November 2022

Towards More Task-Generalized and Explainable AI Through Psychometrics

Alec Braynen
University of South Florida

Towards More Task-Generalized and Explainable AI Through Psychometrics

by

Alec Braynen

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science
Department of Computer Science and Engineering
College of Engineering
University of South Florida

Major Professor: John Licato, Ph.D.
Shaun Canavan, Ph.D.
Lawrence Hall, Ph.D.

Date of Approval:
November 2, 2022

Keywords: Generalization, Evaluation, Metrics, Dimensional, Space

Dedication

This is dedicated to my grandmother, Andree' Braynen, who passed away on August 14, 2022.

Acknowledgments

Thank you to my parents, Andre and Marlo Braynen, who supported and visited me from the Bahamas. And to my siblings, Matthew and Kristin-Grace, who continue to be a source of inspiration, love, and day-brightening humor.

Additionally, a huge thank you to Dr. John Licato for your support, inspiration, guidance, and advice that helped make the completion of my graduate degree coursework possible. And thank you, Antonio Laverghetta, Gene Simmons, Animesh Nighojkar, and Anna Khlyzova, for your insights and advice on the work.

Additionally, Jessica Matos, thank you for your support, encouragement, and companionship throughout the program. And thank you, Alana Clare, for your continued support and friendship throughout my academic career. Thank you to Lauren Cumberbatch, Jonathan Miller, and Trent Kirkendall for your continued friendship and support throughout the program.

Additionally, I'd like to acknowledge Dr. Shaun Canavan and Dr. Robert Karam, who inspired and encouraged me. And finally, thank you, USF faculty, for providing an environment for pursuing knowledge, growth, and fulfillment.

Finally, thank you to Dr. Lawrence Hall, Dr. John Licato, and Dr. Shaun Canavan for being my committee members and professors at USF.

Table of Contents

List of Figures

Abstract

In this work, we propose that adopting the methods, principles, and guidelines of the field of psychometrics can help the Artificial Intelligence (AI) community to build more task-generalizable and explainable AI. Three arguments are presented and explored. These arguments are that psychometrics can help by providing 1) a framework for formulating better datasets, 2) psychometric AI data that can lead to models of generalization in AI, and 3) explainable AI through more informative evaluations.

A review of psychometrics and psychological generalization is performed, along with an overview of evaluation, generalization, and explainability in AI. Various ideas are presented throughout for how psychometrics can lead to more task-generalizable and explainable AI. And where possible, works are presented and discussed that display some of these ideas.

Furthermore, counterarguments to each argument are presented and discussed. And finally, we conclude the work with a summary of the previously discussed and briefly discuss future research directions.

Chapter 1: Introduction

The Artificial Intelligence (AI) and Machine Learning (ML) communities continue to strive for better performance from AI algorithms. On the one hand, researchers work towards Artificial General Intelligence; AI capable of broad-purposed learning, generalization, and abstraction. And on the other hand, researchers pursue improvements in transfer learning and more localized task generalization. Toward these goals, we propose that psychometrics can help with improving AI generalization performance**.**

And nevertheless, despite recent advances in AI due to the adoption of deep neural networks, generalization performance and the ability to extrapolate from the training environment or dataset, continue to be pain points in building "durable" machine learning algorithms [71, 72]. As AI becomes more ubiquitous, it also becomes increasingly necessary for their outputs to be explainable [65, 67]. Explainability allows for AI to be more easily improved, and for the users of these systems to be able to trust, interpret, and explain the algorithm's output; contemporarily, this is known as explainable AI [65-67, 70]. And again, here, we propose that psychometrics can help.

With both these points in mind, we propose our central thesis: adopting the methods, principles, and guidelines of the field of psychometrics, can help the AI community to build more task-generalizing and explainable AI. We support this thesis by arguing that 1) psychometrics can help the AI community to design more feature-rich datasets and more comprehensive evaluations for AI, 2) using psychometrically supported evaluations of AI can lead to improvements in our understanding of the generalization performance of AI, which therefore can help us engineer more generalizable AI models, and 3) psychometrics can help to build more explainable and interpretable AI.

In Chapter 2, we argue that psychometrics can help the AI community to design more feature-rich datasets and more comprehensive evaluations for AI. We present an overview of psychometric principles and methods and then, compare them to the evaluation philosophy implicit in AI practice. In addition, we

discuss how these inadequacies in AI's philosophy of testing, can lead to skepticism, doubt, and even the loss of life in society. Then, we reference and discuss other works that show how psychometrics is used in a wide variety of disparate domains, and that this use in these disparate domains, suggests that psychometrics can help in evaluating black box AI algorithms as well. Then, we hypothesize about how psychometrics can do this and additionally present some examples that show how utilizing the psychometric principles and methods can lead to better AI evaluations. Finally, we explore counterarguments to our thesis, specifically that psychometrics may impede progress and that AI may evade psychometric testing.

In Chapter 3, we argue that psychometrics can improve the learning and generalization performance of AI. Towards this, we present a brief overview of psychological generalization and generalization in AI. Then, we present work that shows that psychometrics, as a field, has a long, successful history of measuring and evaluating generalization in humans. In addition, we hypothesize about how psychometrics can lead to more informative research on AI generalization and present recent work performed by our colleagues in USF's Advancing Machine and Human Reasoning (AMHR) Lab that demonstrates how psychometrics has been successfully applied to AI. Additionally, we show how psychometric evaluations can give useful feedback that improves the education process and hypothesize that it can do the same for AI. After exploring these supportive reasons, we analyze counterarguments that psychometrics cannot help with improving the generalization and learning performance in AI. Specifically, we explore the concern that AI may not generalize in a similar way to humans and the fact that AI has made significant progress without psychometrics.

In Chapter 4, we argue that psychometrics can lead to more explainable AI. We present an overview of explainable AI, along with its goals, its challenges, and its importance. Then, we discuss how the nature of psychometric evaluations results in interpretable and explainable data, which, we hypothesize, can help the community towards more explainable AI. Additionally, we show how psychometrics allows psychometricians to create insightful items/questions (effectively, mini tests) and hypothesize that a combination of these small explainable items, when applied to AI, can lead to systematic explanations and

interpretations. Finally, we explore counterarguments that psychometrics will not directly help to achieve (self) explainable AI.

We present these arguments in support of our thesis, which is, again, that adopting the methods, principles, and guidelines of the field of psychometrics, can help the AI community to build more task-generalizable and explainable AI. In the final chapter of this work, we summarize the points presented and discuss directions for further research.

Chapter 2: Towards Better Model Evaluation with Psychometrics

In this chapter, we argue that adopting psychometric principles and utilizing psychometric methods in AI, will lead to more effective training and more evaluative testing of AI. Firstly, we present a brief overview of the psychometric principles, the psychometric methods, and the testing philosophy in AI. Then, we briefly discuss the consequences of some of the shortcomings of evaluation in the AI community, to then propose, that psychometrics can help alleviate these shortcomings. Furthermore, we hypothesize some specific ways that employing psychometric principles can improve the testing and evaluation of AI. And additionally, we suggest some examples of how utilizing psychometric methods can lead to new AI metrics, evaluations, and datasets. Also, we present some work that demonstrates some of these examples. Afterward, we present and discuss counterarguments; namely, that psychometrics could also lead to degradation in AI research progress, and that AI itself may evade psychometric testing. Finally, we conclude this chapter with a summary.

2.1 Principles of Psychometrics

Psychometrics is the science of psychological measurement and assessment [73]. It emerged from researchers attempting to measure latent psychological constructs and abilities in humans. In other words, the goal of psychometricians is to scientifically measure and predict psychological abilities, traits, and properties. Encapsulated within this science are the guiding psychometric principles of validity, reliability, standardization, and bias-free testing [73][74]. These principles form the foundational goals towards which psychometricians endeavor to create psychometric tests. These principles are explained below.

2.1.1 Validity

Validity is the principle that a test or evaluation measures the construct it purports to measure [73]. Validity requires that a test not only appears to respondents and other observers as measuring a certain concept or ability (face validity) but also that it truly measures it (construct validity) [73][75]. Validity can

be determined in some ways by correlating the results of tests with observable, expected outcomes, correlating results of tests with other tests that measure the same (or similar) construct, and by ensuring the test does not heavily correlate with other tests and measures of contradictory constructs [73].

Validity is important in testing [51, 106], for without it, the test has no specific meaning or use. The test would have no predictive power and no known correlation to an observable outcome. For example, if one wanted to design a psychometric test for the suitability of a student to a particular degree, the test would need to appear to the student to have items on it that seem relevant to the degree, but also, the items would need to actually measure the student's suitability for the degree. If the test doesn't appear to measure the student's suitability for the degree, the student won't interact with the test seriously, which would negatively affect the validity of the results; additionally, if the test doesn't measure the student's suitability, then the test is useless as a measure of this ability. For example, if the test asks questions such as, "How comfortable are you in the degree program?", such a question may capture the student's adjustment to college in general instead of their suitability to the degree program.

Nevertheless, in this example, validity could be achieved by composing the test of known items where a correct answer indicates the possession of an important ability or understanding for the degree, comparing the test to other tests that may exist that have been shown to measure this suitability or unsuitability, and/or giving the test to multiple students as a trial and seeing if the test predicts how well the students perform during the first year of the measured degree. In any case, the validity of a test, which is the result of cumulating evidence over multiple trials of testing, must be established for the test to be of use.

2.1.2 Reliability

Reliability is the principle that requires a test's results to be reproducible and dependable. Ideally, a participant should obtain the same results whenever the test is readministered [73][75]. Some methods to determine reliability are, readministering the same test and correlating the results, administering two correlated tests, or administering the test in equivalent parts and correlating the results [73][76][83]. For determining the internal reliability of a test, which is the consistency of results across the items of the test,

"Cronbach's Alpha" is commonly used. Cronbach's Alpha is a functional average of all possible splits of a test into groups of items. For other reliability measures, such as test-retest reliability, or knowledge/ability measurement reliability, psychometricians correlate test results with readministered testing, other correlated tests, etc.

Reliability is inherently linked with validity.[1] in that, it ensures that a test has some use as a measurement. Reliability strengthens the validity of a test since it shows that the test is invariably measuring some concept. Without reliability, the predictive power of a test or evaluation cannot be trusted since unknown variances can affect the results. Continuing our previous example, for the test of a student's suitability for a degree to be reliable, the test would need to give the same, or close to the same results when readministered later. The student's variable mental state, or their history of having taken the test before, should not affect the results. Ideally, the test's results would be the same across many variable conditions.

2.1.3 Standardization

Standardization is the principle that requires that a test's results be referentially interpretable regarding the expected capabilities of the said test taker. In other words, standardization is the requirement that a test's results are comparable to some baseline or norm and are therefore interpretable. This can be achieved by norm-referencing and criterion-referencing [73, 77]. Norm-referencing requires administering the test to a large number of appropriate respondents (norms) and calculating the average, standard deviation, and ranking of the results. Criterion-referencing requires determining some outside/observable criteria and referencing the results against them. Criterion-referencing is much like observing the test-taker meet some requirement or achieve some observable performance [73].

Regarding standardization and continuing our example of creating a degree suitability test, if one wanted to modify the test to output a standardized suitability score, one would then begin norm-referencing. The test would be given to many respondents, their results received, and the data analyzed either against

[1] Reliability and validity form a symbiotic relationship with each other in that establishing the validity of the measurement of a psychological construct should, in almost every case, coincide with an establishment in the reliability of the measure. Of course, in the cases where the construct is itself inconsistent, i.e. the spontaneity or randomness of some property of the psychological space, the two principles may become independent;

local respondents, or if the distribution of the test was large enough, against the established norms of the large distribution. With the local respondents, the results could be ranked and normalized, or perhaps the score is produced against the average of all scores. With a large norm, the results of an individual test-taker could be compared against the calculated standards based on the norm. Criterion-testing this suitability would be difficult, however, one could create a miniature degree (a course) that captures many of the performances that the overall degree would require from a student and, observe how the student does in this course.[2]

2.1.4 Equivalence / Bias-Free / Fairness

The principle of bias-free/equivalence is that a test is free from discrimination or bias in testing. This means that a test evaluates the construct in a valid, reliable, and standardized way, without discrimination towards groups and individuals [73, 80-82]. In other words, it means that a test is a fair measure. Some ways bias can be identified in tests is by using differential item functioning, which measures the bias of particular items on the test, and measuring the intrinsic test bias, which is observing discrepancies in performance on a whole test across social groups [73].

For example, in our test of suitability for a degree, item analysis may show that the item is formulated in a way that leads to respondents from a certain country being disadvantaged in understanding the item. The item may be biased toward those whose first and primary language is English. Intrinsic test bias may show that groups from a certain county, or more simply different high-school curricula, show clear disparity with results. Tests must be fair since interpretations and inferences that affect a respondent's life are made from them. In these scenarios, a dialogue must constantly be occurring between the interpreters of a test and the test creators, to minimize discrimination and social bias in scenarios where the

[2] Criterion-referencing is a slightly contentious topic in psychometrics. Without norms external to the individual or the test, it becomes difficult to meaningfully interpret the results of some tests. In some cases, the premise of a criterion-referenced test is that the individual either passes or fails the test. More subtly however, criterion-referencing can also capture the intervals in the improvement of the individual toward passing a test. Perhaps the individual fails the criteria initially and needs to be trained/taught more, and then they fail again but less so than before; in these cases, the individual can be referenced against his/her/their self. Nevertheless, norm-referenced tests can be argued to be more meaningful, whereas criterion-referenced tests are more humane.

test naturally captures some discrimination across groups; and in scenarios where the bias is simply unintended discrimination, the items should be reformulated or removed.

2.2 Psychometric Methods

Some of the methods psychometricians use to create valid, reliable, standardized, and bias-free testing include item analysis, norm and criterion referencing, and item response theory. These are discussed briefly below.

2.2.1 Item Analysis

Item analysis involves examining (the facility and discrimination of) each item/question on a test.[3] One item analytic concept is the facility index, which indicates whether multiple respondents answer an item the same way [73] [84] [85]. A high facility index indicates the item is redundant (since every test taker answers it correctly) and a low facility index indicates that an item is too difficult. Discrimination is another item analytic concept, which is the ability of items to discriminate against respondents based on their understanding of a concept or level of skill at an ability; typically, it is measured by correlating each item on the test with the total score from the sum of all other items on the test.

2.2.2 Obtaining Standardization

1. Norm-Referencing

Standardization is achieved in psychometric testing by obtaining good norms (a large sample of results from relevant respondents) and categorizing, ranking, and extracting interval results from the scores [73][86]. From this data, standardized scores are produced with an assumption that performance on the test is normally distributed, then psychometricians transform this data into a variety of interpretable test scores such as T scores, Stanine scores, STEN scores, and IQ scores [73]. There are other assumptions that can be made about the distribution of scores on a test, but as with all good arguments, they should be clearly stated.

[3] Item analysis can be applied in ways outside of facility indices and discrimination, though these analyses seem to be most well delineated. Items can be analyzed for how predictive they are of one's overall test score, how pregnant they are with meaning and information to a respondent, how perplexing they are, how multifaceted they are, etc.

2. Criterion-Referencing

Standardization regarding criterion or domain referencing, is done by measuring the performance of a test taker against some domain, their previous performance, or some defined criteria. Criterion-referenced tests are performative, or observation-based tests. These are tests where a participant's performance is directly observed or compared against some objective domain [86-88].

2.2.3 Item Response Theory

Item Response Theory (IRT) is a method of treating items, questions, challenges, or groups of these on a test, as a falsifiable hypothesis of the test taker's ability [73, 83, 89]. Correctly answering an item leads to a more challenging item being presented, and incorrectly answering an item leads to an easier item being presented. With each answered item, the confidence in the prediction of the test taker's ability is increased. Additionally, IRT helps to inform the psychometric process of good item selection, test size reduction, cross-calibration of tests and items, and adaptive testing.

2.3 Evaluating AI: More Can Be Done

The AI community does not seem to have a guiding set of principles in the testing of AI. There have been calls in the literature for a need to reform testing and evaluation practices in AI [56, 78, 90-93, 101-103]. Additionally, while older ML algorithms allowed for researchers to understand their generalization behavior [61], it is difficult to do the same with newer algorithms [25]. We argue that the way we evaluate AI needs psychometrics to combat its increase in complexity. In this section, we present an overview of the testing philosophy in AI and propose that it can be improved with psychometrics.

2.3.1 The Turing Test

The Turing test (or the imitation game) [122] underlies the testing methodology in AI. The Turing test proposes that a machine is intelligent if its outputted behavior or ability is not distinguishable from that of humans under certain conditions. This test is often either implicitly or explicitly pursued when testing and evaluating AI ability. AI's performance is compared with that of humans on a variety of tasks, with the goal being, first, at the very least, matching human performance and if possible, surpassing it.

2.3.2 Intuition and Tradition: Principles and Customs of AI Tests

Without clear principles of testing, the AI community must resort to intuition and tradition as its principles of evaluation. AI is tested against popular tests of human ability, i.e., human IQ tests, human games, human performance on a particular task, etc. [58 – 60, 113 – 115], or against what seems best for its context of application [116 – 118]. This shows that some of the community uses tradition and intuition as its guiding principles in test creation and not a method of scientific testing of psychological (or high-dimensional) spaces.

These current testing practices lead to members of the community expressing doubt and skepticism of AI performance at best [12-18, 79, 90-93], and can lead to misunderstandings and the loss of life in the general public in the worst cases [70, 100, 102]. Additionally, these current testing practices do not allow for an easy understanding of an AI algorithm's potential performance [95-97]. When tests are built without a validated method supporting them, it is not easy to accurately figure out what their results mean. A high score on an evaluation could mean that the algorithm is well suited to the domain it will be deployed in, but it could also just mean the algorithm did well on the evaluation. We argue that adopting psychometrics can help with situations like these and allow the community at large to better test and evaluate AI capabilities.[4]

2.4 Applying the Psychometric Principles to AI Algorithm Testing Will Lead to Better AI Evaluation

2.4.1 How Psychometric Principles Can Help

The psychometric test principles of validity, reliability, standardization, and equivalence (or test freedom from bias), form the overarching goals that lead to the creation of good psychological tests. Psychometricians' ability to achieve these principles in testing has led to the use and acceptance of psychometric tests to test scholarly aptitude, mental health, and employee suitability [1-11]. If psychometrics is methodical enough to test humans on abilities and concepts in ways that will be used to

[4] We do not purport however that psychometrics can subsume or absolve the Turing Test. Psychometric evaluations would still be subject to the Turing Test however, it provides another modality with which to play the imitation game.

make important determinations about their cognitive abilities, we argue, psychometrics can be methodical enough to ensure we properly test for analogous abilities in AI as well.

Firstly, the principle of validity requires that the test appears to, and measures, the concept, construct, or ability it claims to measure. The application of this principle requires that tests not only appear valid to observers and test-takers, but also that it tests the underlying concept or psychological ability that it claims to test. This principle can benefit the AI community by requiring that metrics of performance believably appear to measure the ability claimed of an algorithm, and by requiring strong evidence from the researchers that the performance metrics and datasets used, will capture the ability or concept claimed of the algorithm. These proposals cannot simply be face-validated, meaning the dataset only seems on the surface to be related to the task. Instead, the datasets, training, and evaluation tests need to be strongly correlated with the criteria/context of the algorithm's application through various experiments and demonstrations. This would reduce debates, skepticism, concerns, and misjudgments about an algorithm's proposed capabilities, which in turn would lead to increased productivity, better collaborative work, and safer algorithm deployment in the community.

Secondly, the principle of reliability requires that test scores are accurate and free from measurement error in invariable conditions. Ideally, an evaluation should return the same result when readministered. The AI community has sort of adopted this principle in one way due to the computerized nature of its evaluations---after all, a deterministic machine will output the same results given the same inputs. However, a different type of reliability is frequently violated in AI: different datasets which purport to measure the same thing often fail to correlate with each other [62, 99]. Ensuring that datasets have some corollary relationship with each other is one step toward more reliable datasets in AI.

Additionally, the principle of reliability could also be applied to making datasets more informative; for example, by adding the inter-annotator reliability of the annotations of a dataset [124]. It is not a priority in the community for the creators of datasets to include a reliability measure, or a measure of how suitable the dataset is for its task [125, 126]. Very often, datasets will be published with no detailed information on what the actual agreement was between annotators on each item (for example, the recent Uncertain NLI

dataset [94]). Psychometrics requires researchers to add these properties and measures to their datasets. These changes would also increase the trust in published work, improve dataset quality, help reduce skepticism, and better inform of an algorithm's performance and ability.

Thirdly, the principle of standardization requires that the tests have norms to be compared against. Standardization is required to have some ground truth to discriminate results against. This could be implemented in the community via community-accepted public datasets or performance evaluations for a particular task or ability. Standardization could also be implemented by researchers presenting their algorithm's performance relative to the performance of other algorithms in the same domain or dataset; however, this would require researchers to adopt the practice of making their code available in the public domain (if not open source, then at the very least, compiled versions of the code.) Instead of the common metrics of accuracy or precision, which are based on a dataset-contained evaluation, an algorithm's score/performance would be presented concerning the performance of other algorithms' performance. An accuracy of 99% on an individual dataset would become less meaningful, but an Artificial Intelligence Quotient (AIQ) score of 150 relative to most algorithms' performance of 130, along with an accuracy of 99% would be meaningful. Following the principle of standardization would make it easier for newcomers to the field to contribute to making progress on a particular task, reduce doubts and skepticism about a model's capabilities, and further bolster productivity, collaborative work, and iterative progress in the community.

Finally, the principle of non-discrimination requires that a test follows the previous principles in a fair manner across different test takers. We interpret this principle in two ways: 1) in that the validified, reliable, and standardized evaluations proposed in the community are publicly distributed so that there aren't any inherent biases to a particular model implementation due to advantaged datasets, and 2) models don't perpetuate bias in society against people [108, 109]. Both these interpretations mean that the quality of datasets being used to train and test models need to be rigorously investigated and will improve in some way.

2.5 Applying the Psychometric Methods Will Lead to New Evaluations and Datasets

Item Analysis, Criterion Referencing, and Item Response Theory are psychometric methods that help psychometricians to create items that measure ability or construct [73, 83-89]. The Item Analysis and Criterion Referencing methods, along with Item Response Theory, if applied to the creation of datasets and evaluation for AI models, can help researchers to gather better data and to create better tests of an algorithm's ability. These methods could help to prioritize challenges in the field with a sort of "AI facility index" of high-dimensional constructs, help guide researchers in building better extrinsic tests (read: criterion references) of a model's ability, help researchers analyze specific items in the dataset for constructs and relatability to each other and, increase the information contained in performance metrics by adopting the Item Response Theory approach to testing as an iterative and falsifiable hypothesis of ability.

With item analysis, an AI facility index could be built that allows researchers to dedicate their time and effort more effectively to achievable constructs or abilities. Instead of spending time on a problem that is, at a particular moment, seemingly insurmountable, psychometrics could help estimate the difficulty of a particular AI challenge. Determining algorithms' performance on certain datasets in a domain and its relation to other domains could allow for an AI facility index that ranks the current problems in AI by how likely they are solvable at that time.

Psychometric approaches to criterion referencing could also inform extrinsic and observable tests of AI algorithms. Perhaps an algorithm needs to achieve a certain score on an evaluative test, along with a score on an observable practical test of ability. Psychometrics can allow for an advanced Turing test to be proposed, where not only must the machine fool the interrogator with its behavior, but with the evaluation of its high-dimensional space via psychometric tests as well.

Additionally, each proposed item for a dataset could be analyzed for specific constructs and qualities to be included in a dataset. Whether the dataset is intended for training or testing could also affect the criteria. Researchers could analyze items by testing the dataset on humans, AI, or both, and select the items in the data that led to the highest learning rate, or best discrimination of some concept. Furthermore,

IRTs testing practices could be applied to evaluations, which would allow researchers to not only score an algorithm, but also provide the co-occurring confidence of the score [54, 120, 127-130].

In [120], Chmait et al., do exactly that. They used IRT to estimate the accuracy of AI on cognitive tasks of comparable complexities. Applying IRT to the evaluation function allowed them to create a lower bound on accuracy concerning the complexity of the task performance being measured. This allowed them to discern a relationship between an agent's selection cost, task difficulty, and accuracy as optimization problems. Additionally, in [127], Plumed et al., applied IRT to classification task evaluation to obtain evaluations of discrimination, difficulty, and guessing concerning instance hardness. Other works have also applied IRT as a method to select better items to use as datasets for NLP tasks [130] and to create more informative evaluations of models [128, 129]

2.6 Counterargument: Applying Psychometrics in AI Research May Slow Down Research Progress

One major disadvantage to applying the psychometric principles in AI model evaluation would be the increase in the amount of time and cost it takes to complete research [73, 104]. Properly identifying the construct one wants to measure, properly creating items that capture it, obtaining good norms, and finding and reducing bias all would increase the time and cost of research.

Establishing validity would require that a researcher investigate thoroughly, the concepts and theories surrounding the domain they will deploy an algorithm. Perhaps the researcher would need to consult with psychometricians and psychometric literature, along with the literature and experts of the proposed domain. Besides validating the algorithm with computerized evaluations, the researcher would also have to perform contextual evaluations, noting where in deployment the algorithm behaves unexpectedly, updating the previous evaluations to coincide with the observed behavior when deployed, and fixing the undesirable behavior. This process would have to occur iteratively until an accurate measure of the algorithm’s abilities emerges.

Alongside this process, the researcher would also have to establish a reliability measure. Perhaps deploying the algorithm in a similar but measurably different environment and observing its results or, evaluating it on a similar but measurably different test. Then to establish standardization, either a norm of

human behavior, a norm of AI behavior or both would need to be created — a costly and time-consuming endeavor of finding participants. Finally, adhering to equivalence would also require the researcher to ensure that the algorithm is not perpetuating societal bias, along with setting up a means of making the dataset public for others to contribute to and build upon.

This long process of imbuing an AI evaluation with psychometric principles and properties could add so much time to the research process, that the progress being made in the community may slow down dramatically which could lead to a reduction in progress (thereby prohibiting the improvement of generalization and explainability in AI). Additionally, incorporating psychometrics into the evaluative work also increases the cost of AI research, which could also lead to a reduction in progress.

Nevertheless, we argue that this increase in research time, if it occurred, would be positive for the field, particularly since it would coincide with an increase in the quality of the works being published. Published results would show more accurately the algorithms' performance in a domain, due to the results being strengthened with psychometric principles, methods, and guidelines, and additionally, bring with it, data for the community at large to use and build upon. As an aside, as shown in Figure 2.1 below, the number of papers published per year in the AI community is extraordinarily high, and a reduction in publication frequency (to make it a bit easier to keep up with good information), along with an increase in publication quality would probably be good for the community.

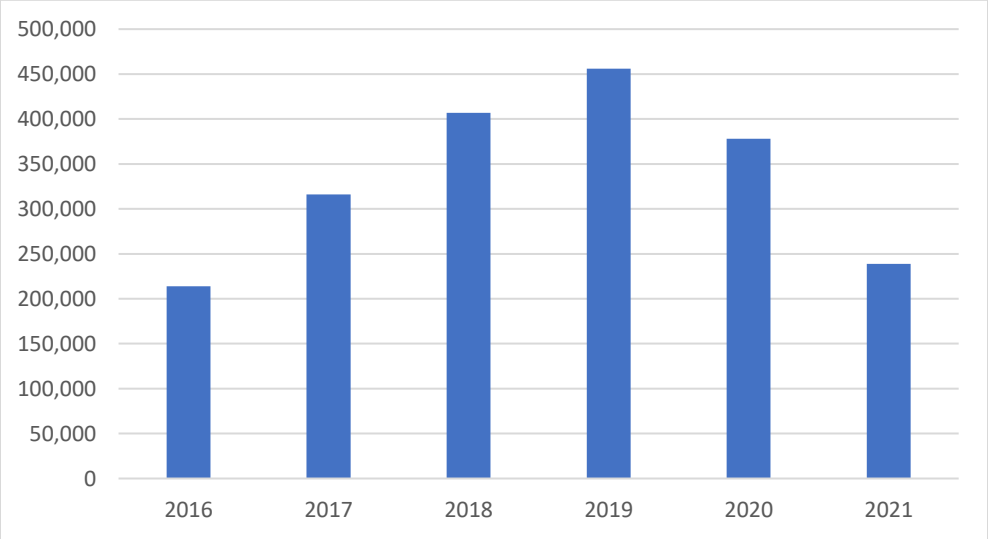

Figure 2.1: Number of Papers Published About AI (Google Scholar "Artificial Intelligence")

2.7 Counterargument: Psychometric Evaluations May Be Gameable by AI Algorithms

A critique of psychometric testing is that, due to the inherently intangible nature of the psychological space, psychometricians cannot be sure that they are measuring anything other than a test taker's ability to take a test [73, 105-107, 134]. Applying psychometrics to the evaluation of models in AI will open the field up to these same critiques levied against psychometrics. One would be able to argue that a measurement/test applied to an AI algorithm is not measuring the purported ability, but instead is only measuring the ability to pass the measurement. Another concern is that algorithms may simply find a pattern that allows them to game a test, instead of the test capturing some specific ability [110]. This pattern recognition ability means that models may end up discovering some pattern that allows for high results on an evaluation, instead of having high performance on the ability being measured. Researchers would not be able to be sure they are measuring an algorithm's specific ability, latent within its high dimensional space, or simply its ability to correctly answer the test.

However, we argue that this situation is the case even with current methods of evaluation in AI and that psychometrics has the methods to address these arguments. This is reflected in what has become known as "Goodhart's law": *When a measure becomes a target, it ceases to become a good measure* [123]. This limitation of benchmarks in AI is well known to practitioners, and this might be a reason why so many new benchmarks are constantly arising. But this is exactly the sort of thing that good psychometric practice is designed to counter: If a test in AI becomes nothing more than a target, then its validity and reliability will drop, and psychometrics measurements will reflect that.

Furthermore, these critiques are applicable today in AI since the measurements applied today only give some insight into an algorithm's performance on a specific test. A measurement of ability in a specific domain, if not rigorously obtained through measurements across multiple correlated datasets, can be argued to only show a model's performance on that specific test. In other words, these arguments against psychometrics apply to current testing methodologies in AI today. However, psychometrics has spent years building methods and theories to ensure that a test is valid, and these are used to counter the claim that a test is only testing ability on a test, or in AI's case, that the model is gaming the test. Psychometricians have

formulated in response, correlating a test's results with direct observations of real-world contextual performance, and other correlated tests of the same ability, and ensuring uncorrelation with tests of contradictory constructs (or abilities antithetical to the measured ability) all help to ensure that a test is validly testing some ability and not simply localized test-taking ability.

2.8 Summary

In this chapter, we argued that adopting psychometric principles and utilizing psychometric methods in AI will lead to more effective training and more evaluative testing of AI. First, we reviewed the psychometric principles, psychometric methods and theories, and evaluation principles in AI. We showed how psychometrics is used across a wide variety of domains and suggested some ways its principles could improve evaluation in AI; mainly, positing how to create psychometrically standardized evaluations in AI with increased validity and reliability. Additionally, we suggested some ways psychometric methods could help with creating new evaluations and presented some work that has utilized IRT to create new evaluations. Finally, we discussed how psychometrics could increase the cost and time-to-completion of research in AI and how AI may evade psychometric testing.

Chapter 3: Towards Better Learning and Generalization in AI with Psychometrics

In this chapter, we argue that psychometric tests and the psychometric data they generate for AI algorithms can lead to improvements in the learning and generalization performance of AI. Namely, by helping to contribute to models and theories of the generalization behavior of AI algorithms, which can be incorporated into the building and research of AI and therefore lead to improvements in their performance. To begin this chapter, we provide a brief overview of psychological generalization and the state the psychology field was in during the 1960s and 1970s; which was, trying to model generalization in humans. Then, we present a brief overview of the state of generalization in AI, which is that the complexity of algorithms has led to difficulties in trying to model and understand generalization in AI. To this fact, we argue that psychometrics can help the AI community to take steps towards modeling generalization in AI, presenting therefore, evidence that:

1. psychometrics has an extensive and variable history of testing generalization,
2. psychometric tests have contributed to the models of human generalization that we have today and
3. psychometrics leads to improvements in education and teaching

From these, we argue that psychometrics can provide similar benefits to the AI community and hypothesize some ways it might. Then, in addition, we present and discuss recent work that has demonstrated some of these ideas. Specifically, this work [133] utilizes psychometric principles, methods, and data in a manner that allows them to show that TLMs could be used to predict item difficulty for linguistic capability tests. From this, we further argue that adoptions and variations of this method, and the type of data it generates, can act as a foundation for a law of generalization in AI algorithms.

Afterward, we discuss counterarguments to this argument for our thesis, such as AI may not generalize similarly to humans, therefore weakening psychometrics' applicability to AI, and that the AI

community has made substantial progress without the adaptation of psychometrics and the various drawbacks its adoption may bring. And finally, the chapter is concluded with a summary.

3.1 Overview of Generalization

3.1.1 Psychological Generalization

Generalization is the ability to use past experiences and knowledge to deal with present situations and contexts [131]; or in other words, to see similarities from the past and apply it to the present. In the late 1960s and early 1970s, generalization ability in humans was generally thought to be unamenable to modeling [32]. Nevertheless, methodical tests and experiments were continually performed, leading to a vast source of human generalization data. This methodically collected data allowed for falsifiable, predictive theories of generalization to emerge from the research [32-34].

Generalization models today are generally built from the assumption that generalization occurs in a multidimensional psychological space of classes or consequential regions and when points occur somewhere in the same multidimensional consequential region/class, those points are said to be generalizable to each other [32-34].

3.1.2 Generalization in AI

State-of-the-art AI algorithms today often employ a deep architecture composed of input, output, and multiple hidden layers composed of neurons [132]. These neurons take in inputs and weights for each input, apply bias, summation, and activation functions to the inputs, and pass these outputs as inputs to other sets of neurons. Deep learning networks are powerful; however, their complexity and the depth of their architecture have made modeling/understanding their generalization behavior elusive [63-64]. State-of-the-art networks have become so big that they potentially overparameterize the data they have been trained on [63], however, this property doesn't change that properly evaluating their generalization capabilities remains difficult.

3.2 Evaluating Generalization Ability: Psychometrics as an Answer

One reason we argue that psychometrics can lead to improvements in generalization performance in AI is that psychometrics has extensive experience with evaluating generalization ability in humans in a

variety of different ways. Psychometrics has produced IQ tests, analogy tests, and generalization tests [35-39], which are all tests that deal with evaluating some generalization ability. Psychometrics' extensive results in evaluating generalization make it most suitable for informing the evaluation of generalization in artificial intelligence models [55, 57]. Psychometrics has created language-based, picture-based, pattern-based, and sensory-modal-based tests of generalization; clearly, these are a testament to the ability of the field to create varied types of generalization tests. Therefore, we argue that psychometrics can help with creating similarly varied evaluations of generalization in AI.

Psychometric tests of generalization abilities that are correlated with psychological constructs in humans can be used as tests for AI to allow researchers to infer the properties of an algorithm's high dimensional space and its similarity to humans' psychological space. Additionally, the items and the techniques of creating them on these psychometric tests can be used to prompt models in an attempt to quantify and qualify what they "understand," [137]. The scoring practices of psychometrics could also be used to create more informative metrics of generalization; for example, IQ or T-scores of generalization ability across AIs, or batteries of psychometric tests that give a unified score of some generalization ability [119].

Additionally, in some sense, it is easier and more convenient to measure generalization in AI than in humans. A researcher's ability to instantiate multiple different AI models and experiment with psychometric tests on them is much more convenient than gathering participants to test [133]. AI makes it easier to rigorously evaluate intelligence and its compositions from a psychometric perspective. In other words, using psychometrics to test AI generalization ability, could help not only the AI community's measurements and understanding of generalization and intelligence, but also, perhaps, the field of psychology.

3.3 Towards Models of Generalization in AI with Psychometrics

Another reason we argue that psychometrics can help the AI field to improve generalization evaluation and performance in algorithms because it has already helped to do it for humans [32-34]. In the next section, we explore a research paper that utilizes psychometrics to analyze Transformer Language

Models, which has implications for the field of Psychology; allowing them to use TLMs to create psychometric tests more cost-effectively; however, we argue, this paper also has implications for the AI community.

3.3.1 A Demonstration

In [133], Laverghetta Jr. et al., use psychometric tests to find similarities in an algorithm's high-dimensional space and humans' psychological space. Laverghetta Jr. et al. utilized the General Language Understanding Evaluation (GLUE) [98] benchmark as their data source for their psychometric items. Specifically, they utilize the broad coverage diagnostic task, a set of natural language inference problems that aims to test the linguistic competencies of lexical semantics, predicate-argument structure, logic and knowledge, and commonsense. From this diagnostic, they chose items that belonged to a single subcategory from a collection of seven sub-categories:

1. Morphological negations – tests for reasoning ability with negation
2. Prepositional phrases – tests for the ability to understand sentences with preposition modifiers
3. Lexical entailment – tests for the relationship between words, i.e., hypernyms, hyponyms,
4. Quantifiers – tests for the ability to reason over operators of quantity
5. Propositional structure – tests for the ability to reason over lexical logical operators, i.e., conjunction, conditionals, etc.
6. Richer logical structure – tests for the ability to reason over higher-order lexical logical forms
7. World knowledge – tests for knowledge of facts about the world

Then they evaluated a number of distinct (across some dimension) transformer language models [121] and long short-term memory (LSTM) [19] models using their selected items. Each language model represented a theoretical individual. In addition to evaluating the language models, they also evaluated humans using the GLUE dataset, gathering in total 240 language model participants and 27 human participants. Finally, these responses are analyzed using psychometric classical test theory and item response theory methods. Specifically, they set about to estimate the *simple difficulty* of questions (how many human participants would get a question correct).

Their results show a significant correlation between the item difficulties (as measured by human responses) and transformer language model predictions of difficulties. With their classical test theory (CTT) and IRT analysis, the data in Figures 3.1 and 3.2 show that TLMs responses predict, with a significant probability above random chance, the psychometric test difficulty for linguistic items on all the subcategory tests of linguistic ability except for, morphological negation and richer logical structure.

This work is significant because it can help us to build a validated understanding of, if not an algorithm's high-dimensional space, its practical similarity to ours. In this specific work, Laverghetta Jr. et al., mention that this correlation between human subjects and LMs means that LM models can be used in place of human participants to build psychometric tests. We argue that furthermore, work like this allows for significant steps to be made towards models of generalization in AI; specifically, that the testing data is supported with psychometrics. This approach to evaluating models, if done across a wide variety of tasks using psychometrics, can provide the interpretive framework to understand and model AI generalization. Additionally, work like this already begins to help the community towards a more practical understanding of generalization in AI. It presents, in regard to linguistic capabilities, the contexts in which it is okay to project the qualities of human thought onto an algorithm's output or computation process and therefore, also, when its processing must be treated as distinct from human thinking.

3.4 Better Training and Learning with Psychometrics

Since psychometric evaluations enforce rigor in testing, the meaning of the data obtained from a psychometric evaluation allows a test administrator to better understand shortcomings in the learning or conceptualization process for the test-taker [49-53]. These understandings have led to improvements in education [105], and we argue that similar improvements would happen if psychometric evaluations were used to test AI model performance. Since the evaluations are more rigorous, the results obtained could help researchers to better understand why a model is not generalizing or fitting to the task it is being trained on.

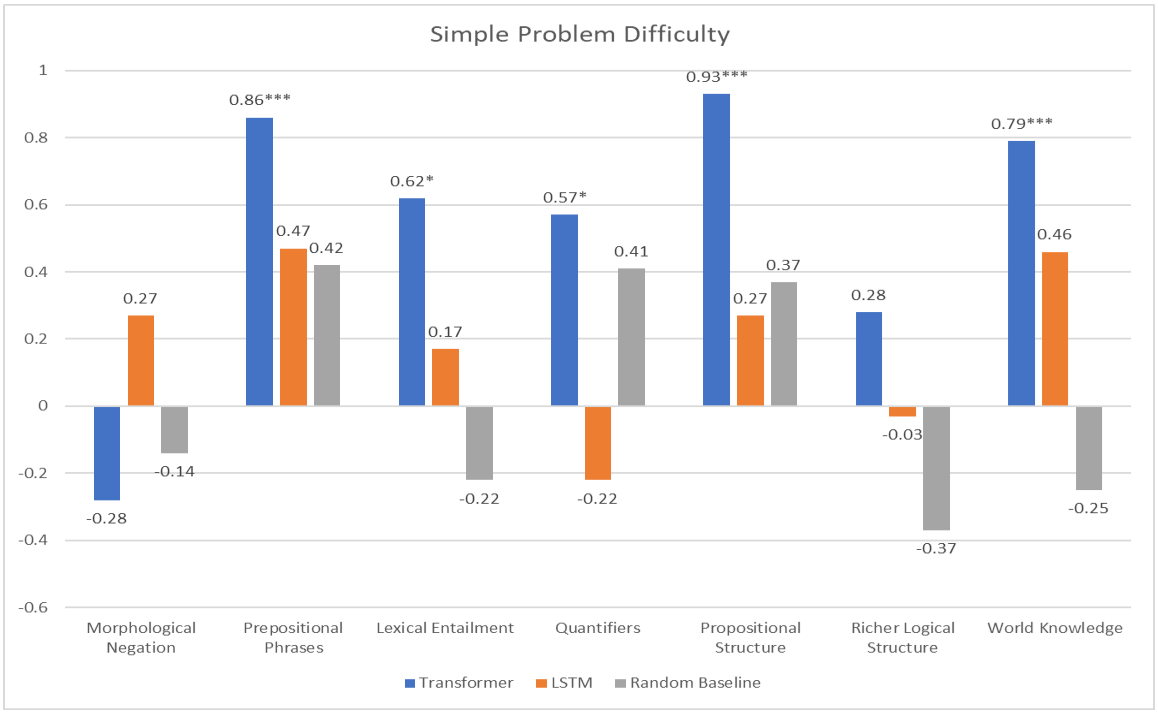

Figure 3.1: Spearman correlation was calculated with TLMs, LSTMs, and a Random LM estimate of problem difficulty for the Classical Test Theory

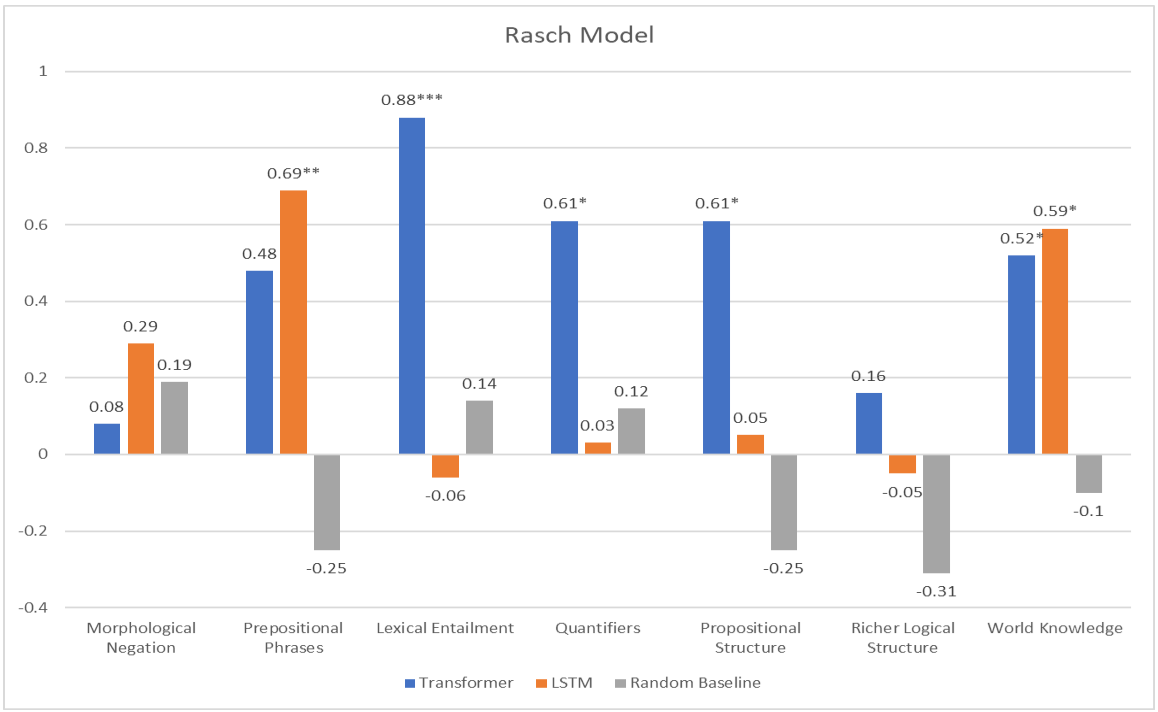

Figure 3.2: Spearman correlation was calculated with TLMs, LSTMs, and a Random LM estimate of problem difficulty for the Item Response Theory

Understanding why algorithms don't learn certain concepts across a variety of domains, with psychometric evaluations, would help the community to understand broadly, what current state-of-the-art algorithms are missing in their architecture. It would also help improve the performance of algorithms in their task-specific domains, giving researchers the insight they need to improve the training environment in an effective way to improve the algorithm's learning.

3.5 Counterargument: But AI May Not Generalize Like Humans Do

One counterargument to the proposed argument is that AI models may not reason, generalize, or abstract in any way similar to humans, therefore psychometrics would be ill-equipped to provide data that would lead to generalization theories for AI models. If AI models don't generalize in any similar way to humans, there may be no underlying theory of generalization, or their generalization ability may just be incomprehensible to humans.

However, we argue that whether AI models mimic the human mind or not, psychometrics still will provide insights into model generalization. Firstly, psychometrics has shown it provides the methodology to measure ability within the psychological space. The psychological space is inherently intangible and, in some ways, unconceptualizable to humans, which makes it difficult to model. The psychological space is theorized to be a high-dimensional space of consequential objects, which is similar in a sense to the fact that state-of-the-art AI models also utilize high-dimensional spaces of consequential objects. In both the mind and AI algorithms' dimensional space, the exact working and occurrences in this space are unable to be observed directly or easily modeled. Yet, psychometrics has been able to glean reliable measurements of constructs within the psychological space of the mind, and we propose that it will be able to glean reliable measurements of the occurrences in the high-dimensional space of modern AI as well. Additionally, the requirements and rigor of psychometric evaluations will nevertheless improve the quality of evaluation occurring in the field which still is a positive contribution to the field.

3.6 Counterargument: The AI Community Has Made Large Strides in Generalization Without Adopting Psychometric Rigor

Another counterargument to the argument that adopting psychometrics can lead to improvements in learning and generalization in AI is that large progress in AI has been made so far without adopting psychometrics. Psychometrics brings with its adoption a decrease in research output, an increase in the difficulty of research, and a decrease in fervor around claims of AI performance. One can ask, "Are these disadvantages worth it when AI has made such big leaps without psychometrics?"

Additionally, one can argue that adopting psychometrics could slow the progress made in generalization performance. The AI community currently tackles problems in this space from a variety of vantage points; however, adopting psychometrics would lead to a more singular approach to improvement. Tackling problems from a variety of vantage points is more likely to lead to scientific paradigm shifts in the community, versus a singular methodical approach.

Nevertheless, we argue that psychometrics can help the community to improve generalization performance in AI, despite any disadvantages it may bring. The methodical evaluation of generalization will provide the community with clear metrics to improve upon. Additionally, there will still be room for creative approaches to solving problems in the field; and additionally, more valid empirical work.

3.7 Summary

In this chapter, we argued that adopting psychometrics in AI can lead to improvements in the learning and generalization performance of AI. First, we overviewed generalization from the psychological and artificial intelligence perspectives. Then, we showed: psychometrics has been extensively used in testing generalization ability, led to theories of generalization in psychology, and helps to improve education. These indicate that psychometrics can similarly help the artificial intelligence community since we are struggling with defining and modeling the generalization capabilities of modern AI networks. Additionally, we discuss [133] and how they utilized psychometrics to show that TLMs could be used to predict psychometric item difficulty for linguistic capability; their work presents a method with which

generalization theories can emerge. Finally, we discuss counterarguments that AI may not generalize in any way similar to humans and that AI has not needed psychometrics thus far to make progress.

Chapter 4: Towards Explainable AI with Psychometrics

In this chapter, we argue that psychometrics, by the nature of its philosophy and its methods, can help the AI community to build more explainable and interpretable AI. Firstly, we present a brief overview of Explainable AI (XAI), namely, its goals, its challenges, and its importance. Then, we reference work that shows how psychometrics allows for the interpretation and explanation of test-respondents abilities. Furthermore, we present references relative to the psychometric methods of item analysis, proposing therefrom, that item analytic methods can provide small constituents of explainable data, that can be composed into mini-tests and tests that express systematic explanations and interpretations of an algorithm. Afterward, we discuss a counterargument that psychometrics does not directly allow an algorithm to explain its understandings, and therefore an interpretation and explanation still heavily depend on the human in the loop and their subsequent knowledge, which in fact, does not guarantee an accurate explanation or interpretation. Finally, the chapter is concluded with a summary.

4.1 Explainable AI

As AI becomes more ubiquitous, its use, especially in critical applications, requires that users can interpret, understand, trust, and explain an AI's output. Explainable AI (XAI) is an AI system with behaviors and outputs that are understandable by humans [65-67].

XAI systems are expected to be able to explain their capabilities, understandings, what they have done, what they are currently and will be doing, and what information they are using for their behavior with a suitable level of explainability/interpretability for the domain/context of the system.

4.1.1 Why XAI Is Important

Achieving good XAI systems is important for the use of AI in safety-critical, health, and financial applications. AI systems that are not explainable cannot be used in domains where an incorrect decision leads to significant financial losses, health damages, or loss of lives. XAI allows users to trust an AI system

and also understand why it is recommending or choosing a particular output [65-67]. Being able to understand this would open new domains to the benefits of AI systems, allowing users to make the correct decision despite an AI shortcoming and allowing researchers and developers to understand and improve limitations in the AI models.

4.1.2 Major Challenges in Building XAI

From a technical standpoint, one major challenge in achieving XAI is that AI models cannot argue, explain, and defend their decisions in a dialogical way [68, 111-112]. Another major challenge is that an objective measure of explainability or interpretability does not exist in the literature [69].

4.2 The Nature of Psychometric Evaluations will Increase the Interpretability and Explainability of AI

One way that psychometrics can help with building more explainable and interpretable AI is that psychometric evaluations allow test creators and test administrators to interpret and explain the capabilities and traits of a test-taker [40-43]. The rigor involved in creating the items and the overall test brings with it an increased interpretability of the results. Psychometric tests are iteratively improved until they believably predict or measure some capability or behavior in the test subject. This allows psychometric tests to be used for job candidates, school admissions, and mental health evaluations [1-11, 40-43]; observers of the results can interpret how well an applicant or patient fits a certain criterion.

Similarly, employing psychometric evaluations in AI can bring similar results. Test suites built with psychometrics, and extrinsic evaluations validated with psychometrics would increase the interpretability of an algorithm's evaluation results. Psychometric evaluations bring with them an increased precision in the evaluation of psychological (and hypothetically high dimensional) conceptualization testing. This increased precision would give the community a more robust foundation with which to interpret and infer algorithms' decision-making.

The previous chapters both capture some of the essences of how psychometrics can improve the interpretability and explainability of models. In Chapter 2, we discussed some works that utilized IRT to create more informative evaluations [120, 127-130]. By utilizing IRT, researchers were able to delineate

from their evaluations, more information with which to interpret and explain models. Also, in [133], using a psychometrically validated dataset allowed the researchers to better analyze and, in some sense, understand and explain some of TLMs' high-dimensional conceptual space. Additionally, we argue, that as researchers create training datasets of items psychometrically discriminated for certain features, the interpretability of a model's output and learning process will also increase.

4.3 Psychometric Methods Allow for a System of Explainability of Evaluations

Another way that psychometrics can help with building more explainable AI is through the use of psychometric item analysis methods. During test creation, psychometricians may utilize item response theory and item analysis to create items on a test that provide an explanation of some specific capability or construct [44-48]. For example, using item analysis would mean testing and analyzing specific items in a dataset to determine how feature-rich they are for an algorithm. Furthermore, because one is using and analyzing a small dataset first, before using big data, one can explain more insightfully, what an algorithm is focusing on and learning from the dataset. Building a theory of what is salient to the algorithm from a small item analyzed dataset, allows for the big dataset to be more informative, since one can find out how correlated the other items in it are to the original item analyzed dataset.

Regarding testing, one can, because of the item analysis performed on the dataset throughout the process, create tests of specific constructs in addition to the overall testing of the algorithm on a test dataset. To put this more generally, psychometric methods allow for individual items and groups of items to be explainable of some concept or ability through its item analysis methods. These smaller items of explainability can be combined to form a more systemized explanation. For example, in [133], Laverghetta Jr. Et al., apply some psychometric item analysis, by testing which items were most discriminative and indicative of learning occurring. This led to them choosing more informative items with which to test humans and the TLMs. We argue that this can be taken even further, where psychometrics is used to rigorously analyze items in large datasets for their meaning to the algorithm to determine whether the item should be included in a more focused dataset of some particular task or ability. These datasets of psychological constructs can then be compiled into a combinatorial test of some practical ability. For

example, psychometrics can inform the creation of a dataset of stop scenarios for self-driving cars. The items compromising this dataset would be psychometrically validated on AI to be feature-rich and discriminatory. Then another dataset of turn scenarios could be created in the same way and so on. These datasets are then combined into a "simulative test of driving ability," with which a score on each scenario context is generated. These generated scores of smaller concepts, with regards to driving, which themselves are composed of items that are salient in some discerned way, would combine into an explanative test of driving ability. This is just but one hypothetical example of how using psychometrics to inform the evaluation of AI will allow researchers to provide a systemized explanation of an algorithm's dimensional space, through inferences of smaller explanations gained from psychometrically created items/mini-tests.

4.4 Counterargument: Why Psychometrics May Not Provide More Explainability to AI

Psychometric evaluations themselves don't allow an AI model to explain its understanding, capabilities, or what it has, is, and will do. Instead, psychometric evaluations make it possible for researchers and other users of AI systems to better infer and interpret model behavior and output. This is similar to what occurs when one does not use psychometric evaluations. Human beings are naturally meaning-finding creatures, and we can overlay meaning even in meaningless contexts. Psychometrics would simply provide a richer context within which users could infer interpretations or explanations; however, it would not provide the AI systems with the capabilities to explain their understandings, capabilities, and output decisions.

However, even though psychometrics will not directly allow AI systems to explain themselves, it would still improve the ability of users to make more valid inferences about an algorithm's capabilities and conceptions, which would still improve the overall explainability and interpretability of AI. Additionally, the increase in explainability and interpretability from psychometrics could also lead to self-explainable XAI since psychometrics would improve the ability of researchers to understand and more critically debug AI algorithms.

4.5 Summary

In this chapter, we argued that psychometrics can help with building more explainable and interpretable AI. We presented an overview of XAI, explaining what it is, and the general goals of the field. Then, we discussed how psychometric evaluations inherently provide a lot of useful information to test administrators of test-taker capabilities and conceptual understandings due to the rigor with which psychometricians create psychometric tests. Additionally, we hypothesize that psychometric items can provide small explainable insights that, when taken together, can provide a systematic explanation of an algorithm. Afterward, we discuss the counterpoint that psychometric evaluations may not contribute directly to an AI's ability to explain itself or argue its point, however, despite this, it would still improve the explainability and interpretability of AI algorithms.

Chapter 5: Conclusion

In this work, we presented and discussed research and presented arguments that show how psychometrics and its methods, principles, and guidelines, can help the AI community to overcome its challenges in building more task-generalized and explainable AI algorithms and in testing and measuring the black box performance of AI models.

First, we discussed works and arguments that show that psychometrics can help the AI field to design better evaluations of models. Across a variety of domains, psychometrics has designed evaluations and tests that can test the abilities, capabilities, and personality traits of the black-boxed human mind. Additionally, psychometrics, and its long history of testing and evaluating psychological concepts, has obtained validity as a science and has garnered acceptance by the scientific and broader community. Furthermore, we discussed and presented works that indicate that current evaluation methods in AI are too general as evaluations, which leads to skepticism and doubts about the performance of AI algorithms and models at best, and catastrophes such as the loss of life at worst. Then, we presented hypothetical examples of how psychometrics can help the AI community in building better datasets and evaluations; additionally, presenting some work that has already began using IRT for more informative evaluations. Afterward, we discussed counterarguments such as that psychometrics can slow down research progress in AI, and that psychometric tests may be gameable by AI algorithms, and in response to these, propose that a reduction in publication rate doesn't have to be all bad for the AI community and that current evaluation methods are gameable as well and therefore can be improved with psychometrics.

Secondly, we discussed works and arguments that indicate that psychometric datasets built from the psychometric evaluation of models can lead to improvements in the learning and generalization capabilities of AI. Firstly, we presented works that show that psychometrics has widely explored generalization concepts via IQ and analogy tests. Then, we showed and discussed how psychometric testing

led to theories of generalization in humans. Additionally, we showed and discussed how psychometric testing has led to measurable improvements in education and how similar results can occur in AI. Then we explored work, which demonstrates how psychometrics can inform generalization models of AI. After all of this, we discussed counterarguments such as that AI algorithms may not generalize in any way similar to humans, and that the AI community has been able to make outstanding progress so far without psychometrics. These counterarguments indicate that psychometrics may not lead to improvements in learning and generalization performance in AI. However, in response, we propose that whether AI models mimic human generalization or not, psychometrics can extract measurements from high-dimensional spaces which will still make it useful in AI. Additionally, in response, we argue that psychometrics can help the field to make continuous, iterative progress without the risks of stagnation occurring due to a non-methodical process of algorithm improvement.

Finally, we discussed works and presented arguments that indicate that psychometrics can provide explainability to AI models. We presented works that show how psychometrics allows for the interpretation and explanation of a test-taker's capabilities. Then, we presented some works on psychometric items and discussed how explainable items created with the psychometric methods, can be combined to create a systematic explanation of capability or understanding, which would be useful for XAI. Then, we tempered expectations with a counterargument discussing how psychometrics do not directly allow algorithms to self-explain or justify their outputs and decision; nevertheless, psychometric evaluations would increase the explainability of models which could indirectly help create AI that can better explain itself.

All of these topics were brought up to show that psychometrics can help the AI community to overcome its challenges in building more task-generalizable and explainable AI algorithms and in more accurately evaluating and measuring the performance of algorithms. We hope that this work encourages further investigation into how psychometric methods and principles can contribute to AI progress and help lead to more adaptive and explainable models for the good of society. Future research directions could be building a corpus composed of psychometrically informed smaller items that combine into larger categories of classification or building a corpus of correlated tasks for a common task in AI, i.e., self-driving cars.

Ultimately, we hope this work has shown the analogical relationship between the high-dimensional space of AI algorithms and the theorized high-dimensional space of the mind [20-31], and how psychometrics has some insights that may help the AI community move towards building better models.

References

[1] Ramdani, Z., Marliani, R., & Rahman, A. A. (2019). The individual work performance scale: A psychometric study and its application for employee performance. *Humanities & Social Sciences Reviews*, *7*(5), 405-414.

[2] Clark, C. M., Sattler, V. P., & Barbosa-Leiker, C. (2018). Development and psychometric testing of the Workplace Civility Index: A reliable tool for measuring civility in the workplace. *The Journal of Continuing Education in Nursing*, *49*(9), 400-406.

[3] Schlegel, K., & Mortillaro, M. (2019). The Geneva Emotional Competence Test (GECo): An ability measure of workplace emotional intelligence. *Journal of applied psychology*, *104*(4), 559.

[4] Osadebe, P. U., & Nwabeze, C. P. (2018). Construction and validation of physics aptitude test as an assessment tool for senior secondary school students. *International Journal of Assessment Tools in Education*, *5*(3), 461-473.

[5] Lee, J., Park, C. G., Kim, S. H., & Bae, J. (2021). Psychometric properties of a clinical reasoning assessment rubric for nursing education. *BMC nursing*, *20*(1), 1-9.

[6] Anunciacao, L. (2018). An Overview of the History and Methodological Aspects of Psychometrics: History and Methodological aspects of Psychometrics. *Journal for ReAttach Therapy and Developmental Diversities*, *1*(1), 44-58.

[7] Chessa, E., Piga, M., Floris, A., Devilliers, H., Cauli, A., & Arnaud, L. (2020). Use of Physician Global Assessment in systemic lupus erythematosus: a systematic review of its psychometric properties. *Rheumatology*, *59*(12), 3622-3632.

[8] Sanchez-Balcells, S., Callarisa Roca, M., Rodriguez-Zunino, N., Puig-Llobet, M., Lluch-Canut, M. T., & Roldan-Merino, J. F. (2018). Psychometric properties of instruments measuring quality and satisfaction in mental health: A systematic review. *Journal of advanced nursing*, *74*(11), 2497-2510.

[9] Moon, S. J., Hwang, J. S., Kim, J. Y., Shin, A. L., Bae, S. M., & Kim, J. W. (2018). Psychometric properties of the Internet Addiction Test: A systematic review and meta-analysis. *Cyberpsychology, Behavior, and Social Networking*, *21*(8), 473-484.

[10] Fabbri, B., Berardi, A., Tofani, M., Panuccio, F., Ruotolo, I., Sellitto, G., & Galeoto, G. (2021). A systematic review of the psychometric properties of the Jebsen–Taylor Hand Function Test (JTHFT). *Hand Surgery and Rehabilitation*, *40*(5), 560-567.

[11] Bennett, R. E., & von Davier, M. (2017). Advancing human assessment: The methodological, psychological and policy contributions of ETS (p. 711). Springer Nature.

[12] Mitchell, M. (2021). Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, *1505*(1), 79-101.

[13] Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006, December). Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence* (pp. 1015-1021). Springer, Berlin, Heidelberg.

[14] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information processing & management*, *45*(4), 427-437.

[15] Tharwat, A. (2020). Classification assessment methods. *Applied Computing and Informatics*.

[16] García, S., Fernández, A., Luengo, J., & Herrera, F. (2009). A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability. *Soft Computing*, *13*(10), 959-977.

[17] Luque, A., Carrasco, A., Martín, A., & de Las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, *91*, 216-231.

[18] Thomas, R. L., & Uminsky, D. (2022). Excerpt from Reliance on Metrics is a Fundamental Challenge for AI. In *Ethics of Data and Analytics* (pp. 342-349). Auerbach Publications.

[19] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[20] Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137-141.

[21] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, *6*, 52138-52160.

[22] Pedreschi, D., Giannotti, F., Guidotti, R., Monreale, A., Ruggieri, S., & Turini, F. (2019, July). Meaningful explanations of black box AI decision systems. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 9780-9784).

[23] Castelvecchi, D. (2016). Can we open the black box of AI?. *Nature News*, *538*(7623), 20.

[24] Durán, J. M., & Jongsma, K. R. (2021). Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics*, *47*(5), 329-335.

[25] Carabantes, M. (2020). Black-box artificial intelligence: an epistemological and critical analysis. *AI & society*, *35*(2), 309-317.

[26] Kripke, S. A. (1982). *Wittgenstein on rules and private language: An elementary exposition*. Harvard University Press.

[27] Hamlyn, D. W. (1990). *In and out of the black box: On the philosophy of cognition*. Basil Blackwell.

[28] Donald, B. B., & Bakies, E. (2016). A Glimpse Inside the Brain's Black Box: Understanding the Role of Neuroscience in Criminal Sentencing. *Fordham L. Rev.*, *85*, 481.

[29] Székely, G. (2001). An approach to the complexity of the brain. *Brain research bulletin*, *55*(1), 11-28.

[30] Niv, Y. (2021). The primacy of behavioral research for understanding the brain. *Behavioral Neuroscience*.

[31] Gallagher, S. (2018). Decentering the brain: Embodied cognition and the critique of neurocentrism and narrow-minded philosophy of mind.

[32] Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317-1323.

[33] Chater, N., & Vitányi, P. M. (2003). The generalized universal law of generalization. *Journal of Mathematical Psychology*, *47*(3), 346-369.

[34] Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and brain sciences*, *24*(4), 629-640.

[35] Schwegman, K. (2022). The use of psychometric test systems as a pre-selection tool for identifying successful harvesting machine operators.

[36] Mühling, A., Ruf, A., & Hubwieser, P. (2015, November). Design and first results of a psychometric test for measuring basic programming abilities. In *Proceedings of the workshop in primary and secondary computing education* (pp. 2-10).

[37] Vierula, J., Talman, K., Hupli, M., Laakkonen, E., Engblom, J., & Haavisto, E. (2021). Development and psychometric testing of Reasoning Skills test for nursing student selection: An item response theory approach. *Journal of Advanced Nursing*, *77*(5), 2549-2560.

[38] Schubert, A. L., & Frischkorn, G. T. (2020). Neurocognitive psychometrics of intelligence: How measurement advancements unveiled the role of mental speed in intelligence differences. *Current Directions in Psychological Science*, *29*(2), 140-146.

[39] Kovacs, K., & Conway, A. R. (2019). A unified cognitive/differential approach to human intelligence: Implications for IQ testing. *Journal of Applied Research in Memory and Cognition*, *8*(3), 255-272.

[40] Hughes, D. J. (2018). Psychometric validity: Establishing the accuracy and appropriateness of psychometric measures. *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development*, 751-779.

[41] Hammond, S. (2006). Using psychometric tests. *Research methods in psychology*, *3*, 182-209.

[42] Litwin, M. S., & Fink, A. (2003). *How to assess and interpret survey psychometrics* (Vol. 8). Sage.

[43] Geisinger, K. F. (1998). Psychometric issues in test interpretation.

[44] Bringsjord, S., & Schimanski, B. (2003, August). What is artificial intelligence? Psychometric AI as an answer. In *IJCAI* (pp. 887-893).

[45] Jensen, A. R. (2002). Psychometric g: Definition and substantiation. In *The general factor of intelligence* (pp. 51-66). Psychology Press.

[46] Kranzler, J. H., & Jensen, A. R. (1991). The nature of psychometric g: Unitary process or a number of independent processes?. *Intelligence*, *15*(4), 397-422.

[47] Bringsjord, S. (2011). Psychometric artificial intelligence. *Journal of Experimental & Theoretical Artificial Intelligence*, *23*(3), 271-277.

[48] Bringsjord, S., & Licato, J. (2012). Psychometric artificial general intelligence: the Piaget-MacGuyver room. In *Theoretical foundations of artificial general intelligence* (pp. 25-48). Atlantis Press, Paris.

[49] Shayer, M. (2008). Intelligence for education: As described by Piaget and measured by psychometrics. *British Journal of Educational Psychology*, *78*(1), 1-29.

[50] Abedi, J. (2006). Psychometric issues in the ELL assessment and special education eligibility. *Teachers College Record*, *108*(11), 2282-2303.

[51] Kolen, M. J. (2017). *Enhancing assessment in higher education: Putting psychometrics to work*. Stylus Publishing, LLC.

[52] Hart, P. L., Spiva, L., & Kimble, L. P. (2011). Nurses' knowledge of heart failure education principles survey: A psychometric study. *Journal of clinical nursing*, *20*(21-22), 3020-3028.

[53] Niksadat, N., Rakhshanderou, S., Negarandeh, R., Ramezankhani, A., Vasheghani Farahani, A., & Ghaffari, M. (2019). Development and psychometric evaluation of Andragogy-based Patient Education Questionnaire (APEQ). *American Journal of Health Education*, *50*(6), 390-397.

[54] Laverghetta Jr, A., Mirzakhalov, J., & Licato, J. (2020, December). Towards a task-agnostic model of difficulty estimation for supervised learning tasks. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop* (pp. 16-23).

[55] Hernández-Orallo, J., Dowe, D. L., & Hernández-Lloreda, M. V. (2014). Universal psychometrics: Measuring cognitive abilities in the machine kingdom. *Cognitive Systems Research*, *27*, 50-74.

[56] Hernández-Orallo, J. (2017). Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. *Artificial Intelligence Review*, *48*(3), 397-447.

[57] Besold, T., Hernández-Orallo, J., & Schmid, U. (2015). Can machine intelligence be measured in the same way as human intelligence?. *KI-Künstliche Intelligenz*, *29*(3), 291-297.

[58] Liu, Y., He, F., Zhang, H., Rao, G., Feng, Z., & Zhou, Y. (2019). How Well Do Machines Perform on IQ tests: a Comparison Study on a Large-Scale Dataset. In *IJCAI* (pp. 6110-6116).

[59] Ohlsson, S., Sloan, R. H., Turán, G., & Urasky, A. (2017). Measuring an artificial intelligence system's performance on a verbal IQ test for young children. *Journal of Experimental & Theoretical Artificial Intelligence*, *29*(4), 679-693.

[60] Wang, H., Tian, F., Gao, B., Bian, J., & Liu, T. Y. (2015). Solving verbal comprehension questions in IQ test by knowledge-powered word embedding. *arXiv preprint arXiv:1505.07909*.

[61] Levin, E., Tishby, N., & Solla, S. A. (1990). A statistical approach to learning and generalization in layered neural networks. *Proceedings of the IEEE*, *78*(10), 1568-1574.

[62] Chatzikyriakidis, S., Cooper, R., Dobnik, S., & Larsson, S. (2017). An overview of Natural Language Inference Data Collection: The way forward?. In Proceedings of the Computing Natural Language Inference Workshop.

[63] Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, *64*(3), 107-115.

[64] Jakubovitz, D., Giryes, R., & Rodrigues, M. R. (2019). Generalization error in deep learning. In *Compressed sensing and its applications* (pp. 153-193). Birkhäuser, Cham.

[65] Gunning, D. (2017). Explainable artificial intelligence (xai). *Defense advanced research projects agency (DARPA), nd Web*, *2*(2), 1.

[66] Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, *6*, 52138-52160.

[67] Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. *Science robotics*, *4*(37), eaay7120.

[68] Čyras, K., Rago, A., Albini, E., Baroni, P., & Toni, F. (2021). Argumentative XAI: a survey. *arXiv preprint arXiv:2105.11266*.

[69] Nguyen, V. B., Schlötterer, J., & Seifert, C. (2022). Explaining Machine Learning Models in Natural Conversations: Towards a Conversational XAI Agent. *arXiv preprint arXiv:2209.02552*.

[70] Raskin S. (2022). US vehicle safety watchdog investigating Tesla crash that killed 2 in Florida. New York Post

[71] Mitchell, M. (2021). Abstraction and analogy-making in artificial intelligence. *Annals of the New York Academy of Sciences*, *1505*(1), 79-101.

[72] Jenkins, O. C., Lopresti, D., & Mitchell, M. (2020). Next Wave Artificial Intelligence: Robust, Explainable, Adaptable, Ethical, and Accountable. *arXiv preprint arXiv:2012.06058*.

[73] Rust, J., & Golombok, S. (2021). *Modern psychometrics: The science of psychological assessment*. Routledge.

[74] Woo, S. E., LeBreton, J., Keith, M., & Tay, L. (2020). Bias, fairness, and validity in graduate admissions: A psychometric perspective. *Preprint*.

[75] Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: theory and application. *The American journal of medicine*, *119*(2), 166-e7.

[76] Shrout, P. E., & Lane, S. P. (2012). Psychometrics.

[77] Kyriazos, T. A., & Stalikas, A. (2018). Applied psychometrics: The steps of scale development and standardization process. *Psychology*, *9*(11), 2531.

[78] Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

[79] Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., ... & Anderljung, M. (2020). Toward trustworthy AI development: mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213*.

[80] Campbell, T., Dollaghan, C., Needleman, H., & Janosky, J. (1997). Reducing bias in language assessment: Processing-dependent measures. *Journal of Speech, Language, and Hearing Research*, *40*(3), 519-525.

[81] Poortinga, Y. H. (1995). Cultural bias in assessment: Historical and thematic issues. *European Journal of Psychological Assessment*, *11*(3), 140.

[82] Helms, J. E. (2006). Fairness is not validity or cultural bias in racial-group assessment: A quantitative perspective. *American Psychologist*, *61*(8), 845.

[83] Peter, J. P. (1979). Reliability: A review of psychometric basics and recent marketing practices. *Journal of marketing research*, *16*(1), 6-17.

[84] Hubley, A. M., & Zumbo, B. D. (2013). Psychometric characteristics of assessment procedures: An overview.

[85] Moses, T. (2017). A review of developments and applications in item analysis. *Advancing Human Assessment*, 19-46.

[86] Ricketts, C. (2009). A plea for the proper use of criterion-referenced tests in medical assessment. Medical education, 43(12), 1141-1146.

[87] Brown, J. D., & Hudson, T. (2002). Criterion-referenced language testing. Cambridge University Press.

[88] Shayer, M. (2008). Intelligence for education: As described by Piaget and measured by psychometrics. British Journal of Educational Psychology, 78(1), 1-29.

[89] Embretson, S. E., & Reise, S. P. (2013). Item response theory. Psychology Press.

[90] Mishra, S., Arunkumar, A., Bryan, C., & Baral, C. (2020). Our evaluation metric needs an update to encourage generalization. arXiv preprint arXiv:2007.06898.

[91] Thomas, R., & Uminsky, D. (2020). The problem with metrics is a fundamental problem for AI. arXiv preprint arXiv:2002.08512.

[92] Cabitza, F., Campagner, A., & Sconfienza, L. M. (2020). As if sand were stone. New concepts and metrics to probe the ground on which to build trustable AI. BMC Medical Informatics and Decision Making, 20(1), 1-21.

[93] Thomas, R. L., & Uminsky, D. (2022). Reliance on metrics is a fundamental challenge for AI. Patterns, 3(5), 100476.

[94] Chen, T., Jiang, Z., Poliak, A., Sakaguchi, K., & Van Durme, B. (2019). Uncertain natural language inference. arXiv preprint arXiv:1909.03042.

[95] Lever, J. (2016). Classification evaluation: It is important to understand both what a classification metric expresses and what it hides. *Nature methods*, *13*(8), 603-605.

[96] Powers, D. M. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

[97] Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, *5*(2), 1.

[98] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461.

[99] Chatzikyriakidis, S., Cooper, R., Dobnik, S., & Larsson, S. (2017). An overview of Natural Language Inference Data Collection: The way forward?. In Proceedings of the Computing Natural Language Inference Workshop.

[100] The Associated Press (2022). Nearly 400 car crashes in 11 months involved automated tech, companies tell regulators. *NPR*

[101] Zhou, Z. Q., & Sun, L. (2019). Metamorphic testing of driverless cars. *Communications of the ACM*, *62*(3), 61-67.

[102] Stilgoe, J. (2020). Who Killed Elaine Herzberg?. In *Who's Driving Innovation?* (pp. 1-6). Palgrave Macmillan, Cham.

[103] Fehlmann, T. (2019, September). Testing artificial intelligence. In *European Conference on Software Process Improvement* (pp. 709-721). Springer, Cham.

[104] Price, L. R. (2016). *Psychometric methods: Theory into practice*. Guilford Publications.

[105] Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics*. Cambridge University Press.

[106] Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *Journal of educational data mining*, *4*(1), 11-48.

[107] Ritchie, S. (2015). *Intelligence: All that matters*. John Murray.

[108] McDuff, D., Cheng, R., & Kapoor, A. (2018). Identifying bias in AI using simulation. arXiv preprint arXiv:1810.00471.

[109] Roselli, D., Matthews, J., & Talagala, N. (2019, May). Managing bias in AI. In *Companion Proceedings of The 2019 World Wide Web Conference* (pp. 539-544).

[110] Thesing, L., Antun, V., & Hansen, A. C. (2019). What do AI algorithms actually learn?-On false structures in deep learning. arXiv preprint arXiv:1906.01478.

[111] Saeed, W., & Omlin, C. (2021). Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. *arXiv preprint arXiv:2111.06420*.

[112] Marques-Silva, J., & Ignatiev, A. (2022). Delivering Trustworthy AI through formal XAI. In *Proc. of AAAI* (pp. 3806-3814).

[113] Risi, S., & Preuss, M. (2020). From chess and atari to starcraft and beyond: How game ai is driving the world of ai. *KI-Künstliche Intelligenz*, *34*(1), 7-17.

[114] Brown, N., & Sandholm, T. (2018). Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, *359*(6374), 418-424.

[115] Liu, Y., He, F., Zhang, H., Rao, G., Feng, Z., & Zhou, Y. (2019). How Well Do Machines Perform on IQ tests: a Comparison Study on a Large-Scale Dataset. In *IJCAI* (pp. 6110-6116).

[116] Pisano, E. D. (2020). AI shows promise for breast cancer screening.

[117] Elkins, K., & Chun, J. (2020). Can GPT-3 pass a Writer's Turing test?. *Journal of Cultural Analytics*, *5*(2), 17212.

[118] Krol, S. J., Llano, M. T., & McCormack, J. (2022). Towards the Generation of Musical Explanations with GPT-3. In *International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar)* (pp. 131-147). Springer, Cham.

[119] Bringsjord, S., & Licato, J. (2012). Psychometric artificial general intelligence: the Piaget-MacGuyver room. In *Theoretical foundations of artificial general intelligence* (pp. 25-48). Atlantis Press, Paris.

[120] Chmait, N., Dowe, D. L., Li, Y. F., & Green, D. G. (2017, August). An information-theoretic predictive model for the accuracy of AI agents adapted from psychometrics. In *International conference on artificial general intelligence* (pp. 225-236). Springer, Cham.

[121] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[122] Turing, A. M. (2009). Computing machinery and intelligence. In *Parsing the Turing test* (pp. 23-65). Springer, Dordrecht.

[123] Strathern, M. (1997). 'Improving ratings': Audit in the British University system. *European Review*, *5*(3), 305-321. doi:10.1002/(SICI)1234-981X(199707)5:33.0.CO;2-4

[124] Hadj Taieb, M. A., Zesch, T., & Ben Aouicha, M. (2020). A survey of semantic relatedness evaluation datasets and procedures. *Artificial Intelligence Review*, *53*(6), 4407-4448.

[125] Wiegreffe, S., & Marasović, A. (2021). Teach me to explain: A review of datasets for explainable NLP. *arXiv preprint arXiv:2102.12060*.

[126] El Dehaibi, N., & MacDonald, E. F. (2020, May). INVESTIGATING INTER-RATER RELIABILITY OF QUALITATIVE TEXT ANNOTATIONS IN MACHINE LEARNING DATASETS. In *Proceedings of the Design Society: DESIGN Conference* (Vol. 1, pp. 21-30). Cambridge University Press.

[127] Martínez-Plumed, F., Prudêncio, R. B., Martínez-Usó, A., & Hernández-Orallo, J. (2019). Item response theory in AI: Analysing machine learning classifiers at the instance level. *Artificial intelligence*, *271*, 18-42.

[128] John P. Lalor, Hao Wu, and Hong Yu. 2016. Building an Evaluation Scale using Item Response Theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 648–657, Austin, Texas. Association for Computational Linguistics.

[129] Martínez-Plumed, F., Prudêncio, R. B., Martínez-Usó, A., & Hernández-Orallo, J. (2016). Making sense of item response theory in machine learning. In *ECAI 2016* (pp. 1140-1148). IOS Press.

[130] Lalor, J., Wu, H., & Yu, H. Beyond Majority Voting: Generating Evaluation Scales using Item Response Theory. Association for Computational Linguistics.

131] Banich, M. T., & Caccamise, D. (2011). *Generalization of knowledge: Multidisciplinary perspectives*. Psychology Press.

[132] Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*.

[133] Laverghetta, A., Nighojkar, A., Mirzakhalov, J., & Licato, J. (2022). Predicting Human Psychometric Properties Using Computational Language Models. In The Annual Meeting of the Psychometric Society (pp. 151-169). Springer, Cham.

[134] Richardson, K. (2002). What IQ tests test. Theory & Psychology, 12(3), 283-314.

[135] Williams, A. E. (2020, September). A model for artificial general intelligence. In *International Conference on Artificial General Intelligence* (pp. 357-369). Springer, Cham.

[136] Pei, J., Deng, L., Song, S., Zhao, M., Zhang, Y., Wu, S., ... & Shi, L. (2019). Towards artificial general intelligence with hybrid Tianjic chip architecture. *Nature*, *572*(7767), 106-111

[137] Sobieszek, A., & Price, T. (2022). Playing Games with Ais: The Limits of GPT-3 and Similar Large Language Models. *Minds and Machines*, *32*(2), 341-364.