BEFORE THE NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY

Re: Responses to Request for Information

The Association of Test Publishers ("ATP") submits its responses to the questions posed in the Request for Information ("RFI") issued on March 17, 2022, by the National Institute of Standards and Technology ("NIST") as part of its initial draft AI Risk Management Framework ("AI RMF" or "Framework"). The ATP welcomes this opportunity to express its views and contribute to developing the eventual AI RMF, which we believe should assist in crafting future policy for regulating Artificial Intelligence ("AI") in the United States. We firmly agree with NIST that focusing on AI risk management is a crucial element – perhaps the most crucial element – amongst the overarching policies surrounding the appropriate regulation of AI systems. This response is being made by the required date of April 29, 2022.

Identity of the ATP

The ATP is a not-for-profit international trade association for the global assessment and lifelong learning industry, which includes regional organizations individually representing North America, Europe, Asia (including China, Japan, South Korea and Australia), India, and the Middle East/Africa. The ATP is comprised of hundreds of publishers, test sponsors (i.e., owners of test content, such as professional certification bodies), delivery vendors of tests used in various settings, including healthcare, employment (e.g., employee selection and other HR decisions), education (e.g., academic admissions), clinical diagnostic assessment, and for certification, licensure, and credentialing, as well as businesses that provide testing services (e.g., test development, proctoring, scoring) or who own and administer test programs ("Members"). Additionally, many Members are global vendors and service providers of technology in assessment and learning, including alleged or actual AI systems, as well as traditional automated decision-making solutions used in assessment.

Since its inception in 1987, the Association has advocated for the use of fair, reliable, and valid assessments, which includes ensuring the security of test content and integrity of test results. Our activities include providing resources and expertise to the U.S. Congress and state legislatures in the United States on legislative proposals affecting the use of testing in education and employment, as well as representing the industry in federal and state regulatory matters and litigation surrounding uses of testing. With the growth of its global representation, the ATP has been active in providing educational guidance materials on many international regulatory issues, including privacy and AI.

Relevant Background

Since 2016, the ATP has expended substantial resources on assisting its Members to understand and comply with emerging international privacy laws and regulations, including those addressing AI. The ATP has provided industry-specific education on the EU's General Data Protection Regulation ("GDPR") for its Members in the EU and the US, plus it has published a "Checklist for EU-US Privacy Shield Registration" (2016) and a "Compliance Guide for the EU General Data Protection Regulation" (2017). We also submitted comments expressing specific concerns of the assessment industry to the European Data Protection Board on its proposed 2019 Guidelines for the Use of Video Surveillance under the GDPR. In June of 2021, the ATP provided comments to the Organization for Economic Co-operation and Development ("OECD") on its proposed risk-based AI Framework. *See*

https://atpu.memberclicks.net/atp-comments-on-oecd-framework.[1] Subsequently, in August 2021, the ATP submitted comments to the European Commission to express the testing industry's concerns about the proposed regulation of AI systems under the Artificial Intelligence Act ("AIA"). *See* http://atpu.membershipclicks.net/atp-comments-on-EC-proposed-regulation.

In general, ATP Members are data-oriented organizations; thus, analysis, predictive analytics, and the increased use of technology-based systems, some of which are claimed to be "AI," have been vital tools in industry research and commercial activities for decades, including some in use since the 1950s. Specific to the uses in the assessment industry of advanced technologies/alleged to be "AI" (e.g., automated scoring of tests, automated generation of items, delivery of online assessments), the ATP published a White Paper in 2021 that provides information on the background of AI systems, the historic uses of advanced technologies in the industry, and the growing regulatory scrutiny being paid to AI. *See* https://atpu.memberclicks.net/assets/ATP%20White%20Paper_AI%20and%20Testing_A%20Primer_1July2021_Final%20R1%20.pdf .

Given its history of advocacy on behalf of its Members and the assessment and learning industry as a whole, the ATP provides these responses to the questions asked by NIST in its RFI. We earnestly hope this information will assist NIST in developing its AI RMF for use in the comprehensive regulation of AI in the United States that cuts through conceptual theories and focuses on the practical implications of providing a balanced, trustworthy, and ethical AI framework that complies with various national privacy laws as to how personal data is used within AI systems. In this way, the ATP desires to promote the establishment of an AI regulatory environment in the United States that will recognize the benefits of AI while ensuring that assessment organizations have the ability to manage the risks associated with the use of AI systems.

General Comments on Initial Draft

The NIST initial draft Framework sets forth a taxonomy of characteristics that should be considered in comprehensive approaches for identifying and managing risk related to AI systems: technical characteristics, socio-technical characteristics, and guiding principles. The ATP largely agrees with this taxonomy.

The ATP accepts the generalized characterization by NIST that technical characteristics refer to factors under the direct control of AI system designers and developers, which may be measured using standard evaluation criteria, such as accuracy, reliability, and resilience. Similarly, we accept that socio-technical characteristics refer to how AI systems are used and perceived in individual, group, and societal contexts, such as privacy, safety, and managing bias/discrimination. Thus, we understand that in the AI RMF taxonomy, guiding principles refer to broader societal norms and values that indicate social priorities, such as fairness, accountability, and transparency.

The ATP acknowledges that the AI RMF is intended to be a private/public collaboration under the National AI Initiative Act of 2020 (P.L. 116-283), and it also should be consistent with the National Security Commission on Artificial Intelligence recommendations and the Plan for Federal Engagement in AI Standards and Related Tools. We applaud NIST for its efforts to involve the broader AI community in developing this Framework. Moreover, we believe that NIST's encouragement of context- specific

[1] The ATP also is considering participation in OECD's development of an approach to assess the capabilities of AI solutions and compare them with human capabilities. OECD plans to use existing human tests to carry out this assessment, supplemented with AI-specific measures developed by the computer science community. The goal is to provide a set of valid and transparent measures of AI capabilities that give policymakers a meaningful way to understand what current AI systems can and cannot do. *See* https://www.oecd.org/education/ceri/future-of-skills.htm.

"profiles" for inclusion in future drafts is a significant, positive mechanism for evaluating risks and therefore, that approach should be utilized as the underpinning for this Framework. The ATP has consistently advocated for reasonable risk management, including in our 2021 comments to the OECD. However, we firmly assert that "one size does not fit all" and that US policy/regulation of AI should be based on, and proportionate to, the risks posed by the different uses of AI. Thus, the concept of profiles is fully consistent with the goal of managing the risks in a manner tailored to reflect how AI systems are actually used within a specific industry, recognizing the granular, discrete nuances present in such uses. Accordingly, the ATP intends to propose such a profile for the use of AI in assessment. As the ATP has commented to both the European Commission and the OECD, management of risks must not be classified in a circumscribed, narrow range of options (e.g., essentially no risk, low-risk, and high-risk), which ignore or fail to evaluate the granularity or proportionality of risks associated with specific activities.

Despite NIST's good intentions, the ATP believes that the Initial Draft Framework does not meet the intended legislative purpose. We believe that Congress expected this Framework would be structured along the same lines as the NIST Cybersecurity Framework and the Privacy Framework (*see* NIST SP 800-53 Rev. 5), where NIST guidance is directly tied to specific points of "control" that allow users to evaluate their relative vulnerabilities measured against those controls. Rather than establishing a coordinated set of guidance around AI controls, NIST proposes that users should "fill in the gaps" by referencing other guidelines, without any directions or guidance on which ones apply or how to use them. As a result, the draft AI RMF needs improvement both in terms of substance and in its ability to be used in a shared and common approach by multiple stakeholders evaluating the same or similar AI systems. Different organizations are likely to "fill in the gaps" in different ways, making the draft AI RMF inappropriate for US regulatory purposes and difficult to evaluate as a national compliance tool, let alone for members of any specific industry. Thus, we suggest that the Draft Framework does not achieve the goals we think Congress intended in the National AI Initiative Act. Moreover, we believe that the Draft Framework is not structured in the same manner as the other NIST frameworks and thus, there is no easy way for users to apply all three of the Frameworks in a coordinated manner. We encourage NIST to align the AI RMF more closely with the other NIST frameworks to ensure consistency across AI, privacy, and cybersecurity, which have significant overlap in issues and governance processes.

We appreciate that the AI RMF is voluntary, and that it "can be used to map compliance considerations," but the Framework does not well-support those objectives when it goes on to state that the Framework "is neither a checklist nor should be used in any way to certify an AI system." *See* Initial Draft at 2. Whether an entity self-certifies its AI systems, or uses an independent third-party certification body, the Framework is needed to fill that objective. Likewise, NIST's assertion that use of the AI RMF "does not substitute for due diligence and judgment by organizations and individuals in deciding whether to design, develop, and deploy AI technologies" (*id.)*, only highlights the lack of a controls checklist, which is precisely the result we believe Congress sought to achieve in order to enable developers and users of AI systems to be able to reference in compliance efforts, specifically including the NIST Cybersecurity Framework/Privacy Framework.

Perhaps the most problematic concern we have with the AI RMF is that NIST defines the term artificial intelligence as "algorithmic processes that learn from data in an automated or semi-automated manner." The ATP contends this definition is both overly-broad and imprecise. In fact, very little current computer-based technology/software used in the assessment industry is truly machine learning or based on algorithms using the personal information of test takers or profiling test takers. Rather, the majority of software is used simply to automate otherwise manual processes (see Assessment Industry Profile in Q9, *infra.* at pp 9-14), where no true AI is involved. Accordingly, the ATP strongly contends that a different definition of an AI system should be adopted – a viewpoint which appears to be growing in acceptance.

For example, on November 29, 2021, the European Council ("EC") agreed to restrict its legal definition of AI to exclude traditional software merely used to automate human actions rather than substitute for human decision-making.[2] The ATP supports this revised EC definition, application of which to the assessment industry would exclude a number of clearly non-AI activities (e.g. automated test scoring where the same scoring rubric is used by a computer program and human scorers, and automated item/test form construction for test delivery where a computer program merely automates processes that would be performed by a human).[3] As the assessment community has begun to incorporate advanced machine learning and algorithmic analytics ("advanced AI") , it is expected to become increasingly important, offering advantages such as scalability and efficiency in assessing higher volumes of test takers, creation of data rich measurement models, and greater data fidelity through technology-based simulation of task performances. The ATP asserts that the assessment industry's ability to utilize and grow such innovations from advanced AI is dependent upon a clear, easy-to-understand definition that reasonably focuses on true AI technology, not merely focusing on computer software that has been mistakenly assumed or perceived to be AI.

Consistent with the revised EC definition, then, the ATP suggests that NIST should adopt a narrower legal definition of AI, namely, that an AI system "engages in learning, reasoning, or data modeling to reach outcomes." We contend that this definition is more accurate than highly conceptual ones, and will avoid confusion in both the scope of AI and its application to risk management and regulatory compliance.

The ATP also agrees at a high level with many aspects of the taxonomy attributes and principles set forth by NIST, specifically concerns related to transparency, cybersecurity, privacy, safety, and infrastructure. However, we urge NIST to provide more clarity as to exactly what the distinction is between principles and attributes – and what the distinction means in the context of the Framework. To that end, the ATP will attempt to provide appropriate clarifications in our responses for NIST's consideration. In this vein, the ATP believes that NIST should seek to avoid premature requirements relating to aspirational standards for AI concepts that have not yet been defined or developed -- and which certainly have not yet matured to a level of common understanding. These include concepts such as "explainability," "auditability," "robust accuracy," and "error-free algorithms." Reliance on such vague and imprecise concepts will place unnecessary burdens on developers, implementers, and users, which could easily impact or even stifle innovation. Of course, the ATP understands and agrees with the need for NIST to focus on specific concrete and/or potential harms of AI systems, such as bias and unlawful discrimination. However, without common definitions and terminology, we firmly believe that states and localities will continue to regulate using their own terminology. For the sake of adoption of a uniform

[2] A "compromise text" to the draft EU AI Regulation under the Artificial Intelligence Act ("AIA"), released by the European Council (November 29, 2021), clarifies that traditional software that merely automates a manual task is not considered AI, in contrast to a system that requires data learning, reasoning, or modeling to reach outcomes. Thus, some types of testing software used today (e.g., scoring, item generation, test monitoring) should not be considered or treated as AI. By comparison, the original EC proposed regulation primarily focused on machine learning concepts to define AI as, "any system that generates content, predictions, recommendations or decisions, based on, *inter alia*, machine learning approaches, logic- and knowledge-based approaches, or statistical approaches." *See* AI, Article 3(1). Many comments to the EC, including by the ATP and the U.S. Chamber of Commerce, criticized the original definition, as well as the scope of its application.

[3] While we recognize that automated decision-making may still require privacy attention under applicable privacy laws and regulations, the ATP takes the position that software purely used for automated decision-making should not be classified as AI.

regulatory approach to AI, we urge NIST to focus the AI RMF on a set of AI controls that can be easily understood and applied by all stakeholders.

SPECIFIC QUESTIONS

NIST specifically asks for comments on eight detailed questions, plus a ninth miscellaneous one. The initial draft of the AI RMF does not include Implementation Tiers as considered in the original NIST concept paper, which tiers may be added later if stakeholders consider them to be a helpful feature in the AI RMF – an outcome that would be consistent with the Cybersecurity and Privacy Frameworks, and upon which NIST seeks comment. The ATP's responses will address the questions as relevant to the assessment industry's perspectives and issues.

Q 1. Does the AI RMF appropriately cover and address AI risks, including with the right level of specificity for various use cases?

As we have summarized in our General Comments, *supra.* at 3, a number of countries are incorporating the concept of risk into their AI frameworks, especially the European Union and Canada. By distinguishing between the levels of risk for different uses of AI, regulators will be better able to focus their limited resources on the uses of AI that could have the most significant impacts on individuals' lives.

This AI RMF is an initial attempt to describe how the risks from AI-based systems differ from other domains and to encourage and equip many different stakeholders in AI to address those risks purposefully. While NIST suggests the initial draft provides a "flexible, structured, and measurable process to address AI risks throughout the AI lifecycle," the ATP contends the guidance in the Framework needs to be precise to cover all of the levels of risk needed by US businesses for the development and use of trustworthy and responsible AI. Moreover, as we also indicated, *id.*, the ATP believes the RMF is not well-aligned with the Cybersecurity Framework and the Privacy Framework, which would enable users to integrate their analyses of AI risks with the appropriate security and privacy controls.

As a fundamental matter, the ATP agrees with the point made by the US Chamber of Commerce in its initial comments to NIST last September, namely, that to "manage" risk requires the ability to identify, assess, prioritize, respond to, or communicate those risks. We believe that there is a vast array of issues related to managing AI-related risk, including individual corporate challenges associated with resource constraints and incentives that make investing in AI risk management difficult, especially given the rapid pace of technological evolution. Thus, the major issue is that the Framework should recognize the nuances and granularity of risks that potentially arise in the specific industries/sectors. NIST has suggested it will consider adding profiles to the RMF to address that concern. To that end, the ATP has provided a Testing Industry Profile in its response to Q9 (*infra.* at pp. 9-14).

Q 2. Is the AI RMF flexible enough to serve as a continuing resource considering evolving technology and standards landscape?

The ATP contends that future disruptive technologies (e.g., ubiquitous AI), are unlikely to be effectively corralled through a unitary risk management framework (RMF). One of the AI RMF's main objectives should be to ensure that future AI systems and applications are ethical, trustworthy, and human centric. We assert that this objective will be extremely difficult to achieve using a single risk template. As stated in our "General Comments" (*supra.* at pp. 3-5), we think it is doubtful that a single AI RMF can possess "the granularity, and discrete nuance" attributes to provide clear guidance on ways to mitigate the potential harm to persons and society.

However, assessing the flexibility of the AI RMF for use with unknown future technologies appears to be premature. The definition of the Framework's flexibility in the context of risk management for current and future use is unclear. We believe NIST should first understand how effective the initial draft AI RMF will be in addressing today's concerns and issues.

So, addressing what is missing in the current approach that may hinder its usefulness with newer/emerging technologies, the ATP strongly suggests that NIST undertake a risk management "stress test" to assess the Framework's value on such dimensions as regulation, ethics/trustworthiness, and relevant mitigation strategy. The Framework's capability to add value across industry sectors (i.e., use cases) would demonstrate the Framework's generalizability and offer encouraging evidence of its flexibility when applied to tomorrow's breakthrough technologies. On the other hand, failure to meet the predefined benchmarks of the stress test would demonstrate that a unitary risk management framework is less likely to possess the flexibility and clarity necessary to meet the complexity of most emerging technologies. This latter outcome would seemingly point to the need to craft sector specific frameworks.

As we also have summarized in our General Comments, *supra.* at 3-5, the ATP believes that NIST should consider that some aspirational or presently unknown innovative capabilities of AI are not ripe for risk evaluation. Instead, NIST should focus on developing a concrete foundation of AI definitions, terminology, and applicable controls that can be easily understood and applied by all stakeholders as AI capabilities continue to evolve. Just as reliance on vague concepts and standards will place unnecessary burdens on developers and users of AI, and possibly stifle innovation, loose and/or ill-defined terminology and resulting standards related thereto creates additional risks where organizations and regulators alike are provided with little or no guiding compass. In the face of loosely defined terms and vague standards there is no minimum baseline for organizations to operate within creating fragmented interpretations of the Framework – this lack of certainty creates a substantial risk that states and localities will adopt their own understandings and application of AI regulation which will pose additional burdens related to doing business across disconnected or even conflicting legal or regulatory frameworks.

Q 3. Does the AI RMF enable decisions about how an organization can increase understanding of, communication about, and efforts to manage AI risks?

In order to avoid being caught "off guard" about reasonable steps towards regulatory compliance, the ATP has developed and released a document identifying core "principles" that the assessment industry should keep in mind and work towards related to the development and use of AI systems. *See* https://atpu.memberclicks.net/ai-principles. The assessment industry has always embraced regulation when applied in a consistent manner across the board, and the regulation bears a rational relationship to the purposes and strikes a balance between the costs and benefits of regulation. If an assessment organization truly relies on AI/machine learning and/or data modeling systems, then the ATP supports requiring transparency and trust that those systems are not biased or discriminatory in their application to individuals, whether in employment education, or training settings – as long as regulatory reporting systems are reasonably related to the goals of risk management and compliance is not unduly burdensome.

At any level of utility of AI systems, the assessment industry will benefit greatly from the availability of a robust, comprehensive AI RMF that enables every organization to understand and manage its risks surrounding the development and/or use of AI, as well as to effectively communicate to all stakeholders how its AI functions and may impact them directly.

Q 4. Are the functions, categories, and subcategories in the RMF complete, appropriate, and clearly stated?

While we agree with NIST that answers to the question of what makes an AI technology trustworthy will differ based on the key characteristics NIST cites, they all generally support the concept of trustworthiness. These characteristics include transparency, interpretability, privacy, accuracy, reliability, robustness, safety, security (resilience) and mitigation of harmful bias. There also are key guiding principles to take into account such as accountability, fairness, and equity. Despite our general understanding of these terms, the ATP has serious concerns about exactly how these attributes should be used in the real world of AI risk management. That is precisely why we advocate for establishment of a meaningful Framework based on well-enunciated terms and controls.

The ATP also generally agrees with NIST that "cultivating trust and communication about how to understand and manage the risks of AI systems will help create opportunities for innovation and realize the full potential of this technology." However, we think it is critical to recognize that what makes AI risk different is the ethical nature of the risk and therefore the realization that ethical risk elevates to a societal level of concern. Ensuring that AI developers and users focus on these ethical questions should be an underpinning of the Framework *(see* Q7, *infra.* at p. 8-9).

Q 5. Is the AI RMF in alignment with or leverage other frameworks and standards such as those developed or being developed by IEEE or ISO/IEC SC42?

The ATP notes that ISO, IEC, and IEEE are engaged in drafting cross-sector and sector-specific voluntary, consensus-driven AI standards that could be sources of information for crafting AI compliance guidelines.[4] To date, ISO/IEC JTC1/SC42 has adopted 11 standards since its creation in 2017; however, the focus of the majority of these standards is on data quality (ISO/IEC AWI 5259 – Parts 1-5), data lifecycle management (ISO/IEC CD 8183), and on market efficiency, rather than on topics that may be more relevant to establishing a risk management Framework or using risk-based categorization of AI systems.

Nevertheless, the ATP is aware of a single standard that is under development on risk management, ISO DIS 23894. As the DIS designation reflects, it is a "draft international standard" – the current ballot on this DIS closed on April 15, so there is still a need for the Committee to review comments, which could revise the standard, and could likely result in a reballot as a "final draft international standard." Accordingly, the ATP will reserve judgment on ISO 23894 until we can actually analyze the final standard; it is premature to evaluate its utility and consistency with the NIST Framework. But even if this standard covers specific topics that bear on effective risk management, the ATP notes the ISO standard is also a consensus voluntary standard, for which users must pay ISO. We contend that NIST should publish its Framework for the public to use as a self-contained document, in the same way NIST published the Cybersecurity Framework and the Privacy Framework.

The ATP strongly favors standardization to inform the conduct of AI research, setting forth guidelines for the responsibility of developers and users of AI, and for how safety and security procedures should be achieved. We also encourage greater reliance at least one element of the global OECD classification of AI systems, the concept of Action Autonomy Level (human support, human-in-the-loop, human-on-the-loop, human-out-of-the-loop).

[4] *See* "Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development," P. Cihon, Future of Humanities Institute, University of Oxford (2019).

Q 6. Is the AI RMF in alignment with existing practices, and broader risk management practices?

Perhaps the most fundamental concern the ATP has with existing risk management approaches, including the AI-RMF, is the apparent lack of understanding of the science of psychometrics that underlies all of assessment/measurement. Psychometric terms may sound identical to AI characteristics, but they are quite distinct and need to be treated separately so as not to confuse every professional assessment (i.e., "test") with an AI system. The ATP urges NIST to recognize the implicit differences between the science of psychometrics and the science of AI – these differences should be incorporated in the Testing Industry Profile as articulated in the following paragraphs in the response to Q9 (*see infra.* at pp. 9-14).

Since the 1950s, rigorous professional standards have governed the development, administration, and scoring of all psychometric assessments, especially in the areas of education and employment. *See Standards for Educational and Psychological Testing,* American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (ed. 201)4) ("Joint Standards").[5] These standards establish how assessments are professionally built, evaluated, and used -- based on their accuracy ("validity") and repeatability ("reliability"), as well as their fairness to test takers. The intent is to promote the sound and ethical use of tests and to provide a rigorous professional basis for the quality of testing and assessment practices. *See* Eignor, D. R., "Standards for the development and use of tests: The Standards for Educational and Psychological Testing," *European Journal of Psychological Assessment*, vol. 17(3), 157–163 (2001) https://doi.org/10.1027/1015-5759.17.3.157. Accordingly, the ATP requests that the NIST affirmatively recognize the Joint Standards as an example of the appropriateness of creating an industry-specific risk profile for assessments, which would assist in providing a reasonable compliance path for testing organizations to follow.

In the same vein, we urge that NIST should recognize that classical statistical algorithms in psychometrics (e.g., linear regression, multiple regression analysis) have been commercially deployed within the testing industry for decades with no documented negative impact on test takers' rights. In spite of the negative connotation the term "artificial intelligence" has acquired, in reality AI in psychometrics is not very "intelligent" but only applies technology – namely, mathematical and statistical functions performed on data. Accordingly, the ATP asserts there is truly no reason for regulating the mathematical and statistical functions themselves, and, therefore, the Framework should merely focus on the data analytics involved. The application of AI practices, as with the previously discussed application of discrete probability, statistical, and predictive practices and procedures, requires a transparency of data and algorithmic function origin, clarity of application, human oversight, and appropriate review mechanisms. These AI principles – which the ATP sees as the core themes of the proposed AI regulation – equally sit at the heart of the ATP's focus on fairness and transparency using psychometric principles. When the assessment industry adheres to those psychometric principles in practice, we believe that most applications of psychometric principles to AI techniques should not be deemed to be or classified as AI –

[5] Six versions of the *Standards* have been produced, with the most recent version published in 2014. These *Standards* are "joint" in nature in that they have been written by a joint committee of testing experts representing the three sponsoring organizations: the AERA; APA; and the NCME. The Joint Standards are widely accepted and followed by testing professionals around the world – they have been cited by the US Congress and the Supreme Court. Despite the title, these standards are widely acknowledged to apply equally to assessments used in certification/licensure, workforce and professional credentialing, and clinical/diagnostic settings.

and they certainly should not be considered high-risk, whether in the assessment industry or more broadly.[6]

Q 7. What considerations are missing from the AI RMF?

The ATP encourages NIST to provide more concrete discussion about ethics into the Framework. For example, assessment professionals recently have focused on best practices for ethical HR algorithms as a timely addition to the growing body of professional literature on the responsible use of artificial intelligence ("AI") in talent management. *See* Jones, J.W., & Cunningham, M.R. (2022). Ethical HR algorithms: Best practices for talent management, printed in Jones e*t al*., *Personnel risk assessment: Research advances, applications, impact, and compliance,* 156-170, Chicago, IL: FifthTheory.

This paper discusses how the advent of large scale database mining techniques and sophisticated data analytics has transformed not only customer online targeted marketing models but also employment-related talent management decision models for greater scalability, efficiency, and rapidity. Indeed, personnel tests, high-stakes licensure or certification exams, and other evaluation methods are often utilized as a source of information about a person to assess candidate capabilities, readiness and job-fit in determining overall suitability. The paper explains that data generated by ethical algorithms may also be integrated with other reliable and valid sources of candidate information in automated decision models to provide overall indicators of potential for performance. The paper provides a sensible framework and practical guidance for HR and Talent Management practitioners to consider in adopting or using AI solutions for employment-related purposes, whether it is in the context of applicant interviewing, test administration, scoring or reporting, pulse surveys, qualifications examinations, or succession planning – and enables stakeholders to ask the right questions of the right people to help reduce risks of using AI.

Finally, as we mention in response to Q5, the ATP believes that the role, timing, and logistics of human intervention deserves more attention. As such, the OECD concept of Action Autonomy Level (human support, human-in-the-loop, human-on-the-loop, human-out-of-the-loop) should be included in the AI RMF.

Q 8. Is the soon-to-be-published draft companion document citing AI risk management practices useful as a complementary resource and what practices or standards should be added?

NIST has stated that Part III of the Framework will include a companion Practice Guide to assist in adopting the AI RMF. The ATP will reserve comments on any examples and practices that NIST includes to assist in using the AI RMF. We appreciate that NIST has committed that the Guide will be part of a NIST AI Resource Center that is being established.

[6] For example, the ATP has noted that the EC's risk management approach automatically treats all computerized technology related to tests used in education, clinical, and employment settings as "high risk" activities. The ATP has strongly objected to this conclusion, which is based: on (1) the mistaken assumption that characterizes all knowledge, predictive, analytical, and logic-based practices as AI; (2) the overly-broad definition of AI; and (3) a lack of familiarity with long-standing testing standards and practices, including the well-documented and safe historical uses of data-based methods and computer technology.

Q 9. Other comments.

We indicated at the outset of this document that we would propose an industry-specific risk profile for NIST to incorporate into the next draft of Part II of the AI RMF. Consistent with our comments, this profile is predicated on specific functionalities in assessment and learning, including whether the process involves automated decision-making (ADM) or ML/AI and analyzes the specific risks we contend are associated with them. In each case, the suggestion (by the EC) that the risk is high is demonstrated to be in error – indeed, analysis of the risks presented in different assessment functionalities, demonstrates that a risk determination is highly dependent on the granularity and nuances of how the AI system is used. Accordingly, the ATP urges NIST to include this Profile in the Framework.

ASSESSMENT INDUSTRY PROFILE

A. Test Content Analysis

• Question construction. Some forms of AI are used in generating/writing items/questions for use in tests (e.g., by taking some instructional text and using language analysis to construct new, different questions based on that analysis).

Risk Analysis: Significantly, it is well-accepted best practice that every item that is used in a test, whether it is written by a human item writer or is generated by a computer (i.e., some form of automated item generation or "AIG"), is subjected to bias study, other psychometric research, and pilot testing, to make sure that items evidencing any form of bias are removed from the pool of items eventually used in constructing tests (whether by humans or computer). Equally significant, such item construction does NOT involve any use of current test taker personal data: if past test usage is used for research purposes, that information has been de-identified and aggregated so no one involved in the process has any access to individual test taker information. Based on ATP's proposed definition of AI (*see supra.* at p. 4), AIG is not an AI system. Based on this analysis, the ATP believes AIG never rises to the level of a "high risk" activity and therefore should not automatically be regulated as such.

• Question selection/ordering of items. Some organizations use algorithms to select questions to be included in a particular test or separate forms of the test (e.g., for a regional or national administration), or for establishing the order in which items appear on a test form (either paper-based or delivered by computer). A similar process is used to present a unique set of test questions to individual test-takers (e.g., in fixed format using Linear on the Fly ("LOFT") testing, in variable forms using Computer Adaptive Testing ("CAT"), or in other situations where each test-taker receives a personalized/ customized assessment). AI may be used to make more effective selections.

Risk Analysis: Generic item selection processes for creating many tests or comparable fixed forms of tests are conducted before any test administrations – and are performed without any reference to specific test taker personal information. Selection decisions for a test are based solely on considerations of ensuring appropriate test content/coverage of subject areas and related psychometric principles for each test. Creating different forms of the same test are similarly performed without reference to test taker personal information to ensure that those multiple forms are equitable (i.e., same level of difficulty, same level of content coverage, and same level of validity/reliability -- so that scores on all forms of a test can be compared). These types of algorithms are built by trained psychometricians to achieve those results – and those algorithms are applied exactly the same way a human would apply them if performing the same scientific work by hand. By comparison, if a psychometric algorithm is employed in tailoring test items for a test administration to individual test takers (i.e., Computer Adaptive Testing ("CAT")), then the next question asked of each test taker is dependent on his/her previous answers, which provides a more

efficient test administration and a more accurate scoring methodology. While it is appropriate to explain this process to test takers in reasonable terms, nothing about it is prejudicial or discriminatory – all questions ultimately given to each test taker were previously equated with all other items in the pool of possible questions from which items are selected and all have been pre-determined to be valid and reliable and free from bias. Based on this analysis, the ATP believes this functionality rarely if ever rises to the level of a "high risk" AI activity and therefore should not automatically be regulated as such.

• Data analysis. Data analytics techniques that may include AI are used to analyze assessment data sets and make predictions or evaluative analyses (e.g., predicting job competence, identification of learning deficiencies, evaluation of compliance risks).

Risk Analysis: As mentioned earlier, while prediction and statistical analysis are mathematical components of ML/AI, in these applications those principles are critical hallmarks of the psychometric process and managed through the history of psychometric governance. Whether used in an educational or employment setting, these types of data analytics (e.g., as opposed to profiling of a person for targeted marketing purposes) enable an organization to evaluate uniformly every candidate against a pre-determined set of common criteria (be they job-related or education-based). Based on this analysis, the ATP believes this functionality does not rise to the level of a "high risk" AI activity and therefore should not be regulated as such.

• AI in learning. Edtech companies and other testing organizations focused on various functions (e.g., reading, training, language learning) are using AI to aid in helping individuals to learn, whether that is through traditional instruction or e-learning/e-assessment, at every level of education, including social-emotional learning, life-long learning, and employment training. By definition, personalized learning is intended to adjust the program to the specialized needs of the individual, using systematic, step-by-step methodologies by which the person is able to advance towards identified educational goals.

Risk Analysis: Machine learning (ML) and data analytics enable a testing organization to create more effective personalized learning instructional content (e.g., courses, curriculum), and to assess a person's competence/skills or to assist in making career choices, whether that is to identify education weaknesses or positive pathways. Nevertheless, some aspects of personalized learning may also involve ADM to address how the learning program operates. Any program structure is tied to the psychometric principles to demonstrate validity, reliability, and fairness. Critically, when the testing organization gives notice to the individual about how the personalized learning works, the use of personal information is directly related to the profiling used to create the personalized plan and is exactly what the individual student expects/has agreed to; in other words, the AI solution is co-extensive with the outcomes sought by the individual. In these use cases, the ATP agrees that relevant test taker protections and privacy considerations need to be taken into account in determining how to regulate this functionality, using a more granular risk analysis to evaluate where on the scale of risk any specific AI system falls.

B. Test administration integrity/security[7]

[7] Standardized test administration is required to assure that everyone who takes a test has the same opportunity to be measured on a test given under the same conditions to achieve fair results. See *Standards for Educational and Psychological Testing* (2014) (*see supra.* at p. 8). The Joint Standards have been recognized in most countries around the world. *See also,* ISO 10667 -- Parts 1 and 2 (2011) Section 5.4 (Note), which requires that "… when administering an assessment to one or more individuals, assessment administrators follow the standardized procedures for the delivery of the assessment and document any deviations from those procedures." Standard administration requires observing the test administration, to identify any irregularities that may occur (e.g., use of cheating devices, instance of a power failure, medical emergency, disruption of test takers), as well as to protect the test content from being copied and illegally distributed (e.g., infringing the owner's copyright).

• Analyzing photographic images. Another useful application of AI supports a testing organization verifying the identity of a test taker. Here, AI is used to compare a form of identification/photographic image provided by a test taker at the time of registration with the identification provided at the time of testing. A match ensures that the proper person is taking the test, and not a surrogate/imposter.

Risk Analysis: This "one-to-one" match function is merely an electronic image evaluation of whether the identification provided by a test taker at two different times match one another, so that only the person who registered (or is eligible) to take a test actually takes it. This type of AI function does not actually constitute (practically or even legally in some jurisdictions) biometrics/facial recognition – and the individual was given notice about and consented to provide the testing organization (or its vendor) with personal identification. Further, the testing organization (or its vendor) provides notice to the test taker about the image matching process, and the individual is asked to provide consent prior to the testing organization (or its vendor) collecting the individual's identification/photographic image at the time of registration. Then the previously provided identification is matched with the identification presented by the individual before the test administration begins. Even if a digital match is performed, it almost always occurs under human supervision, allowing for a digital match to be overruled. This match serves the same exact function as using one's own biometrics to open a mobile device or laptop – to ensure that only the right person is able to get access. Indeed, this image matching process ensures the integrity of the testing event so that all persons involved can be assured that a surrogate/imposter is not cheating the system. No other use of the identification/photographic image is made and images are not retained longer than necessary to resolve challenges; the identification/photographic image is not used as part of the test, to change the test administration or scoring, or conduct any profiling of the test taker. Based on this analysis, the ATP believes this functionality does not rise to the level of a "high risk" AI activity and therefore should not be regulated as such.

By comparison, in other types of online proctoring systems, AI is used to perform digital facial recognition or other analysis of biometrics to help in identification of test takers. Some of these situations involve a "one-to-many" analysis, where in fact personal profiling of individual test takers occurs. When profiling occurs, the ATP agrees that it is important that the AI system provides accurate profiles for test takers of varying demographics.[8] In these use cases, the ATP agrees that relevant test taker protections and privacy considerations need to be taken into account in determining how to regulate this functionality, using a more granular risk analysis to evaluate where on the scale of risk any specific AI system falls.

• Analyzing video/audio. Using AI enables the analysis of a testing event in real time (either during in person or online administration) with test proctoring/monitoring by one or more humans to determine if any test taker has cheated on the test, or has stolen test content.

Risk Analysis: Such "hybrid" proctoring systems use algorithms to analyze video and/or audio recordings, often along with other data (e.g., observation by a human proctor in either real time or subsequent to the testing event), to identify test taker actions that could raise questions about the integrity of the test administration (e.g., using a mobile phone, talking to someone through an earpiece, persistent looking away from the screen, seeing a second person in the room who could assist in taking the test). Such issues are flagged, typically for direct review by human proctors or reviewers, to determine if any genuine integrity violations have occurred. Moreover, testing organizations are careful to provide procedures for any test taker to challenge a ruling/score where analysis of video has occurred. Based on

---

[8] Significantly, even this use of biometrics in testing is not equivalent to public surveillance (e.g., for law enforcement) inasmuch as test takers have registered for the testing event and have been notified that using an imposter is a violation and that profiling will occur as part of the process.

this analysis, the ATP believes that when there is human involvement this functionality does not rise to the level of a "high risk" AI activity and therefore should not automatically be regulated as such.

• Fraud detection. Machine learning and other AI solutions can also be used to look for and analyze patterns in data collected during the test administration to identify anomalies that could represent cheating or other test integrity issues (e.g., forensic data analytics, keystroke analysis).

Risk Analysis. In some cases, the AI is capable of identifying a statistical rationale for a potential anomaly, which establishes the person has not cheated, while in other cases, machine learning or other AI systems are capable of identifying a potential issue that has no apparent rationale or explanation. As implemented, these AI systems generally produce information that is escalated for review – either in real time or subsequently – by a human being to resolve whether a particular action was an attempt to cheat, including a procedure for challenge or appeal of the decision. Consequently, the ATP agrees that relevant test taker protections and privacy considerations need to be taken into account in determining how to regulate this functionality, using a more granular risk analysis to evaluate where on the scale of risk any specific AI system falls.

C. Test Scoring

• Scoring answer sheets. Automation, in the conversion and computerization of data on paper-based assessments, using optical readers to read "fill-in-the-bubble" answer sheets and convert them to digital information, has been used since the 1950s. Identical automated scoring occurs on computer-based assessments, by converting on-screen responses to digital information for scoring against the scoring key. Such scoring is usually associated with multiple choice test items.

ATP Analysis: Critically, such computerized functionality is ADM, not AI; moreover, no test taker's personal information is involved in the process, inasmuch as the scoring completely relies on a human-developed scoring key ('rubric"), comprised of correct/desirable responses based on scientific research. As discussed above, *supra*. at p. 7, the ability of the optical/computerized scoring system to provide more accurate results in a more efficient manner and timeframe benefits all stakeholders. Based on ATP's proposed definition of AI (*supra*. at p. 4), automated scoring is not an AI system. Based on this analysis, the ATP believes this ADM functionality does not rise to the level of a "high risk" AI activity and therefore should not be regulated as such.

• Scoring written test answers. One of the most established uses of AI in the testing industry is to automatically score certain types of questions (e.g., fill-in the blank, short answer, essays), whether those answers are handwritten or electronically captured in a digital format by a computer, using software designed to identify key words or phrases in a test taker's written response, digitize that information, and then provide scores. Computer-based systems for this purpose have been used by testing organizations since the late 1990s.

Risk Analysis: Similar to scoring multiple-choice test items discussed above, scoring other written test answers (whether handwritten or computer-entered) results in ADM that relies on a human-developed rubric comprised of key word or phrases. Here again, the scoring rubric uses no personal information from the test taker but the ADM merely "reads" the test taker's written answers. As with scoring answer sheets, this computer-based scoring performs the function faster and more accurately than human scoring. Accordingly, testing organizations are able to provide test scores on many tests taken on a computer at the end of the testing event, or within a much shorter "turn-around" time. The speed of scoring using this form of ADM is now commonplace, demanded by test takers who expect scores quickly, often to enable reporting those scores to an entity (e.g., educational institution, employer, certificate/credential issuer) that uses the scores to make a decision those test takers want or have paid for. All those final decisions

(e.g., admissions, hiring, issuance of a certification/credential) are made by the third-party entity, not the testing organization providing the test scores. Finally, virtually every testing organization provides test takers with notice about their right to challenge/appeal a score, so human intervention is anticipated and available in those challenges. Based on ATP's proposed definition of AI (*supra.* at p. 4), then, automated scoring is not defined as an AI system. Based on this analysis, the ATP believes this functionality does not rise to the level of a "high risk" AI activity and therefore should not automatically be regulated as such.

• Scoring audio responses. AI systems have been developed that are capable of recognizing speech to enable the scoring of verbal responses (e.g., in spoken English and other language proficiency exams). For example, a test-taker is asked a question, s/he speaks the answer, and the AI analyzes the response, evaluates it, and determines a score or grade, based on an analysis of the response.

Risk Analysis: Test taker engagement/speech analytics platforms that leverage AI and machine learning (ML) operate to capture, transcribe, and evaluate outcomes from those verbal interactions – those outcomes may range from native language speaking proficiency, to foreign language proficiency, to evaluating personal traits/characteristics based on speech patterns. Some of these AI solutions may utilize the speaker's personal information to profile or predict the test taker's abilities, while other solutions redact sensitive biometric data and focus exclusively on the words that are spoken. Consequently, the ATP agrees that relevant test taker protections and privacy considerations need to be taken into account in determining how to regulate this functionality, using a more granular risk analysis to evaluate where on the scale of risk any specific AI system falls.

• Scoring video responses. AI systems also score video recordings (e.g., a job applicant asked to respond to a series of recorded questions), where AI is used to evaluate and score the responses, and in some cases to screen out applicants who do not meet set job qualifications necessary for the job, or who fail to demonstrate sufficient skills necessary for a particular job (e.g., communications skills).

Risk Analysis: Although some AI systems are used to assess test takers' job-related skills and attributes, they may also predict how individuals will perform in a specific job. To some extent, such analyses are fully consistent with psychometric principles; in other instances they go beyond the scientific bases for assessment.[9] Other AI solutions are also being made available directly to job candidates, to assist them in preparing for interviews by evaluating them against typical attributes used by employers. Especially in these latter instances, AI producers are striving to deliver computer-based assessments powered by AI without infringing on people's privacy or security through the use of privacy-by-design, anonymization of all data to protect the sensitive information, and avoidance of any facial recognition profiling function. Consequently, the ATP agrees that relevant test taker protections and privacy considerations need to be taken into account in determining how to regulate this functionality, using a more granular risk analysis to evaluate where on the risk scale any specific AI system falls.

Conclusion

The ATP appreciates NIST's attention to the important issues surrounding the need for an AI risk management framework. First and foremost, the assessment industry needs an appropriate classification system by which to evaluate risks that helps the public understand where and what to be concerned about when an AI system is used, rather than lumping every test using computer software that has some kind of mathematical formula attached to it as AI and therefore automatically considered as a "high risk" activity.

[9] These uses of AI should not be confused with those performing video surveillance of testing events for the purpose of evaluating if test takers are attempting to cheat on the test or to identify other irregularities in the test administration (*see, supra.* at p. 11).

The assessment industry needs – and supports – a simple unified US regulatory standard, not a hodgepodge of state and local standards, policies, laws, and regulations with onerous reporting requirements. The assessment industry is not afraid of regulation, if the regulations make sense, are applied in a consistent manner across the board, and the cost of regulation bears a rational relationship to the benefits and is consistent with the regulatory purpose(s). If organizations are truly relying on AI/machine learning systems, then the ATP supports requiring transparency, assuming that reporting is simple and easy to use and compliance is not unreasonably burdensome.

The ATP is willing to answer any questions NIST may have about its responses; we would be happy to arrange for such an opportunity at a convenient time for NIST. We appreciate that NIST has scheduled a series of workshops to further these discussions. The ATP certainly hopes to participate in those hearings.

Sincerely,

ASSOCIATION OF TEST PUBLISHERS

William G. Harris, Ph.D.
CEO

Alina A. von Davier
2021 Chair of the Board of Directors
Chief of Assessment
Duolingo
and
President of the International Association for Computer Adaptive Testing
https://www.iacat.org

Alan J. Thiemann
General Counsel
Partner,
Han Santos, PLLC
225 Reinekers Lane, Suite 525
Alexandria, VA 22314