Abstracts

PP1A.1
The purpose of this study is to examine measurement invariance in a self-efficacy scale used in international research. Differential item functioning (DIF) is routinely conducted for cognitive items, but DIF analyses are less frequently conducted for affective items that are included in cross-cultural studies. This research examines DIF related to gender and country using data from PISA 2012.

PP1A.2
The high costs associated with conducting face-to-face cut score studies have resulted in many standard setting practitioners exploring the feasibility of using virtual methods for such workshops. The aim of this study was to explore whether reliable and valid cut scores could be set in two synchronous virtual environments - audio and video. One way to investigate which virtual media (audio vs. video) is best suited to conduct a synchronous virtual standard setting workshop is to place all instruments and all judgments on the same linear scale through the Rasch measurement model. The focus of this presentation is on the framework used to evaluate cut scores set through the Many-Facet Rasch measurement (MFRM) model.

PP1A.3
Polytomous item explanatory item response theory (IRT) models have methodological advantages in explaining and predicting polytomous item difficulties, in extracting meaningful elementary components, and for validating hypothesized constructs in instrument development, by incorporating item properties that are designed, hypothesized, and observed in educational and psychological measurement. Using the polytomous item explanatory IRT models, we aim to investigate internal structure and test underlying rationale for the integrative inferential reasoning (IIR) comic assessment to support its development and validation. For data analysis, the item location explanatory many-facet Rasch model and the step difficulty explanatory linear partial credit model were applied to the IIR comic assessment data. The results show that the two polytomous item explanatory IRT models worked well to explain and predict the overall item difficulties or the step difficulties of the IIR comic items by the item properties that are used to design the assessment with underlying theoretical hypotheses.

PP1B.1
Osteoarthritis is a leading cause of pain and disability worldwide. Recently, there has been growing evidence regarding associations between osteoarthritis and oral-health conditions. Consequently, the assessment of oral-health-related quality-of-life (OHRQoL) in osteoarthritis patients is essential and could help identifying patients at risk of worsening their oral-health status. The Oral Health Impact Profile (OHIP)-14 is the most used instrument measuring OHRQoL in adults. Despite its popularity, to date OHIP-14 psychometric properties have been predominantly investigated based on classical test theory. Therefore, the aim of our study was to assess the psychometric properties of the OHIP-14 in osteoarthritis patients using the Rasch model. Data were obtained from a sample (N=360) drawn from a multi-center osteoarthritis registry in Austria. The OHIP-14 was found unidimensional. Person Separation Index was low (0.46) which also relates to the poor targeting of the scale, indicating that the scale was not sensitive enough to assess incipient decreasing OHRQoL.

PP1B.2
At the June, 2015 meeting of the Clinton Global Initiative America, the STEM Funders Network officially launched the STEM Learning Ecosystems Initiative. Though STEM typically focuses on Science, Technology, Engineering, and Mathematics, a wide range of educational programs across areas encompassing the arts, coding, social skills, literacy, and other fields are also included under the STEM heading. Educators, policymakers, funders, and other key stakeholders convinced of the value of cross-sector collaborations defined five pillars of ecosystem success: partnerships, measures, administrative systems, teaching and learning practices, and workforce development connections. The ecosystem success construct was mapped, existing items from an earlier ecosystem assessment were revised and augmented, four parallel forms with common items were devised, and an online administration system was created in early 2020. A data set of over 250 responses became available in early February, and will be reported in April. The measurement system design, calibrations, stakeholder comparisons, interpretation guidelines, and reports are reported in a partner presentation by the same authors.

PP1B.3
Proposed Caliper items were reviewed by a group of about 50 stakeholders at a meeting of the Measure STEM initiative at the October 2019 convening of the STEM learning ecosystems in Cleveland, Ohio. Per the terms described in the accompanying presentation proposal, the items were categorized as falling into one of the five ICICE levels and as addressing the conditions for success specified in one of the five pillars. To reduce the burden on those willing to take a role in the calibration pilot study, given 161 items, keeping the number of items per respondent to around 60, with about half of them in common, requires four forms. Linking the forms via 31 items in common on all of them leaves 130 items for distribution, and 32 or 33 items on each form. Preliminary data from about 135 respondents reported here will be augmented by a larger data set of over 250 respondents available as of this writing (early February). Results suggest broad conformity with theoretical, model fit, and parallel form expectations.

PP1C.1
Despite rapid increase in the enrollment rates experienced in Chile during the last two decades, there still are important differences in the proportion of students attending higher education by socioeconomic levels. We believe these differences may be in part related to the knowledge and perception high school students have about college. The research presented in this study explores the latent variables that are involved in the process of acquiring information and making decisions about college.This is done for the case of Chile using a sample of 1,600 high school students and information collected using an instrument designed specially for this project.

PP1C.2
The calculation of type-token ratios (TTR) is a common attempt to measure lexical proficiency, which is in turn, an important predictor of overall linguistic proficiency. There are at least five reasons, however, why Item Response Theory (IRT) estimation of ability levels may be a more valid measurement approach. This IOMW presentation introduces these reasons and shares a multistep process of measuring second language lexical ability that consists of lexical essay data pre-processing, dichotomous scoring of individual words, IRT modeling, and item fit analysis.

Relevant research investigates whether such an instrument measurement is, not only more valid, but also a more accurate and reliable predictor of human lexical proficiency ratings than TTRs.

PP1C.3
In this paper we demonstrate how Rasch Measurement Theory (RMT) can be used by program evaluators as a method for constructing measures that reflect a program's specific goals. Likewise, we show how the logic of a program theory can provide the necessary input for constructing meaningful measures, more closely tying program evaluation theory to measurement theory. As an example of this approach, we document the measure construction process for an evaluation of a university-based biology mentorship program. We discuss the path of moving from a logic model to a construct map (Wilson, 2005), and to use of the Rasch model for this evaluation. We further show how representation of the program theory through the logic model and the construct map can inject meaning into measurement results, which, in turn, can provide feedback to the program of interest. We hope to present this as a podium paper presentation.

SP01
Outcome phenomena are typically measured at the binary level: a comment is toxic or not, an image has sexual content or it doesn't, a patient is healthy or deceased. But the real world is more complex: most target variables are inherently continuous in nature. Physical quantities such as temperature and weight can be measured as interval variables where magnitudes are meaningful. How can we achieve that same interval measurement for arbitrary outcomes - creating continuous scales with magnitudes? We propose a method for measuring phenomena as continuous, interval variables by unifying deep learning with the Constructing Measures approach to Rasch item response theory (IRT). The crux of our method is decomposing the target construct into multiple constituent components measured as ordinal survey items, which are then transformed via an IRT non-linear activation into a continuous measure of unprecedented quality. In particular, we estimate first-order labeler bias and eliminate its influence on the final construct when creating a training dataset, which renders obsolete the notion of inter-rater reliability as a quality metric. To our knowledge this IRT bias adjustment has never before been implemented in machine learning but is critical for algorithmic fairness. We further estimate the response quality of each individual labeler, allowing responses from low-quality labelers to be removed. Our IRT scaling procedure fits naturally into multi-task, weight-sharing deep learning architectures in which our theorized components of the target outcome are used as supervised, ordinal latent variables for the neural networks' internal representation learning, improving sample efficiency and promoting generalizability. Built-in explainability is an inherent advantage of our method, because the final numeric prediction can be directly explained by the predictions on the constituent components. We demonstrate our method on a new dataset of 50,000 online comments labeled to measure a spectrum from hate speech to counterspeech, and sourced from YouTube, Twitter, and Reddit. We evaluate Universal Sentence Encoders, RoBERTa, XLNet, and ULMFiT as contextual representation models for the comment text, and benchmark our predictive accuracy against Google Jigsaw's Perspective API models.

PP2A.1
Because modern, simultaneously estimated longitudinal Rasch models are unable to handle many timepoints, new methods of producing person and item estimates and evaluating test function are

necessary. Longitudinal anchoring is a potential solution. By linking separate models together with a common scale of item parameters, trait levels can be estimated over many occasions. With proper anchoring procedures, person and item estimates can be obtained without limiting the number of timepoints that can be analyzed. A simulation study examining the performance of six longitudinal anchoring methods (Floated, Racked, Time One, Mean, Random, and Stacked) was conducted. The Mean and the Stacked anchoring methods best recovered the population change over time, person and item estimates, and model fit. Longitudinal anchoring shows promise as a low computation method of producing latent trait estimates in a Rasch-model framework.

PP2A.2
Early grade assessments are usually oral and administered on a one-on-one basis. Assessments in later grades—as students become better readers—are mostly written. In order to have student learning continuum from early to later grades, it is necessary to bring both modes of assessments into a common scale. A common way to vertically link assessments administered at multiple grade levels is by using common-item equating. However, an item presented to a student in two different modes—oral and written—has different item difficulty parameters. In this paper, we describe methods to link oral and written assessments into a common student learning outcome scale using Rasch model framework and dimensional alignment methodology.

PP2A.3
This paper presents two extensions of the Rasch model applicable to repeated measures in longitudinal studies. Adaptations of the Rasch model to experiments with repeated measures either ignore possible existing time effects or specify them in the model, usually with some latent distribution. We instead amend the dichotomous Rasch model and partial credit model by adding a subject-dependent time parameter. This reflects the assumption that the item parameters should be fixed to provide a stable frame of reference against which the change of the measured construct can be assessed for each individual. Adding time effects in this way naturally models the evolution of individuals, enables for parameter separation and
allows for rigorous item calibration. Sufficient statistics are also available for all parameters in the model. The theoretical implications of the model assumptions are further discussed and compared to current Rasch-related methods used in longitudinal studies.

PP2B.1
When developing psychological scales, researchers typically do not collect evidence that participants (1) understand items as intended and (2) respond to items for the expected reasons. Indeed, validation studies often lack formal documentation of the process through which items were generated and the expected manner in which they are to be interpreted by respondents. This paper introduces a mixed-methods approach to item-level validation called Response Process Evaluation (RPE). This large-scale, iterative method incorporates the population of interest into the creation and refinement of items, enabling researchers to present qualitative and quantitative evidence of validity based on item interpretations in service of the larger goal of validating a psychological instrument.

PP2B.2
In social measurement, constructs are measured that reside in the mind of patients, students, or consumers. To this end, items are administered forming a scale supposed to epitomize the

content of the latent variable representing the construct. Traditionally, scales are interpreted as measurement instruments. Recently, this view has been challenged arguing that man be the instrument. The respondent is indeed instrumental in the process of measurement. Since the self-perception of the object of measurement is performed by the subject, the perspective of man as the instrument does seem to be plausible. However, it is the items that enable the self-perception in the first place. It is therefore suggested that both man, as detectors, and the items, as lenses, conjointly constitute the measurement instrument. The thought-provoking perspective helps illustrate not only the importance of the unique characteristics of the Rasch model but also visualizes the requirements and properties of social measurement.

PP2B.3
Explanatory and predictive construct theories enable more fit-for-purpose, better targeted, and better administered measures. Construct specification equations (CSEs) provide the highest level of construct theory in social, psychological and health measurements, such as person-centered outcome measures (PCOMs). CSEs enable a deep understanding of how a collection of items works together by developing explicit theories explaining what is being measured. Such explanations offer the highly practical convenience of efficient automatic item generation, which greatly reduces measure development and quality assurance costs. Earlier work on CSE involves quantitative explanatory variables (such as number of digits), which are not necessarily relevant or applicable to all PCOMs. Therefore, new methods are needed to be able to fully understand and implement PCOM constructs. We demonstrate how a CSE can be obtained with qualitative explanatory variables for a measure of patient experiences of participating in care and rehabilitation. Results advance the understanding of construct theories and unit definitions in social, psychological, and health measurements.

SP02
What is educational measurement, and how does it relate – if at all – to concepts of measurement found in other disciplines? Despite the apparent fundamentality of this question for our field, it is not addressed at all in the most recent two editions of the authoritative volume Educational Measurement. This omission may have inadvertently reinforced the perception that it is not important or valuable for scholars interested in educational measurement to think about how their work fits in with more general scientific and philosophical frameworks, or even to be able to define their own central terms. As we write a new chapter on the nature of measurement for the forthcoming fifth edition of Educational Measurement, we hope to counter-argue that developing a more encompassing and interdisciplinary understanding of measurement is valuable not just philosophically, but has important practical ramifications for test design, analysis, interpretation, and use.

RS1A.1
Researchers generally employ rule-of-thumb critical values of the infit and outfit mean square error (MSE) statistics to detect rater effects. Unfortunately, prior studies have shown these values may not be appropriate. Parametric bootstrap method is an alternative approach to identify item and person misfit. However, its performance for detecting rater misfit has not been examined. We conducted a simulation study to assess its performance and we observed that the false positive rates of infit and outfit MSE statistics were highly inflated, and the true positive rates were relatively low. Thus, we proposed an iterative parametric bootstrap procedure to overcome

these limitations. The results indicated that using the iterative procedure to establish 95% CIs of infit and outfit MSE statistics had better-controlled false positive rates and higher true positive rates compared to using traditional parametric bootstrap procedure and rule-of-thumb critical values.

RS1A.2
To assess the predictive capacity of selection tests is a challenge because the response variable is observed only in selected individuals. In this paper we propose to evaluate the predictive capacity of selection tests through marginal effects under a partial identification approach. Identification bounds are defined for the marginal effects under monotonicity assumptions of the response variable. The performance of our method is assessed using a real data set from the university selection test applied in Chile and compared with the marginal effect of the traditional model used in Chile to evaluate the predictive capacity of the selection test.

RS1A.3
Three attitudes to identification problems can be recognized: the first, which we call illusory, is the one that claims that identification problems are not a problem for inferences, especially the predictive ones. This attitude underlies an important part of the Bayesian paradigm. But they only convey an illusion. A second attitude, which we call naive, is what sees in these problems a mere technical problem that does not allow obtaining “good” estimates of parameters of interest (e.g., consistent, unbiased) and that is necessary to solve by restricting the parametric space. This is a naive attitude because it fails to understand the significance of identification problems in the specification of a statistical model. A third attitude is one that realizes that identification problems correspond to the interaction between theory and empirical evidence: the evidence itself is not explanatory of its generation and therefore explanatory theories are required. Identification problems appear here. In order to understand the relevance of this last attitude, we want to show how identification problems emerged precisely as a barrier to an interpretation “without control” of empirical evidence. We will revisit above all a memorandum written by Ragnar Frisch in 1948.

RS1B.1
Clothing serves essential functions ranging from basic protection against the elements to symbolic expressions of social status and individual creativity. To create its products, the clothing, textile, and fashion industry taps into a wide range of resources, from agriculture (cotton, wool, leather, etc.) to petroleum (synthetics) to mining (metals for buckles, zippers, rivets, etc.) and water. The industry has recently begun to recognize that it is a major polluter, with new efforts focusing on enhanced long term sustainability. Data from the Kering Group's 2018 Environmental Profit and Loss (EP&L) statement were examined for their capacity to meet the demand for meaningful and manageable sustainability metrics. A huge effort and significant resources were invested in creating the data reported in this EP&L statement, as Kering's operations in 104 countries were evaluated in ways separable into almost 1,500 different indicators. The data system was not, however, designed as a measurement system. That is, it was not set up as specifically positing the possibility of estimating separable parameters for comparing company location performances across sustainability challenges. Of particular importance is the lack of information in the EP&L on the overall consistency of the data reported, on the uncertainties associated with the metrics given, and on the meaningfulness of

comparisons across environmental impacts, processes, and materials. The results reported here showing far from perfect data consistency and large uncertainties comprise an effort at constructing meaningful measures that offers important lessons for the redesign of the data and reporting system.

RS1B.2
This study measures three distinct constructs of superintendents' beliefs regarding school climate data (Importance, Capacity, and Trustworthiness). IRT analysis confirm the multidimensionality of the scale, and reveal the patterns of superintendents' responses across these three belief constructs providing a foundation for improving assessment practice through understanding and influencing leaders' beliefs. We use the Testing Standards (2014) to examine validity evidence for score interpretation and future uses of the survey with other stakeholders. The reliability evidence for the subscales is acceptable for research purposes but suggest the need for instrument modification.

RS1B.3
Psychometricians will often remove items from a test or assessment if the items show evidence of differential item functioning (DIF). However, Borsboom, Mellenbergh, and Van Heerden (2002) note that this may be unnecessary. They show that while items may show evidence of DIF, the notion of relative measurement invariance may still hold so that within a sub-group measurement results may be compared. However, measurement invariance procedures typically use only observed subgroups as a basis for analysis. Identifying meaningful subgroups thus remains a challenge. This paper uses item response data from a reading test to explore how comparable observable subgroups may be. Using a subset of data from a multidimensional reading measure, a latent class analysis is used to compare subgroup probabilities of belonging to particular classes and compares item response probabilities for different latent classes based on subsets of students. We hope to present this paper at a roundtable.

RS1C.1
As part of a NSF grant examining covariational reasoning in university College Algebra courses, this study examines a way to measure mathematical graphic reasoning. Within all mathematics courses there is a struggle between measuring correct answers verses correct reasoning. Using multiple Rasch models, we demonstrate differences in modeling Correctness (demonstrated in multiple choice problems) verses modeling Reasoning (demonstrated in written responses). We also gained useful insight into the items via Wright Maps when modeling the interaction between these. While this is a work in progress, we recommend continued coding of written responses for type of reasoning demonstrated using a 4 point coding framework for the scale.

RS1C.2
Student evaluations of teaching (SET) have been adopted worldwide as standard practice to enhance teaching and learning in higher education. In this study, we provide examination and exploration between conceptual perceptive and empirical data on SET, including a framework for measuring SET and empirical data supporting the framework. A coherent system for documenting SET measures was developed in order to aggregate horizontally, across units of divisions and faculties and simultaneously, vertically across hierarchical levels within and across disciplines.

RS1C.3
This study builds upon previous research investigating the degree to which students provide accurate Grade Point Averages, Class Ranks, and/or Test Scores (Cole & Gonyea, 2010; Frucot & Cook, 1994; and Kuncel, Crede, and Thomas, 2005) and extends previous studies by including demographic data such as grade level, class period, and gender. This study is important because the use of self-reported data are heavily used in research without knowing the extent of the accuracy of student-provided data. If students provide inaccurate responses about their accomplishments and/or characteristics and then those inaccurate responses are used, analyses and interpretations of results could be adversely affected. This paper presents a comparison between demographic data collected during a low-stakes science test from middle and high school students and data provided through official school records. Matched data from 731 students from a diverse school district on the Western US. Results indicate that student-reported demographic data do not necessarily correspond accurately with the official data source in many cases.

SP03
The speed-accuracy tradeoff suggests that responses generated under time constraints will be less accurate. Using a large corpus of 29 response time datasets, we probe its ability to explain behavior in non-experimental settings across a variety of tasks using idiosyncratic within-person variation in speed. We find inconsistent relationships between marginal increases in time spent responding and accuracy. However, we do observe time pressures (in the form of time limits) to consistently reduce accuracy and for more rapid responses to typically show the anticipated relationship (i.e., they are more accurate if they are slower). We next transition to analysis of items and individuals. We find substantial variation in the item-level associations between speed and accuracy, such variation could potentially be informative about the functioning of the measure. On the person side, respondents who exhibit more changes in response speed are also typically of lower ability. Finally, we consider the predictive power of a person's response time in predicting out-of-sample responses; it is generally a weak predictor. Collectively, our findings suggest the speed-accuracy tradeoff may be limited as a conceptual model in its application in non-experimental settings and, more generally, offer important empirical findings that will be useful as more response time data is collected.

SP04A
A robust body of evidence supports the finding that particular teaching and assessment strategies in the K-12 classroom can improve student achievement (Hattie, 2012). While experts have identified many effective teaching and learning practices in the assessment for learning literature, teachers' knowledge and use of "high leverage" formative assessment (FA) practices are difficult to model in novice teacher populations. By employing advances in construct modeling (Wilson, 2005), the theoretical underpinnings of learning progressions research, and four principles of evidence-centered design, teacher educators working with psychometricians can test hypotheses about student teacher learning progressions. Utilizing a FA moves-based framework (Authors), the paper examines how preservice teachers' posing, pausing, and probing practices in clinical placements can be measured using Rasch modeling techniques. Internal structure and rater reliability evidence are examined in light of intended uses for embedded assessment in credentialing programs.

SP04B

How can more meaningful, rigorous, practical, and convenient measurement be made more widely available in psychology and the social sciences? A perhaps unexpected source of possible answers can be found in the notion that measurement model identifiability might plausibly be rooted in the semiotics of everyday language. The first of two suggestive clues in this regard concerns the way the semiotic triangle of things, words, and ideas structures what Mach called the labor-saving economy of thought. The second clue is suggested by Mach when he notes the way science extends the economy of thought, bringing new things into words via rigorous conceptual determinations. Clear concepts represented in defined words shared throughout a community facilitate local conversations in which situated meanings are negotiated. Following through the implications of these clues points in new directions for psychology and the social sciences.

PP3A.1

Active learning classrooms are becoming more and more frequently used in higher education to increase engagement and inclusivity, thereby enhancing the learning experience, particularly in female, underrepresented minority, and first generation students (Baepler, et al., 2016). Flipped classrooms are also becoming more frequently used (Estes, Ingram, & Liu, 2014), using classroom time for hands-on learning activities and creative discussion has grown increasingly popular in higher educational settings.

Within quantitative psychology, courses such as measurement, statistics, and research methods are frequently fraught with undergraduate disengagement, fear, and loathing. Much research supports the administration of an integrated approach to teaching methods and statistics within the same semester course (e.g., Turiano, 2017, Barron & Apple, 2014). Additionally, integrating the lab component within the classroom setting, using open-source software such as RStudio, allows for accessibility, reproducibility, and inclusivity for students whose ease and comfort with quantitative topics and software applications has not been traditionally extensive. Additionally, flipping a quantitative course within an integrated, active learning classroom allows for even greater communication and synergy of statistics, measurement, and methods content within the learning environment. Delivery of statistical software scripts and learning exercises, all used for immediately applying quantitative knowledge in the active learning classroom, enables complex concepts to be grasped more readily.

Redesigning a complex quantitative curriculum to a flipped, active learning, integrated course is a gradual, lengthy process. Changes and techniques involved in the renovation in the quantitative content, open-source software for lab components, assessments, and exercises are described and their successes will be discussed. Greater overall class success, increased interest in advanced quantitative course enrollment, and a more relaxed classroom atmosphere have been evident, with promising indicators improving all the time.

PP3A.2

The presentation will describe at a high level, a process for developing, communicating, and maintaining, descriptions of growth in student achievement. The method relies on item mapping

as a starting point, but shows what else can be done to develop, communicate, and maintain effective and reliable explanations of growth in student achievement for test developers, teachers, parents, and other stakeholders. Higher levels of structure and rigor than practicable with item maps alone are needed to develop the depth of understanding and clarity of communication that test developers, teachers, and policy makers need to recognize and fill gaps in assessment instruments, decide the next step or topic in a sequence of instruction for students, and establish meaningful standards for achievement levels.

PP3A.3
Rasch measurement is an influential method in educational practice and psychometric modeling. Presently, there is an increasing trend of running Rasch models using R. The purpose of this study is to compare R packages' that offer Rasch analyses in terms of their statistical features, modeling availabilities, and operational efficiencies. We first provide a brief comparison of the critical features for these selected R packages, including descriptive information about the package and technical estimation details that affect their performance for complicated data structures. Then, we use simulated datasets to compare statistical outcomes from running three different Rasch models including the Dichotomous Rasch Model, the Rating Scale Model, and the Partial Credit Model. Lastly, we provide recommendations for Rasch users as well as Rasch Package developers.

PP3B.1
This study investigates the validity and reliability of a formative, digitally-administered mathematics assessment with items designed to align to two learning trajectories. To evaluate the validity of the assessment, we examine the extent to which empirical item difficulties align with learning trajectory levels. Because the assessment produces data visualizations intended to guide instruction, one conception of reliability in the context of this assessment is the extent to which the visualized location of students on the two learning trajectories is contingent upon the number and learning trajectory location of the items administered to each student. We estimate the impact of this assessment's length on these data visualizations by simulating variations in assessment length and item selection, then calculating two metrics: the most common learning trajectory location in each class and Loevinger's Ht, which indicates how well the responses fit the intended Guttman pattern that underlies both learning trajectories.

PP3B.2
To begin to meet the challenges to the field of measurement set forth by current US policy reforms like the Next Generation Science Standards (NGSS), researchers need to leverage the affordances of computer-based delivery and feedback effectively. This study builds on previous research and provides results from our investigation of how middle school and high school teachers interpret and use science assessment materials developed to be responsive to the NGSS. One of our goals was to develop an on-line system that provides timely, actionable, and relevant information about student performance to teachers. We employed an iterative design to create and test seven score reports for use by science teachers. This paper describes an Institute of Education Sciences (IES) funded study that we conducted to investigate how teachers construct meaning and use assessment results to support student learning and improve their practice.

PP3B.3

Visual Information Literacy lacks a measurement-related cognitive model, despite the large volume of research reported on it, and even though such models for textual literacy (e.g., the PISA test) and mathematical literacy have long circulated. We argue for the characterization of Visual Information Literacy in terms of developmental assessment. To this aim, after a short review of the state of the art, we describe the cognitive levels of Visual Information Literacy, by specifying the abilities that may be reached and the integration steps that may be at play in each of the levels. We finally outline the need to apply a robust assessment method also to Visual Information Literacy, likewise those envisaged in other fields. This assessment will be used to test the theory by constructing and calibrating an instrument, and guide future research.

RS2A.1
A novel measure was designed to examine knowledge of word spelling in a nuanced way. In the recall tasks, learners' spelling is scored based on its Levenshtein distance from the target word (i.e., the least number of single-letter changes needed to change a word into another via addition, deletion, or substitution) instead of in an all-or-nothing way. The recognition tasks ask learners to choose the word (e.g., "kindergarten") from five 1st-layer neighbors (e.g., "kindergarden"), five 2nd-layer neighbors (e.g., "kindegarden"), and five 3rd-layer neighbors (e.g., "kidegarden"), different from extant measures where learners choose from only two options (e.g., Conrad, 2008). This test has been administered to 59 English learners in China. Its reliability and validity are supported by item response data analysis. Across items, learners' mean locations consistently increased from the lowest step (spell or choose a 3rd-layer neighbor) to the highest step (correct word).

RS2A.2
Current dementia therapies focus mainly on early-stage disease. This necessitates fit-for-purpose measures to quantify early decline in memory. However, currently widely used memory tests are limited in providing accurate and sensitive measurement to discriminate between people with mild (or 'subjective') cognitive decline. Here, we describe research emerging from the European Union funded projects EMPIR NeuroMET 15HLT04/8HLT09, which has brought together clinical, academic, and national metrology institutes to improve measurements to better enable the early diagnosis of dementia. Specifically, we introduce the steps leading to the NeuroMET Memory Metric, which itself is the combination of legacy memory tests (i.e., blocks, digits, word recall), Rasch Measurement Theory and construct specification equations brought together to help 'tell a better measurement story' about memory and memory decline.

RS2B
Wolfe & Smith (2007a, b) provided the Rasch community with exceptional guidance relative to instrument development and documentation of validity evidence, combining the tenets of Rasch measurement with Messick's (1995) validity framework. Two recent developments have provided those in the measurement community with the opportunity to update this work and have discussions within the Rasch community about best practices for documenting validity evidence for instruments used within an educational context. Possibly most important to the measurement field and educational researchers as a whole, researchers and practitioners are left to make their connections as to how to integrate the work of Wolfe & Smith with the sources of evidence listed

in the updated Standards for Educational and Psychological Testing (2014). Secondly, within the last five years, the National Science Foundation has funded two projects to study the state of validity evidence for existing measures in science and mathematics education research fields and work towards the improvement of measurement in these areas: Exploration of Validity Evidence Gaps in Science Educational Achievement Testing (Gardner) and two grants for the Validity Evidence for Measurement in Mathematics Education project (VM2ED; Bostic, J., Carney, M., Krupa, E., and Shih, J.). This brings about an opportunity to educate a new generation of researchers about Rasch measurement tenets. Hence, the IOMW attendees seem uniquely positioned to contribute to a discussion regarding best practices for different types of measures (i.e., forced-choice knowledge assessments, Likert-scale assessments of motivation, classroom-observation protocols) based jointly on the 2014 Standards and Rasch measurement.