0.85 (0.85)

Demographic Group Differences. Demographic group differences in scores were assessed across 8 variables: race, gender, age, disability status, parental income (used as a proxy for applicant socio-economic status), community size, English proficiency level, and international/domestic student status. Group differences were assessed via standardized mean difference scores (*d*) which can be interpreted such that difference scores of 0.20, 0.50, and 0.80 correspond to small, moderate, and large effect sizes, respectively (Cohen, 1992). We provide demographic differences in test performance for both the typed-response section (scores used in previous years) and the combined Casper test score. As evidenced in Table 32, the combined score either reduces or maintains the magnitude of the demographic differences in group performance relative to scores from the typed-response sections alone.

Table 32

Comparison Of Demographic Differences Across All Tests

		Typed Response Score			Combined Score		
Group	**n**	**mean**	***d***	**Interpretation**	**mean**	***d***	**Interpretation**
Asian	3808	5.22	0.05	Negligible	5.34	0.01	Negligible
Black, African, Caribbean, or African American	735	4.33	0.77	Moderate/ High	4.86	0.54	Moderate
Hispanic, Latinx, or Spanish origin	1272	4.72	0.40	Small/ Moderate	4.96	0.42	Small/ Moderate
Middle Eastern or Northern African	666	4.99	0.16	Negligible	5.20	0.16	Negligible
White or European	7365	5.17	-	-	5.34	-	-
Female	9763	5.14	0.14	Negligible	5.34	0.21	Small
Male	4890	4.98			5.14		
28 Years Old and Under	12145	5.14	0.19	Negligible/ Small	5.31	0.18	Negligible/ Small
Over 28 Years Old	3587	4.93			5.14		
Identifies with a Disability	2389	5.11	0.02	Negligible	5.30	0.02	Negligible
Does not Identify with a Disability	7219	5.09			5.28		
Parental Income $100,00 or More	4514	5.26	0.31	Small	5.41	0.31	Small
Parental Income Under $100,000	4484	4.90			5.13		
Rural or Remote Community Size	1600	4.94	0.15	Negligible	5.15	0.16	Negligible
Non-Rural Community Size	9667	5.11			5.29		
Non-Native English Speaker	1027	4.29	0.77	Moderate/ Large	4.65	0.75	Moderate/ Large
Native English Speaker	10432	5.17			5.33		
International Applicant	623	4.69	0.36	Small	5.06	0.25	Small
Domestic Applicant	13301	5.11			5.29		

Note. Performance of each racial group is being compared to the performance of White or European applicants.
Note. This table represents information collected from 16 test instances from all geographic regions.
Note. Guidelines for interpretation of Cohen's d: Negligible |d| < .2, Small |d| ≈ .2, Moderate |d| ≈ .5, Large |d| ≈ .8 (Cohen, 1988)

2023 Study. While enhancing fairness and equity within the test was the primary driving force of this initiative, it was equally important to establish strong validity evidence for the new format, particularly predictive validity. To do this we partnered with two US Medical Schools to assess the extent to which scores from the new Casper format could predict future behaviour.

Admissions data was collected from 1,011 applicants who were invited to participate in an MMI at either UTMB John Sealy School of Medicine or Rutgers Robert Wood Johnson Medical School. These applicants had also taken the updated Casper test which included both the typed-response and video-response format.

To evaluate the extent to which Casper scores could predict performance on a similar measure, the MMI, a single-predictor bivariate logistic regression model was fitted for each school. Results should that for every one-unit increase in raw Casper score (scale from 1-9), odds of receiving a high MMI score (as determined by each program) increased by 159.5% for UTMB (OR: 2.56, 95%CI[2.07, 3.30]) and by 66.78% for Rutgers (OR: 1.67, 95%CI[1.08, 2.65]).

Key Terms

- ***Application Cycle.*** Application cycle refers to the time at which applicants write the Casper test to prepare their application package. Each application cycle straddles two calendar years with several dates available to write the Casper test so that applicants can submit their scores prior to application deadlines. For example, an applicant who plans to attend a program in September 2024 would write their Casper test in the 2023-2024 application cycle.
- ***Vertical.*** Each Casper test is tailored to program type and geographic location; the term *vertical* is used to describe each program in each geography. For example, United States Undergraduate Medicine programs, Canadian Occupational Therapy programs, and Australian Teachers Education programs are all distinct verticals with unique content.
- ***Test Instance.*** A test instance refers to each unique test date. To ensure accessibility of the test, several test dates are available to applicants within each vertical, thus each program receives Casper scores from multiple test instances. The content of each test instance is unique to ensure applicants do not have access to test material ahead of time.

References

Albanese, M. A., Snow, M. H., Skochelak, S. E., Huggett, K. N., & Farrell, P. M. (2003). Assessing personal qualities in medical school admissions. *Academic Medicine, 78*(3), 313-321. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1049.4723&rep=rep1&type=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.

Association of American Medical Colleges. (2020). *Using MCAT Data in 2021 Medical Student Selection*. https://www.aamc.org/system/files/2020-07/services_mcat_using-mcat-data-in-2021-medical-student-selection-guide_07082020_0.pdf

Association of American Medical Colleges. (2021). *Core competencies for entering medical students*. *https://www.aamc.org/services/admissions-lifecycle/competencies-entering-medical-students*

Australian Institute for Teaching and School Leadership (2017). *Australian professional standards for teachers*. https://www.aitsl.edu.au/teach/standards

Ajzen, I. (1985). *F*rom intentions to actions: A theory of planned behavior. In J. Kuhl & J. Beckmann (Eds.), *Action control: From cognition to behavior* (pp.11-39). Springer Berlin Heidelberg. doi:10.1007/978-3-642-69746-3_2

Boston University. (2023). *Boston University Chobanian & Avedisian School of Medicine*. https://www.bumc.bu.edu/busm/

Brown, T., & Bonsaksen, T. (2019). An examination of the structural validity of the Physical Self-Description Questionnaire-Short Form (PSDQ-S) using the Rash Measurement Model. *Cogent Education, 6*(1). https://www.tandfonline.com/doi/epub/10.1080/2331186X.2019.1571146

Burgess, R., Barber, C., Mountjoy, M., Whyte, R., Vanstone, M., & Grierson, L. (2020, September 7-9). *Evaluating the validity evidence of admissions and medical school performance variable on student outcomes at McMaster University* [Conference presentation]. AMEE 2020 Virtual Conference. https://amee.org/getattachment/Conferences/AMEE-Past-Conferences/AMEE-2020/AMEE-2020-Virtual-Abstract-Book-FINAL-resize.pdf

Burgos, L. M., DeLima, A. A., Parodi, J., Costabel, J. P., Ganiele, M. N., Durante, E., Arceo, M. D., & Gelpi, R. (2020). Reliability and acceptability of the multiple mini-interview for selection of residents in cardiology. *Journal of Advances in Medical Education and Professionalism, 8*(1), 25-31. doi: https://dx.doi.org/10.30476%2Fjamp.2019.83903.1116

Busche, K., Elks, M. L., Hanson, J. T., Jackson-Williams, L., Manuel, R. S., Parsons, W. L., Wofst, D., & Yuan, K. (2020). The validity of scores from the new MCAT exam in predicting student performance: Results from a multisite study. *Academic Medicine*, *95*(3), 387-395. doi: 10.1097/ACM.0000000000002942

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, *56*(2), 81. doi: 10.1037/h0046016

Carlson, K. D., & Herdman, D. O. (2012). Understanding the impact of convergent validity on research results. *Organizational Research Methods, 15*(1). 17-32. doi:10.1177/1094428110392383

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95*, 1005–1018.

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

Cohen, B. (2013). *Explaining psychological statistics* (4th ed.). John Wiley & Sons.

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

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334. http://cda.psych.uiuc.edu/psychometrika_highly_cited_articles/cronbach_1951.pdf

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, *52*(4), 281.

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

Dalhousie University (2023). *Rowe School of Business: About*. https://www.dal.ca/faculty/management/rsb/about.html

Dore, K. L, Reiter, H. I., & Baskwill, A. (2016, August 27-31). *Online situational judgement tests: Implication and perspectives of group test taking in Casper.*

[Conference presentation]. Barcelona, Spain. https://amee.org/getattachment/Conferences/AMEE-2016/AMEE-2016-App-Data/Session-8-k.pdf

Dore, K, L., Reiter, H. I., Eva, K. W., Krueger, S., Scriven, E., Siu, E., Hilsden, S., Thomas, J., & Norman, G. R. (2009). Extending the interview to all medical school candidates-Computer-Based Multiple Sample Evaluation of Noncognitive Skills (CMSENS). *Academic Medicine, 84*(10), S9-S12. doi: 10.1097/ACM.0b013e3181b3705a

Dore, K. L., Reiter, H. I., Kreuger, S., Norman, G. R. (2017). CASPer, an online pre-interview screen for personal/professional characteristics: Prediction of national licensure scores. *Advances in Health Sciences Education, 22*(2), 327-336. doi: 10.1007/s10459-016-9739-9

Dowell, J., Lynch, B., Till, H., Kumwenda, B., & Husbands, A. (2012). The multiple mini-interview in the UK context: 3 years of experience at Dundee. *Medical Teacher, 34*(4), 297-304. doi:10.3109/0142159X.2012.652706

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British journal of psychology, 105*(3), 399-412. http://irep.ntu.ac.uk/id/eprint/4853/1/215051_Dunn.pdf

Eva, K. W., Reiter, H. I., Rosenfeld J, Norman G. (2004). The relationship between interviewers' characteristics and ratings assigned during a multiple mini-interview. *Academic Medicine, 79*(6), 602–609.

Eva, K. W., Reiter, H. I., Rosenfeld, J., Trinh, K., Wood, T. J., Norman, G. R. (2012). Association between a medical school admission process using the multiple mini-interview and national licensing examination scores. *American Medical Association, 308*(21), 2233-2240. doi: 10.1001/jama.2012.36914

Eva, K. W., Rosenfeld, J., Reiter, H. I., & Norman, G. R. (2004b). An admissions OSCE: The multiple mini-interview. *Medical Education, 38*, 314-326. doi: 10.1046/j.1365-2923.2004.01776.x

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods*, *4*(3), 272. http://personal.psu.edu/jxb14/M554/articles/Fabrigaretal1999.pdf

Frey, B. (2018). *The SAGE encyclopedia of educational research, measurement, and evaluation* (Vols. 1-4). SAGE Publications, Inc. doi: 10.4135/9781506326139

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.

Gauer, J. L., Wolff, J. M., & Jackson, B. (2016). Do MCAT scores predict USMLE scores? An analysis on 5 years of medical student data. *Medical Education Online, 21*(1), 1-7. doi:10.3402/meo.v21.31795

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

Haider, S. I., Bari, M. F., & Ijaz, S. (2020). Using multiple mini-interviews for students' admissions in Pakistan: a pilot study. *Advances in medical education and practice*, *11*, 179-185. doi: https://dx.doi.org/10.2147%2FAMEP.S246285

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238-247. doi: 10.1037/1040-3590.7.3.238

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational research methods*, *7*(2), 191-205. https://journals.sagepub.com/doi/pdf/10.1177/1094428104263675

Hecker, K., Donnon, T., Fuentealba, C., Hall, D., Illanes, O., Morck, D. W., & Muelling, C. (2009). Assessment of applicants to the veterinary curriculum using a multiple mini-interview method. *Journal of Veterinary Medical Education*, *36*(2), 166-173. doi: 10.3138/jvme.36.2.166

Henning, C. T., Chapman, R .I., MacIntosh, A., Sitarenios, G., Parker, J. D.A. (2023, June, 23-25). *Elite performance on a text-based situational judgment test for medical school admissions: Relationships with emotional and social competency* [Printed Poster]. Canadian Psychological Association, Toronto, Ontario, Canada.

Hofstra University. (2021). Hofstra University: *About Hofstra.* https://www.hofstra.edu/about/

Howard, M. C. (2016). A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human-Computer Interactions*, *32*, 51-62. doi:10.1080/10447318.2015.1087664

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

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

Jerant, A., Griffin, E., Rainwater, J., Henderson, M., Sousa, F., Bertakis, K. D., Fenton, J. J., Franks, P. (2012). Does applicant personality influence multiple mini-interview performance and medical school acceptance offers? *Academic Medicine, 87*(9), 1250–1259. doi: 10.1097/ACM.0b013e31826102ad

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied measurement in education*, *14*(4), 329-349.

Juster, F. R., Baum, R. C., Zou, C., Risucci, D., Ly, A., Reiter, H., Miller, D.D., Dore, K.L. (2019). Addressing the diversity-validity dilemma using situational judgement tests. *Academic Medicine, 94*(8), 1197-1203. doi: 10.1097/ACM.0000000000002769

Kent and Medway Medical School. (2021). *About*. https://kmms.ac.uk/about/

Kline, P. (2014). *An easy guide to factor analysis*. Routledge.

Knorr, M., & Hissbach, J. (2014). Multiple mini-interviews: Same concept, different approaches. *Medical Education, 48*, 1157-1175. doi: 10.1111/medu.12535

Koo, T. K., Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Kulasegaram, K., Reiter, H. I., Wiesner, W., Hackett, R. D., & Norman, G. R. (2010). Non-association between Neo-5 personality tests and multiple mini-interview. *Advances in Health Science Education, 15*(3), 415-423. doi:10.1007/s10459-009-9209-8

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., 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

Mahon, K. E., Henderson, M. K.., & Kirch, D. G. (2013). Selecting tomorrow's physicians: The key to the future health care workforce. *Academic Medicine, 88*(12), 1806-1811. doi:10.1097/ACM.0000000000000023

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=H3

yIazT_5csAAAAA:xdjdFTYgH-PED9PrGYPsC_iJt3extBOuZs7KRlfF2GEW5lFKTgOxc6loBjgUb6KfZs4KOicl_8GyhKG2tw

McDonald, R. P. (1999). Test theory: A unified treatment. Lawrence Erlbaum Associates. doi:10.1111/j.2044-8317.1981.tb00621.x

McMaster University. (2021). *Postgraduate Medical Education*. https://pgme.mcmaster.ca/

Monash University. (2012). *Monash at a Glance*. https://www.monash.edu/about/who/glance

Mortaz Hejri, S., Ivan, R., & Jama. N. (2022). Assessment through a cross-cultural lens in North American higher education. *Frontiers in Education, 7*. doi:10.3389/feduc.2022.1012722

Moskowitz, J. B., Yan, W., Ho, J. L., Robb, C., MacIntosh, A., Sitarenios, G. (2022, November 11-15). *Projecting validity and reliability for shortened educational assessments* [Oral Presentation Submission]. AAMC Learn Serve Lead 2022. Nashville, Tennessee, United States.

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgements. *Journal of Personal and Social Psychology, 35*(4), 250-256. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/92158/TheHaloEffect.pdf?sequence=1

O'Brien, A., Harvey, J., Shannon, M., Lewis, K., & Valencia, O. (2011). A comparison of multiple mini-interviews and structured interviews in a UK setting. *Medical Teacher, 33*(5), 397-402. doi:10.3109/0142159X.2010.541532

Oliver, T., Hecker, K., Hausdorf, P. A., Conlon, P. (2014). Validating MMI scores: Are we measuring multiple attributes? *Advances in Health Sciences Education, 19*, 379-392. doi:10.1007/s10459-013-9480-6

Papadakis, M, A., Teherani, A. T., Banach, M. A., Knettler, T. R., Rattner, S. L., Stern, D. T. Veloski, J. J., & Hodgson, C. S. (2005). Disciplinary action by medical boards and prior behavior in medical school. *The New England Journal of Medicine, 353*(25), 2673-2682.

Parker, J. D. A. (2022). *Multidimensional Inventory of Personal Intelligence (MIPI): Technical manual*. Adaptimist Insights.

Parker-Newlyn, L., Mansfield, L., & Dore, K. L. (2019, July 1-4). *Future directions in selection: Pilot outcomes of a video-based online SJT in Australian medical student selection* [Conference Presentation]. ANZAHPE Conference, Canberra, Australia. https://anzahpe.org/resources/Documents/Conference/Past%20Conference%20documentation/2019%20Proceedings.pdf

Patterson, F., Zibarras, L., & Ashworth, V. (2016). Situational judgement tests in medical education and training: Research, theory and practice: AMEE Guide No. 100. *Medical Teacher*, *38*(1), 3-17. https://.doi.org/10.3109/0142159X.2015.1072619

Price, L. R. (2017). Psychometric Methods: Theory into Practice. The Guilford Press.

Rees, E. L., Hawarden, A. W., Dent, G., Hays, R., Bates, J., & Hassell, A. B. (2016). Evidence regarding the utility of multiple mini-interview (MMI) for selection to undergraduate health programs: A BEME systematic review: BEME guide No. 37. *Medical Teacher, 38*(5), 443-455. doi:10.3109/0142159X.2016.1158799

Roberts, C., Clark.T., Burgess, A., Frommer, M., Grant, M., & Mossman, K. (2014). The validity of a behavioural multiple-mini-interview within an assessment centre for selection into specialty training. *BMC Medical Education, 14*(169). doi: 10.1186/1472-6920-14-169

Roberts, C., Walton, M., Rothnie, I., Crossley, J., Lyon, P., Kumar, K., & Tiller, D. (2008). Factors affecting the utility of the multiple mini-interview in selecting candidates for graduate-entry medical school. *Medical Education, 42*, 396-404. doi:10.1111/j.1365-2923.2008.03018.x

Royal College of Physicians and Surgeons of Canada. (2021). *CanMEDS: Better standards, better physicians, better care*. *https://www.royalcollege.ca/rcsite/canmeds/canmeds-framework-e*

Sebok, S. S., Luu, K., & Klinger, D. A. (2014). Psychometric properties of the multiple mini-interview used for medical admissions: Findings from generalizability and Rasch analysis. *Advances in Health Science Education, 19*, 71-84. doi: 10.1007/s10459-013-9463-7

Shipper, E. S., Mazer, L. M., Merrell S.B., Lin, D. T., Lau, J. N., Melcher, M. l. (2017). Pilot evaluation of the computer-based assessment for sampling personal characteristics test. *Journal of Surgical Research, 215*, 211-218. doi: 10.1016/j.jss.2017.03.054

Smith, E., & Reeves, R. V. (2020). "SAT Math scores mirror and maintain racial inequity". US Front, Brookings. https://www.brookings.edu/blog/up-front/2020/12/01/sat-math-scores-mirror-and-maintain-racial-inequity/

Teresi, J. A., & Fleishman, J. A. (2007). Differential item functioning and health assessment. *Quality of Life Research*, *16*(1), 33-42. https://link.springer.com/content/pdf/10.1007/s11136-007-9184-6.pdf

Texas A&M University (2021). *About the College of Medicine.* https://medicine.tamu.edu/about/index.html

The Trustees of Indiana University (2021). *Indiana University School of Dentistry*. https://dentistry.iu.edu/about/index.html

Tulane University (2021). *School of Medicine*. https://medicine.tulane.edu/

University of Alberta (2021). *Faculty of Rehabilitation Medicine: Department of Occupational Therapy*. https://www.ualberta.ca/occupational-therapy/index.html

University of Evansville (2021). *Academic Programs: Physician Assistant*. https://www.evansville.edu/majors/physicianassistant/goals.cfm

University of Illinois College of Medicine (2023). *Our Mission*. https://medicine.uic.edu/

University of Ottawa (2021). *Undergraduate Medical Education*. https://med.uottawa.ca/undergraduate/admissions

University of Wollongong (2021). *Medicine*. https://www.uow.edu.au/science-medicine-health/schools-entities/medicine/disciplines/medicine/

Van Eck, M. E., Lameijer, C. M., & El Moumni, M. (2018). Structural validity of the Dutch version of the disability of arm, shoulder and hand questionnaire (DASH-DLV) in adult patients with hand and wrist injuries. *BMC musculoskeletal disorders*, *19*(1), 1-10.

Watters, C., & Sitarenios, G. (2021). Measurement invariance across gender and ethnicity: Extending psychometric evidence for equitability of the Casper test [Conference presentation]. AERA 2022 Annual Meeting, San Diego, California, United States.

Weir, J. P. (2005). Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. *Journal of Strength and Conditioning Research, 19*(1), 231-240. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.457.2590&rep=rep1&type=pdf

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

Whitcomb K.M., Cwik S., Singh C. (2021). "Not All Disadvantages Are Equal: Racial/Ethnic Minority Students Have Largest Disadvantage Among Demographic Groups in Both STEM and Non-STEM GPA". AERA Open. doi:10.1177/23328584211059823

Williams, T. S. (1983). Some issues in the standardized testing of minority students. *The Journal of Education, 165*(2), 192-208. https://www.jstor.org/stable/pdf/42772833.pdf?refreqid=excelsior%3A82f5ca73da2119aa4ca9b9499cbe2f6c

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

Yingling, S., Park, Y. S., Curry, R. H., Monson, V., Girotti, J. (2018). Beyond cognitive measures: Empirical evidence supporting holistic medical school admissions practices and professional identity formation. *MedEdPublish 7*(4), 1-11. https://doi.org/10.15694/mep.2018.0000274.1

Zou, C., McConnell, M., Leddy, J., Antonacci, P., & Lemay G. (2018). Comparison of the English and French versions of the Casper test in a bilingual population. *MedEdPublish, 7*(4), 1-10. https://doi.org/10.15694/mep.2018.0000281.1

Appendix 1
Demographic Group Differences For Previous Years

Mean Scores of Female Applicants Compared to Mean Scores of Male Applicants

Application Year	Country	Female	Male	Cohen's *d* [95%CI]	*p*
2018-2019	Canada (English)	0.03 (*n*=15,544)	-0.07 (*n*=6,554)	0.10 [0.07, 0.13]	<.001
	Canada (French)	0.07 (*n*=4,277)	-0.14 (*n*=1,700)	0.20 [-0.26, -0.15]	<.001
	United States	0.05 (*n*=24,110)	-0.07 (*n*=18,680)	0.13 [0.11, 0.15]	<.001
	Australia	0.03 (*n*=7,569)	-0.06 (*n*=2,730)	0.08 [0.04, 0.13]	<.001
2019-2020	Canada (English)	0.03 (*n*=17,950)	-0.08 (*n*=6,749)	0.11 (0.09, 0.14)	<.001
	Canada (French)	0.07 (*n*=4,535)	-0.14 (*n*=1,778)	0.21 [0.16, 0.27]	<.001
	United States	0.04 (*n*=31,157)	-0.07 (*n*=21,029)	0.12 [0.10, 0.14]	<.001
	Australia	0.02 (*n*=7,112)	-0.01 (*n*=2,548)	0.03 [-0.01, 0.07]	.170
2020-2021	Canada (English)	0.05 (*n*=25,656)	-0.08 (*n*=8,618)	0.13 [0.11, 0.16]	<.001
	Canada (French)	0.08 (*n*=5,173)	-0.11 (*n*=1,929)	0.19 [o.14, 0.24]	<.001
	United States	0.08 (*n*=47,088)	-0.13 (*n*=25,911)	0.21 [0.19, 0.22]	<.001
	Australia	0.05 (*n*=8,651)	-0.05 (*n*=3,274)	0.10 [0.06, 0.14]	<.001

Mean Scores Of Household Incomes Of $100,000 Or More Compared To Mean Scores Of Household Incomes Below $100,000

Application Year	Country	Household Income Above $100,000	Household Income Below $100,000	Cohen's *d* [95%CI]	*p*
2018-2019	Canada (English)	0.18 (*n*=7,877)	-0.11 (*n*=11,324)	0.29 [0.26, 0.32]	<.001
	Canada (French)	0.18 (*n*=2,549)	-0.10 (*n*=2,742)	0.29 [0.24, 0.34]	<.001
	United States	0.13 (*n*=21,189)	-0.17 (*n*=17,624)	0.30 [0.28, 0.32]	<.001
	Australia	0.15 (*n*=3,191)	-0.10 (*n*=5,366)	0.25 [0.20, 0.29]	<.001
2019-2020	Canada (English)	0.21 (*n*=7,661)	-0.08 (*n*=11,994)	0.29 [0.26. 0.32]	<.001
	Canada (French)	0.19 (*n*=2,608)	-0.11 (*n*=2,603)	0.31 [0.26, 0.37]	<.001
	United States	0.15 (*n*=22,384)	-0.16 (*n*=22,088)	0.31 [0.29, 0.33]	<.001
	Australia	0.15 (*n*=2,405)	-0.07 (*n*=4,570)	0.22 [0.17, 0.26]	<.001
2020-2021	Canada (English)	0.23 (*n*=11,442)	-0.08 (*n*=15,572)	0.32 [0.29, 0.34]	<.001
	Canada (French)	0.21 (*n*=3,121)	-0.09 (*n*=2,683)	0.31 [0.26, 0.36]	<.001
	United States	0.17 (*n*=30,416)	-0.16 (*n*=31,497)	0.34 [0.32, 0.36]	<.001
	Australia	0.21 (*n*=3,951)	-0.09 (*n*=5,088)	0.31 [0.27, 0.35]	<.001

Mean Scores Of Applicants Whose Parents Possess A Bachelor's Degree Or Higher Compared To Mean Scores Of Applicants Whose Parents Do Not Possess A Bachelor's Degree

Application Year	Country	Parents with Bachelor's Degree Applicants	Parents without Bachelor's Degree Applicants	Cohen's *d* [95%CI]	*p*
2018-2019	Canada (English)	0.04 (*n*=13,939)	-0.08 (*n*=6,988)	0.12 [0.09, 0.15]	<.001
	Canada (French)	0.06 (*n*=4,175)	-0.08 (*n*=1,603)	0.14 [0.08, 0.20]	<.001
	United States	0.07 (*n*=32,016)	-0.26 (*n*=9,239)	0.34 [0.31, 0.36]	<.001
	Australia	0.06 (*n*=4,652)	-0.03 (*n*=4,888)	0.09 [0.05, 0.13]	<.001
2019-2020	Canada (English)	0.06 (*n*=14,797)	-0.07 (*n*=7,909)	0.13 [0.10, 0.15]	<.001
	Canada (French)	0.07 (*n*=4,227)	-0.07 (*n*=1,709)	0.14 [0.08, 0.20]	<.001
	United States	0.07 (*n*=37,353)	-0.25 (*n*=12,017)	0.33 [0.31, 0.35]	<.001
	Australia	0.06 (*n*=4,361)	-0.02 (*n*=4,112)	0.08 [0.10, 0.15]	<.001
2020-2021	Canada (English)	0.08 (*n*=19,178)	-0.03 (*n*=11,834)	0.12 [0.10, 0.14]	<.001
	Canada (French)	0.11 (*n*=4,766)	-0.09 (*n*=1,801)	0.21 [0.15, 0.26]	<.001
	United States	0.09 (*n*=49,839)	-0.21 (*n*=18,342)	0.30 [0.29, 0.32]	<.001
	Australia	0.09 (*n*=5,773)	-0.01 (*n*=5,174)	0.10 [0.06, 0.14]	<.001

Mean Scores of Applicants Under the Age of 28 Compared to Mean Scores of Applicants Over the Age of 28

Application Year	Country	Age 28 and Under Applicants	Over Age 28 Applicants	Cohen's *d* [95%CI]	*p*
2018-2019	Canada (English)	0.08 (*n*=19,119)	-0.54 (*n*=3.020)	0.64 [0.60, 0.68]	<.001
	Canada (French)	0.16 (*n*=75)	-1.18 (*n*=10)	1.50 [0.79, 2.21]	.03
	United States	0.04 (*n*=39,624)	-0.43 (*n*=3,226)	0.47 [0.44, 0.51]	<.001
	Australia	0.04 (*n*=8,259)	-0.12 (*n*=2,052)	0.17 [0.11, 0.21]	<.001
2019-2020	Canada (English)	0.09 (*n*=21,567)	-0.59 (*n*=3,263)	0.71 [0.67, 0.74]	<.001
	Canada (French)	0.08 (*n*=5,710)	-0.66 (*n*=613)	0.76 [0.64, 0.84]	<.001
	United States	0.05 (*n*=47,894)	-0.51 (*n*=4,607)	0.57 [0.54, 0.61]	<.001
	Australia	0.06 (*n*=7,908)	-0.21 (*n*=1,779)	0.28 [0.22, 0.33]	<.001
2020-2021	Canada (English)	0.13 (*n*=26,736)	-0.51 (*n*=4,286)	0.67 [0.64, 0.70]	<.001
	Canada (French)	0.11 (*n*=5,817)	-0.59 (*n*=613)	0.75 [0.66, 0.83]	<.001
	United States	0.06 (*n*=58,391)	-0.38 (*n*=6,378)	0.45 [0.43, 0.48]	<.001
	Australia	0.10 (*n*=8,707)	-0.23 (*n*=2,102)	0.34 [0.29, 0.39]	<.001

Mean Scores of Rural Applicants' Compared To Mean Scores of Non-Rural Applicants

Application Year	Country	Non-Rural Applicants	Rural Applicants	Cohen's *d* [95%CI]	*p*
2018-2019	Canada (English)	0.02 (*n*=18,226)	-0.10 (*n*=2,776)	0.12 [0.08, 0.16]	<.001
	Canada (French)	0.02 (*n*=4,841)	-0.03 (*n*=900)	0.05 [-0.02, 0.12]	.110
	United States	0.02 (*n*=35,200)	-0.13 (*n*=5,970)	0.15 [0.12, 0.18]	<.001
	Australia	0.00 (*n*=8,016)	0.07 (*n*=1,543)	-0.07 [-0.13, -0.02]	.004
2019-2020	Canada (English)	0.03 (*n*=19,689)	-0.07 (*n*=3,230)	0.11 [0.07, 0.14]	<.001
	Canada (French)	0.03 (*n*=4,786)	-0.03 (*n*=995)	0.06 [0.01, -0.12]	.060
	United States	0.02 (*n*=41,450)	-0.15 (*n*=7,545)	0.18 [0.15, 0.20]	<.001
	Australia	0.00 (*n*=7,045)	0.05 (*n*=1,463)	-0.05 [-0.11, 0.01]	.070
2020-2021	Canada (English)	0.05 (*n*=26,372)	-0.02 (*n*=4,751)	0.07 [0.04, 0.10]	<.001
	Canada (French)	0.07 (*n*=5,358)	-0.03 (*n*=1,036)	0.11 [0.04, 0.18]	.001
	United States	0.03 (*n*=56,409)	-0.09 (*n*=10,981)	0.12 [0.10, 0.14]	<.001
	Australia	0.05 (*n*=8,959)	0.00 (*n*=1,671)	0.05 [0.00, 0.10]	.041

Mean Scores Of Applicants Who Primarily Speak English At Home Compared To Mean Scores Of Applicants Who Primarily Speak Another Language At Home

Application Year	Country	Primarily Speak English at Home	Primarily Speak Another Language at Home	Cohen's *d* [95%CI]	*p*
2018-2019	Canada (English)	0.08 (*n*=16,698)	-0.29 (*n*=4,847)	0.38 [0.35, 0.41]	<.001
	United States	0.03 (*n*=34,630)	-0.13 (*n*=7,280)	0.16 [0.13, 0.18]	<.001
	Australia	0.19 (*n*=8,107)	-0.77 (*n*=1,881)	1.04 [0.98, 1.09]	<.001
2019-2020	Canada (English)	0.10 (*n*=18,394)	-0.29 (*n*=5,717)	0.40 [0.36, 0.43]	<.001
	United States	0.02 (*n*=41,761)	-0.11 (*n*=9,262)	0.13 [0.11, 0.15]	<.001
	Australia	0.18 (*n*=7,613)	-0.75 (*n*=1,642)	0.99 [0.93, 1.04]	<.001
2020-2021	Canada (English)	0.12 (*n*=25,545)	-0.28 (*n*=7,339)	0.42 [0.39, 0.45]	<.001
	United States	0.06 (*n*=57,347)	-0.23 (*n*=12,865)	0.30 [0.28, 0.32]	<.001
	Australia	0.17 (*n*=9,499)	-0.64 (*n*=1,950)	0.86 [0.81, 0.91]	<.001

Mean Scores Of Applicants Who Primarily Speak French At Home Compared To Mean Scores Of Applicants Who Primarily Speak Another Language At Home

Application Year	Country	Primarily Speak French at Home	Primarily Speak Another Language at Home	Cohen's *d* [95%CI]	*p*
2018-2019	Canada (French)	0.08 (*n*=4,891)	-0.32 (*n*=1,011)	0.41 [0.34, 0.48]	<.001
2019-2020	Canada (French)	0.09 (*n*=5,040)	-0.33 (*n*=1,130)	0.42 [0.36, 0.49]	<.001
2020-2021	Canada (French)	0.09 (*n*=5,592)	-0.20 (*n*=1,247)	0.30 [0.24, 0.36]	<.001

Mean Scores of Native English Speaking Applicants Compared to Mean Scores of Non-Native English Speaking Applicants

Application Year	Country	Native English Speakers	Non-Native English Speakers	Cohen's *d* [95%CI]	*p*
2020-2021	Canada (English)	0.21 (*n*=22,407)	-0.38 (*n*=9,250)	0.63 [0.61, 0.66]	<.001
	United States	0.08 (*n*=61,688)	-0.59 (*n*=7,352)	0.69 [0.67, 0.72]	<.001
	Australia	0.24 (*n*=7,221)	-0.40 (*n*=3,351)	0.68 [0.64, 0.72]	<.001

Mean Scores of Native French Speaking Applicants Compared to Mean Scores of Non-Native French Speaking Applicants

Application Year	Country	Native French Speakers	Non-Native French Speakers	Cohen's *d* [95%CI]	*p*
2020-2021	Canada (French)	0.13 (*n*=5,540)	-0.39 (*n*=1,174)	0.55 [0.49, 0.62]	<.001

Mean Scores Of Applicants With 10 Years or Less of Program Relevant Work Experience Compared To Mean Scores Of Applicants With Over 10 Years

Application Year	Country	Applicants with 10 Years or Less Work Experience	Applicants with Over 10 Years Work Experience	Cohen's *d* [95%CI]	*p*
2021-2022	Canada (English)	0.04 (*n*=31,638)	-0.67 (*n*=836)	0.73 [0.66, 0.80]	<.001
	Canada (French)	0.06 (*n*=6,261)	-0.95 (*n*=123)	1.04 [0.86, 1.22]	<.001
	United States	0.01 (*n*=59,226)	-0.55 (*n*=983)	0.57 [0.51, 0.63]	<.001
	Australia	0.07 (*n*=9,139)	-0.22 (*n*=329)	0.29 [0.18, 0.40]	<.001
2022-2023	Canada (English)	0.04 (*n*=23,856)	-0.66 (*n*=528)	0.72 [0.63, 0.81]	<.001
	Canada (French)	0.08 (*n*=4,988)	-0.66 (*n*=118)	0.77 [0.59, 0.95]	<.001
	United States	0.01 (*n*=46,076)	-0.69 (*n*=699)	0.71 [0.64, 0.79]	<.001
	Australia	0.06 (*n*=6,702)	-0.28 (*n*=169)	0.34 [0.19, 0.50]	<.001

Mean Scores Of Applicants With 10 Years or Less of General Work Experience Compared To Mean Scores Of Applicants With Over 10 Years

Application Year	Country	Applicants with 10 Years or Less Work Experience	Applicants with Over 10 Years Work Experience	Cohen's *d* [95%CI]	*p*
2018-2019	Canada (English)	0.03 (*n*=18,910)	-0.25 (*n*=2,296)	0.29 [0.24, 0.33]	<.001
	Canada (French)	0.04 (*n*=5,521)	-0.44 (*n*=299)	0.49 [0.37, 0.60]	<.001
	United States	0.03 (*n*=36,368)	-0.34 (*n*=3,456)	0.37 [0.34, 0.41]	<.001
	Australia	-0.01 (*n*=8,154)	0.07 (*n*=1,665)	-0.07 [-0.13, -0.02]	.005
2019-2020	Canada (English)	0.04 (*n*=20,755)	-0.28 (*n*=2,478)	0.33 [0.28, 0.37]	<.001
	Canada (French)	0.05 (*n*=5,553)	-0.39 (*n*=432)	0.45 [0.35, 0.54]	<.001
	United States	0.04 (*n*=44,893)	-0.41 (*n*=4,335)	0.45 [0.42, 0.48]	<.001
	Australia	0.06 (*n*=6,762)	0.00 (*n*=1,494)	0.07 [0.01, 0.12]	.020
2020-2021	Canada (English)	0.08 (*n*=27,439)	-0.24 (*n*=4,028)	0.33 [0.29, 0.36]	<.001
	Canada (French)	0.09 (*n*=6,096)	-0.48 (*n*=495)	0.59 [0.50, 0.68]	<.001
	United States	0.04 (*n*=61,039)	-0.25 (*n*=6,465)	0.29 [0.26, 0.31]	<.001
	Australia	0.05 (*n*=9,180)	-0.01 (*n*=1,800)	0.06 [0.01, 0.11]	.015

Mean Scores of Domestic Applicants Compared to Mean Scores of International Applicants

Application Year	Country	Domestic Applicants	International Applicants	Cohen's *d* [95%CI]	*p*
2020-2021	Canada (English)	0.07 (*n*=31,260)	-0.59 (*n*=2,064)	0.68 [0.64, 0.73]	<.001
	Canada (French)	0.09 (*n*=6,557)	-0.90 (*n*=334)	1.03 [0.92, 1.15]	<.001
	United States	0.03 (*n*=67,234)	-0.31 (*n*=3,850)	0.34 [0.31, 0.37]	<.001
	Australia	0.10 (*n*=10,102)	-0.51 (*n*=1,304)	0.63 [0.57, 0.69]	<.001

Mean Scores of Applicants Who Identify as a Person with a Disability Compared to Mean Scores of Applicants Who Do Not Identify as a Person with a Disability

Application Year	Country	Applicants Who Identify as Person with a Disability	Applicants Who Do Not Identify as Person with a Disability	Cohen's *d* [95%CI]	*p*
2021-2022	Canada (English)	0.12 (*n*=3,050)	0.02 (*n*=30,605)	-0.10 [-0.13, -0.06]	<.001
	Canada (French)	0.12 (*n*=529)	0.03 (*n*=6,169)	-0.08 [-0.17, 0.00]	.071
	United States	0.08 (*n*=4,949)	0.01 (*n*=58,079)	-0.07 [-0.10, -0.05]	<.001
	Australia	0.19 (*n*=672)	0.05 (*n*=9,183)	-0.15 [-0.22, -0.07]	<.001
2022-2023	Canada (English)	0.13 (*n*=6,807)	0.01 (*n*=13,723)	-0.12 [-0.15, -0.09]	<.001
	Canada (French)	0.10 (*n*=1,224)	0.07 (*n*=3,087)	-0.03 [-0.10, 0.03]	.302
	United States	0.06 (n=11,053)	0.00 (*n*=30,116)	-0.06 [-0.08, -0.04]	<.001
	Australia	0.17 (*n*=1,770)	0.03 (*n*=4,117)	-0.14 [-0.20, -0.09]	<.001

Black, African, Caribbean, Or African American Applicants' Mean Scores Compared To White Or European Applicants' Mean Scores

Application Year	Country	White or European Applicants	Black, African, Caribbean, or African American Applicants	Cohen's *d* [95%CI]	*p*
2018-2019	Canada (English)	0.11 (n=10,059)	-0.76 (n=970)	-0.94 [-1.01, -0.87]	<.001
	United States	0.06 (n=20,857)	-0.54 (n=2,834)	-0.62 [-0.66, -0.58]	<.001
	Australia	0.23 (n=6,371)	-0.63 (n=57)	-0.99 [-1.25, -0.73]	<.001
2019-2020	Canada (English)	0.14 (*n*=11,694)	-0.72 (*n*=1,320)	-0.96 [-1.02, -0.90]	<.001
	Canada (French)	0.14 (*n*=4,257)	-0.93 (*n*=254)	-1.18 [-1.31, -1.05]	<.001
	United States	0.04 (*n*=26,378)	-0.51 (*n*=4,102)	-0.57 [-0.61, -0.54]	<.001
	Australia	0.22 (*n*=6,395)	-0.54 (*n*=53)	-0.87 [-1.14, -0.59]	<.001
2020-2021	Canada (English)	0.17 (*n*=17,078)	-0.63 (*n*=1,842)	-0.88 [-0.93, -0.83]	<.001
	Canada (French)	0.15 (*n*=4,750)	-0.80 (*n*=349)	-1.04 [-1.15, -0.93]	<.001
	United States	0.12 (*n*=35,831)	-0.60 (*n*=6,003)	-0.76 [-0.79, -0.74]	<.001
	Australia	0.22 (*n*=7,423)	-0.63 (*n*=115)	-0.96 [-1.15, -0.78]	<.001

Hispanic, Latinx, Or Spanish Origin Applicants' Mean Scores Compared To White Or European Applicants' Mean Scores

Application Year	Country	White or European Applicants	Hispanic, Latinx, or Spanish origin Applicants	Cohen's *d* [95%CI]	*p*
2018-2019	Canada (English)	0.11 (*n*=10,059)	0.07 (*n*=1,920)	-0.04 [-0.09, -0.01]	.110
	United States	0.06 (*n*=20,857)	-0.10 (*n*=6,787)	-0.17 [-0.19, -0.14]	<.001
	Australia	0.23 (*n*=6,371)	0.02 (*n*=436)	-0.24 [-0.33, -0.14]	.110
2019-2020	Canada (English)	0.14 (*n*=11,694)	-0.21 (*n*=342)	-0.39 [-0.50, -0.29]	<.001
	Canada (French)	0.14 (*n*=4,257)	-0.48 (*n*=97)	-0.70 [-0.89, -0.49]	<.001
	United States	0.04 (*n*=26,378)	-0.22 (*n*=4,625)	-0.27 [-0.30, -0.24]	<.001
	Australia	0.22 (*n*=6,395)	-0.17 (*n*=64)	-0.45 [-0.69, -0.20]	<.001
2020-20201	Canada (English)	0.17 (*n*=17,078)	-0.19 (*n*=568)	-0.40 [-0.48, -0.31]	<.001
	Canada (French)	0.15 (*n*=4,750)	-0.45 (*n*=124)	-0.67 [-0.85, -0.50]	<.001
	United States	0.12 (*n*=35,831)	-0.27 (*n*=7,141)	-0.41 [-0.44, -0.39]	<.001
	Australia	0.22 (*n*=7,423)	-0.23 (*n*=75)	-0.51 [-0.73, -0.28]	<.001

Asian Applicants' Mean Scores Compared To White Or European Applicants' Mean Scores

Application Year	Country	White or European Applicants	Asian Applicants	Cohen's *d* [95%CI]	*p*
2018-2019	Canada (English)	0.11 (*n*=10,059)	-0.02 (*n*=5,499)	-0.13 [-0.16, -0.10]	<.001
	United States	0.06 (*n*=20,857)	0.12 (*n*=9,013)	0.07 [0.04, 0.09]	<.001
	Australia	0.23 (*n*=6,371)	-0.66 (*n*=2,149)	-0.96 [-1.01, -0.91]	<.001
2019-2020	Canada (English)	0.14 (*n*=11,694)	0.00 (*n*=6,239)	-0.15 [-0.18, -0.12]	<.001
	Canada (French)	0.14 (*n*=4,257)	-0.12 (*n*=402)	-0.29 [-0.39, -0.18]	<.001
	United States	0.04 (*n*=26,378)	0.18 (*n*=11,673)	0.15 [0.12, 0.17]	<.001
	Australia	0.22 (*n*=6,395)	-0.59 (*n*=1,949)	-0.86 [-0.91, -0.81]	<.001
2020-2021	Canada (English)	0.17 (*n*=17,078)	-0.05 (*n*=8,546)	-0.23 [-0.26, -0.21]	<.001
	Canada (French)	0.15 (*n*=4,750)	-0.03 (*n*=442)	-0.19 [-0.29, -0.10]	<.001
	United States	0.12 (*n*=35,831)	0.11 (*n*=14,663)	-0.01 [-0.03, 0.01]	.209
	Australia	0.22 (*n*=7,423)	-0.48 (*n*=2,041)	-0.74 [-0.79, -0.69]	<.001

Indigenous Applicants' Mean Scores Compared To White Or European Applicants' Mean Scores

Application Year	Country	White or European Applicants	Indigenous Applicants	Cohen's *d* [95%CI]	*p*
2018-2019	Canada (English)	0.11 (*n*=10,059)	-0.14 (*n*=455)	-0.27 [-0.36, -0.18]	<.001
	United States	0.06 (*n*=20,857)	-0.17 (*n*=387)	-0.23 [-0.33, -0.13]	<.001
	Australia	0.23 (*n*=6,371)	0.04 (*n*=132)	-0.22 [-0.39, -0.05]	<.001
2019-2020	Canada (English)	0.14 (*n*=11,694)	-0.13 (*n*=273)	-0.30 [-0.42, -0.18]	<.001
	Canada (French)	0.14 (*n*=4,257)	-0.03 (*n*=17)	-0.18 [-0.66, 0.29]	.560
	United States	0.04 (*n*=26,378)	-0.27 (*n*=147)	-0.33 [-0.49, -0.17]	<.001
	Australia	0.22 (*n*=6,395)	-0.10 (*n*=70)	-0.37 [-0.60, -0.13]	.003
2020-2021	Canada (English)	0.17 (*n*=17,078)	-0.05 (*n*=625)	-0.24 [-0.32, -0.16]	<.001
	Canada (French)	0.15 (*n*=4,750)	-0.11 (*n*=29)	-0.29 [-0.65, 0.08]	.235
	United States	0.12 (*n*=35,831)	-0.21 (*n*=324)	-0.35 [-0.46, -0.24]	<.001
	Australia	0.22 (*n*=7,423)	-0.01 (*n*=104)	-0.26 [-0.45, -0.06]	.014

Note. In the Canadian context, Indigenous refers to Inuit, Métis, or Indigenous applicants and in the Australian context, Indigenous refers to Torres Strait Islander and Māori applicants.

Middle Eastern Or Northern African Applicants's Mean Scores Compared To White Or European Applicants' Mean Scores

Application Year	Country	White or European Applicants	Middle Eastern or Northern African Applicants	Cohen's *d* [95%CI]	*p*
2020-2021	Canada (English)	0.17 (*n*=17,078)	-0.19 (*n*=2,170)	-0.39 [-0.43, -0.34]	<.001
	Canada (French)	0.15 (*n*=4,750)	-0.05 (*n*=967)	-0.21 [-0.28, -0.14]	<.001
	United States	0.12 (*n*=35,831)	-0.04 (*n*=2,526)	-0.17 [-0.21, -0.13]	<.001
	Australia	0.22 (*n*=7,423)	-0.27 (*n*=346)	-0.55 [-0.66, -0.44]]	<.001

Native Hawaiian Or Other Pacific Islander Applicants' Mean Scores Compared To White Or European Applicants' Mean Scores

Application Year	Country	White or European Applicants	Native Hawaiian or Other Pacific Islander Applicants	Cohen's *d* [95%CI]	*p*
2020-2021	Canada (English)	0.17 (*n*=17,078)	0.03 (*n*=97)	-0.16 [-0.38, 0.07]	.217
	Canada (French)	NA	NA	NA	NA
	United States	0.12 (*n*=35,831)	-0.03 (*n*=147)	-0.16 [-0.33, 0.00]	.046
	Australia	0.22 (*n*=7,423)	-0.13 (*n*=77)	-0.39 [-0.62, -0.17]	.001

Appendix 2
Regression Analyses for 2023-2024 Demographic Groups

US Regression Analysis (*n*=53,638)
Adjusted R2 = 0.09, F = 159.90, $p < .001$

Demographic Group	Estimate	Standard Error	*t*	*p*	η_p^2
Intercept	-0.11	0.02	-5.97	<.001	
Gender (Female)	0.23	0.01	17.53	<.001	0.01
Age (Over 28)	-0.29	0.02	-12.47	<.001	0.02
Race/Ethnicity (Black, African, Caribbean, or African American)	-0.48	0.02	-211.06	<.001	0.02
Race/Ethnicity (Hispanic, Latinx, or Spanish)	-0.21	0.02	-10.31	<.001	0.01
Race/Ethnicity (Asian)	0.00	0.02	0.14	.890	0.00
Race/Ethnicity (Indigenous)	-0.21	0.09	-2.35	.019	0.00
Race/Ethnicity (Middle Eastern or Northern African)	0.02	0.03	0.58	.565	0.00
Race/Ethnicity (Native Hawaiian or Pacific Islander)	-0.05	0.13	-0.37	.714	0.00
Household Income ($100,000 or more)	0.11	0.01	8.07	<.001	0.01
Parental Education (Bachelor's Degree or more)	0.13	0.02	8.36	<.001	0.00
Community Size (Rural)	-0.08	0.02	-4.74	<.001	0.00
Language Spoken at Home (Non-English)	-0.10	0.02	-5.22	<.001	0.00
English Proficiency (Non-Native English Speaker)	-0.41	0.02	-19.38	<.001	0.02
Relevant Work Experience (Over 10 Years)	-0.21	0.05	-3.90	<.001	0.00
Domestic or International Status (International)	0.08	0.04	2.15	.031	0.00
Ability Status (Does Identify as a Person with a Disability)	0.01	0.01	1.12	.264	0.00

Canadian Regression Analysis (English) (*n*=28,235)
Adjusted R2 = 0.11, F = 94.80, $p < .001$

Demographic Group	Estimate	Standard Error	*t*	*p*	η_p^2
Intercept	0.13	0.03	4.90	<.001	
Gender (Female)	0.12	0.02	6.42	<.001	0.00
Age (Over 28)	-0.33	0.03	-11.38	<.001	0.03
Race/Ethnicity (Black, African, Caribbean, or African American)	-0.45	0.04	-12.31	<.001	0.02
Race/Ethnicity (Hispanic, Latinx, or Spanish)	0.04	0.06	0.71	.480	0.00
Race/Ethnicity (Asian)	-0.02	0.02	-0.75	.456	0.00
Race/Ethnicity (Indigenous)	-0.20	0.06	-3.34	.001	0.00
Race/Ethnicity (Middle Eastern or Northern African)	-0.05	0.04	-1.52	.128	0.00
Race/Ethnicity (Native Hawaiian or Pacific Islander)	0.19	0.27	0.70	.482	0.00
Household Income ($100,000 or more)	0.13	0.02	6.89	<.001	0.01
Parental Education (Bachelor's Degree or more)	0.07	0.02	3.95	<.001	0.00
Community Size (Rural)	-0.06	0.02	-2.65	.001	0.00
Language Spoken at Home (Non-English)	-0.08	0.02	-3.68	<.001	0.01
English Proficiency (Non-Native English Speaker)	-0.42	0.02	-19.81	<.001	0.04
Relevant Work Experience (Over 10 Years)	-0.17	0.07	-2.42	.015	0.00
Domestic or International Status (International)	-0.35	0.04	-8.15	<.001	0.01
Ability Status (Does Identify as a Person with a Disability)	-0.02	0.02	-1.02	.309	0.00

Australian Regression Analysis (*n*=10,182)
Adjusted R2 = 0.11, F = 30.55, $p < .001$

Demographic Group	Estimate	Standard Error	*t*	*p*	η_p^2
Intercept	0.07	0.04	1.63	.103	
Gender (Female)	0.25	0.03	7.71	<.001	0.01
Age (Over 28)	-0.02	0.04	-0.55	.585	0.00
Race/Ethnicity (Black, African, Caribbean, or African American)	-0.46	0.14	-3.23	.001	0.00
Race/Ethnicity (Hispanic, Latinx, or Spanish)	-0.12	0.14	-0.82	.414	0.00
Race/Ethnicity (Asian)	-0.15	0.04	-3.54	<.001	0.03
Race/Ethnicity (Indigenous)	-0.22	0.14	-1.56	.120	0.00
Race/Ethnicity (Middle Eastern or Northern African)	-0.14	0.08	-1.69	.090	0.00
Race/Ethnicity (Native Hawaiian or Pacific Islander)	-0.31	0.22	-1.40	.163	0.00
Household Income ($100,000 or more)	0.10	0.03	3.24	.001	0.01
Parental Education (Bachelor's Degree or more)	0.10	0.03	3.20	.001	0.00
Community Size (Rural)	-0.19	0.04	-4.40	<.001	0.00
Language Spoken at Home (Non-English)	-0.22	0.05	-4.58	<.001	0.02
English Proficiency (Non-Native English Speaker)	-0.38	0.04	-9.90	<.001	0.03
Relevant Work Experience (Over 10 Years)	-0.16	0.10	-1.56	.120	0.00
Domestic or International Status (International)	-0.21	0.05	-4.07	<.001	0.00
Ability Status (Does Identify as a Person with a Disability)	0.01	0.03	0.40	.691	0.00

Canadian Regression Analysis (French) (*n*=6,284)
Adjusted R2 = 0.10, F = 20.84 , *p <.001*

Demographic Group	Estimate	Standard Error	*t*	*p*	η_p^2
Intercept	-0.05	0.05	-0.85	.396	
Gender (Female)	0.11	0.04	2.73	.006	0.00
Age (Over 28)	-0.21	0.07	-3.11	.002	0.02
Race/Ethnicity (Black, African, Caribbean, or African American)	-0.46	0.07	-6.10	<.001	0.03
Race/Ethnicity (Hispanic, Latinx, or Spanish)	-0.06	0.13	-0.44	.658	0.00
Race/Ethnicity (Asian)	-0.01	0.08	-0.16	.874	0.00
Race/Ethnicity (Indigenous)	-0.27	0.18	-1.45	.147	0.00
Race/Ethnicity (Middle Eastern or Northern African)	-0.05	0.06	-0.92	.358	0.00
Household Income ($100,000 or more)	0.16	0.04	4.29	<.001	0.02
Parental Education (Bachelor's Degree or more)	0.17	0.04	4.05	<.001	0.01
Community Size (Rural)	-0.11	0.05	-2.40	.016	0.00
Language Spoken at Home (Non-French)	0.12	0.06	1.99	.046	0.00
French Proficiency (Non-Native French Speaker)	-0.25	0.06	-4.24	<.001	0.01
Relevant Work Experience (Over 10 Years)	-0.32	0.16	-2.03	.042	0.00
Domestic or International Status (International)	-0.80	0.10	-7.96	<.001	0.02
Ability Status (Does Identify as a Person with a Disability)	-0.02	0.04	-0.46	.644	0.00

Appendix 3
Fit Indices & Difference Statistics for Measurement Invariance Models by Gender (Test 1)

Fit Indices and Difference Statistics for Measurement Invariance Models by Gender (Female, n = 1526, Male, n = 1124; Test 1)

Model	χ2	df	CFI	RMSEA (90%CI)	SRMR	Δdf	ΔCFI	ΔRMSEA	ΔSRMR	Decision
Female Baseline	43.43	54	1.000	.000 (.000, .009)	.012	-	-	-	-	-
Male Baseline	70.27	54	.996	.016 (.000, .026)	.017	-	-	-	-	-
Configural (structure)	113.8	108	.999	.006 (.000, .016)	.013	-	-	-	-	-
Metric (loadings)	127.67	120	.999	.007 (.000, .016)	.021	12	.000	.001	.008	accept
Scalar (intercepts)	156.89	131	.997	.012 (.000, .019)	.023	11	-.002	.005	.002	accept
Residual (item residuals)	165.83	143	.998	.011 (.000, .018)	.023	12	.000	-.001	.000	accept

Note: χ2 = chi square test statistic; df = degrees of freedom; CFI = comparative fit index; RMSEA = root-mean-square error of approximation; 90% CI = 90% confidence interval for RMSEA; SRMR = standardized root mean square residual; Comparison analyses include: 1. metric versus configural model (metric = more restricted model); 2. scalar versus metric model (scalar = more restricted model); residual error versus scalar model (residual error = more restricted model). An accept decision is based on the combined results of ΔCFI ≤ -.010, ΔRMSEA ≤ .015, and ΔSRMR ≤ .030 (for metric invariance) or ≤ .010 for scalar and residual invariance. Fit Indices reflect robust estimates corrected for nonnormality. * χ2, $p < .05$

Appendix 4

Fit Indices & Difference Statistics for Measurement Invariance Models by Gender (Test 2)

Fit Indices and Difference Statistics for Measurement Invariance Models by Gender (Female, n = 1356; Male, n = 976; Test 2)

Model	χ2	df	CFI	RMSEA (90%CI)	SRMR	Δdf	ΔCFI	ΔRMSEA	ΔSRMR	Decision
Female Baseline	69.47	54	.997	.015 (.000, .024)	.017	-	-	-	-	-
Male Baseline	70.31	54	.995	.018 (.000, .029)	.019	-	-	-	-	-
Configural (structure)	139.78*	108	.996	.016 (.007, .023)	.016	-	-	-	-	-
Metric (loadings)	159.42*	120	.995	.017 (.009, .024)	.025	12	-.001	.001	.008	accept
Scalar (intercepts)	181.73*	131	.994	.018 (.011, .024)	.026	11	-.001	.001	.001	accept
Residual (item residuals)	211.24*	143	.992	.020 (.014, .026)	.026	12	-.002	.002	.000	accept

Note: χ2 = chi square test statistic; df = degrees of freedom; CFI = comparative fit index; RMSEA = root-mean-square error of approximation; 90% CI = 90% confidence interval for RMSEA; SRMR = standardized root mean square residual; Comparison analyses include: 1. metric versus configural model (metric = more restricted model); 2. scalar versus metric model (scalar = more restricted model); residual error versus scalar model (residual error = more restricted model). An accept decision is based on the combined results of ΔCFI ≤ -.010, ΔRMSEA ≤ .015, and ΔSRMR ≤ .030 (for metric invariance) or ≤ .010 for scalar and residual invariance. Fit Indices reflect robust estimates corrected for nonnormality. * χ2, $p < .05$

Appendix 5

Fit Indices & Difference Statistics for Measurement Invariance Models by Ethnicity (Test 1)

Fit Indices and Difference Statistics for Measurement Invariance Models by Ethnicity (Test 1)

Model	χ2	df	CFI	RMSEA (90%CI)	SRMR	Δdf	ΔCFI	ΔRMSEA	ΔSRMR	Decision
White Baseline (*n* = 400); Reference group	62.32	54	.994	.020 (.000, .039)	.028	-	-	-	-	-
Asian Baseline (*n* = 458)	61.19	54	.996	.017 (.000, .034)	.025	-	-	-	-	-
Configural (structure)	123.50	108	.995	.018 (.000, .031)	.024	-	-	-	-	-
Metric (loadings)	139.29	120	.994	.019 (.000, .032)	.037	12	-.001	.001	.013	accept
Scalar (intercepts)	161.73	131	.990	.023 (.007, .034)	.039	11	-.004	.004	.002	accept
Residual (item residuals)	174.50	143	.990	.022 (.006, .033)	.041	12	.000	-.001	.001	accept
Black Baseline (*n* = 380)	58.16	54	.997	.014 (.000, .036)	.026	-	-	-	-	-
Configural (structure)	120.45	108	.996	.017 (.000, 032)	.025	-	-	-	-	-
Metric (loadings)	143.50	120	.992	.022 (.000, .035)	.045	12	-.004	.005	.020	accept
Scalar (intercepts)	150.05	131	.993	.019 (.000, .032)	.046	11	.002	-.003	.001	accept
Residual (item residuals)	159.06	143	.994	.017 (.000, 030)	.050	12	.001	-.002	.004	accept
Hispanic Baseline (*n* = 315)	50.51	54	1.000	.000 (.000, .031)	.027	-	-	-	-	-

Configural (structure)	112.74	108	.998	.011 (.000, .030)	.026	-	-	-	-	-
Metric (loadings)	128.04	120	.997	.014 (.000, .030)	.038	12	-.001	.003	.013	accept
Scalar (intercepts)	139.39	131	.997	.013 (.000, .030)	.040	11	.000	.000	.002	accept
Residual (item residuals)	150.18	143	.997	.012 (.000, .028)	.042	12	-.002	-.002	.002	accept
Mixed Race Baseline (*n* = 258)	47.16	54	1.000	000 (.000, .030)	.028	-	-	-	-	-
Configural (structure)	109.43	108	.999	.006 (.000, .029)	.026	-	-	-	-	-
Metric (loadings)	123.00	120	.999	.009 (.000, .029)	.040	12	-.001	.002	.014	accept
Scalar (intercepts)	129.95	131	1.000	.000 (.000, .026)	.041	11	.001	-.009	.001	accept
Residual (item residuals)	152.51	143	.996	.014 (.000, .030)	.042	12	-.004	.014	.001	accept

Note: χ2 = chi square test statistic; df = degrees of freedom; CFI = comparative fit index; RMSEA = root-mean-square error of approximation; 90% CI = 90% confidence interval for RMSEA; SRMR = standardized root mean square residual; Comparison analyses include: 1. metric versus configural model (metric = more restricted model); 2. scalar versus metric model (scalar = more restricted model); residual error versus scalar model (residual error = more restricted model). An accept decision is based on the combined results of ΔCFI ≤ -.010, ΔRMSEA ≤ .015, and ΔSRMR ≤ .030 (for metric invariance) or ≤ .010 for scalar and residual invariance.* χ2, $p < .05$

Appendix 6

Fit Indices & Difference Statistics for Measurement Invariance Models by Ethnicity (Test 2)

Fit Indices and Difference Statistics for Measurement Invariance Models by Ethnicity (Test 2)

Model	χ2	df	CFI	RMSEA (90%CI)	SRMR	Δdf	ΔCFI	ΔRMSEA	ΔSRMR	Decision
White Baseline (*n* = 300); Reference group	48.21	54	1.000	.000 (.000, .029)	.029	-	-	-	-	-
Asian Baseline (*n* = 529)	63.23	54	.995	.018 (.000, .034)	.023	-	-	-	-	-
Configural (structure)	111.58	108	.999	.009 (.000, .027)	.024	-	-	-	-	-
Metric (loadings)	125.96	120	.998	.011 (.000, .027)	.043	12	-.001	.002	.019	accept
Scalar (intercepts)	144.52	131	.996	.016 (.000, .030)	.044	11	-.003	.005	.002	accept
Residual (item residuals)	154.75	143	.996	.014 (.000, .028)	.044	12	.001	-.002	-.001	accept
Black Baseline (*n* = 165)	53.50	54	1.000	.000 (.000, .047)	.040	-	-	-	-	-
Configural (structure)	101.52	108	1.000	.000 (.000, .028)	.031	-	-	-	-	-
Metric (loadings)	109.97	120	1.000	.000 (.000, .025)	.041	12	.000	.000	.010	accept
Scalar (intercepts)	126.62	131	1.000	.000 (.000, .029)	.047	11	.000	.000	.005	accept
Residual (item residuals)	146.84	143	.997	.011 (.000, .033)	.049	12	-.003	.011	.002	accept
Hispanic Baseline (*n* = 203)	76.99*	54	.968	.046 (.018, .067)	.044	-	-	-	-	-

Configural (structure)	125.03	108	.990	.025 (.000, .042)	.033	-	-	-	-	-
Metric (loadings)	129.45	120	.994	.018 (.000, .037)	.038	12	0.004	-.007	.005	accept
Scalar (intercepts)	144.39	131	.992	.020 (.000, .038)	.041	11	-.002	.002	.003	accept
Residual (item residuals)	166	143	.986	.025 (.000, .041)	.046	12	-.006	.005	.004	accept
Mixed Race Baseline (*n* = 280)	63.89	54	.989	.026 (.000, .048)	.035	-	-	-	-	-
Configural (structure)	112.15	108	.998	.012 (.000, .033)	.030	-	-	-	-	-
Metric (loadings)	123.88	120	.998	.011 (.000, .032)	.041	12	.000	-.001	.011	accept
Scalar (intercepts)	138.55	131	.996	.014 (.000, .032)	.043	11	-.002	.004	.002	accept
Residual (item residuals)	155.26	143	.993	.017 (.000, .034)	.044	12	-.003	.003	.001	accept

Note: χ2 = chi square test statistic; df = degrees of freedom; CFI = comparative fit index; RMSEA = root-mean-square error of approximation; 90% CI = 90% confidence interval for RMSEA; SRMR = standardized root mean square residual; Comparison analyses include: 1. metric versus configural model (metric = more restricted model); 2. scalar versus metric model (scalar = more restricted model); residual error versus scalar model (residual error = more restricted model). An accept decision is based on the combined results of ΔCFI ≤ -.010, ΔRMSEA ≤ .015, and ΔSRMR ≤ .030 (for metric invariance) or ≤ .010 for scalar and residual invariance. * χ2, $p < .05$