Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.



Volume 26 Number 12, June 2021 ISSN 1531-7714

# Evaluating Parent Comprehension of Measurement Error Information Presented in Score Reports

Priya Kannan, *Educational Testing Service*
Diego Zapata-Rivera, *Educational Testing Service*
Andrew D. Bryant, I*BM*

Individual-student score reports sometimes include information about precision of scores (i.e., measurement error). In this study, we specifically investigated if parents understand this information when presented. We conducted an online experimental study where 196 parents of middle school children, from various parts of the country, were randomly assigned to three conditions with different amounts of measurement error information. Parents in all conditions answered a series of comprehension questions about a student's performance on a hypothetical test. Results indicate that when information about error was presented, parents showed a significantly better understanding of score variability. Moreover, when asked about their preference for such information, parents across all three conditions indicated that they would like such information to be included in their child's report. Results from this study highlight the importance of clear communication of technical information to stakeholders, particularly parents, who are a diverse stakeholder group.

The correct interpretation and use of test performance results by various stakeholders is an essential part of the argument-based approach to validation (Kane, 2006; 2013). These test performance results in the form of scores and their intended meaning are communicated to stakeholders (including parents) through some form of a score report. Reporting scores meaningfully to stakeholders, including parents, so that they are accurately interpreted and appropriately used, is critical to the validity arguments supporting the assessment (Tannenbaum, 2019).

Parents are a uniquely heterogenous group of stakeholders who vary in their levels of education and language proficiency among other factors, which determines what and how much information they seek from their child's test score report. Regardless of their background, of course, the main goal for most parents is to be able to understand how their child performed on any given assessment and to help their child obtain any support needed relative to their performance. However, in order to be able to effectively help their child, parents must first understand the information communicated about their child's performance. But it is clear from previous research (Barber, Paris, Evans & Gadsden, 1992; Kannan, Zapata-Rivera, & Leibowitz, 2018) that very few parents understand all of the information presented in their child's score report. In particular, parents have been found to struggle with complex concepts such as measurement error (Kannan, Zapata-Rivera, & Leibowitz, 2018), which is a challenging concept for all stakeholders and sometimes misunderstood even by technical experts. This information about measurement error is important for parents to understand since, in practice, a number of high-stakes placement decisions (e.g., for additional academic support) that affect their child will be made based on standardized assessment scores.

In the context of assessments, measurement error corresponds to the difference between the test score and the student's underlying knowledge and skills. This error

in measurement may be introduced due to various factors such as the specific selection of items on the test or the specific conditions under which the test was administered. Because some degree of random measurement error is inevitable in all testing contexts, best practices set forth by the *American Educational Research Association, American Psychological Association, and National Council on Measurement in Education* Standards (AERA, APA & NCME Standards, 2014; specifically, standards 6.10 and 6.12) indicate that information about precision / reliability of scores (or measurement error) should be reported in terms appropriate to the audience.

The guidelines clearly recommend that information about measurement error be included in all score reports. Furthermore, it is also clear that such information about measurement would be very useful to parents as they consider high-stakes placement decisions that affect their child. Therefore, it is important to consider *how* such information can be presented in score reports intended for parents so that it leads to more appropriate interpretations and use for these score users. In this study, we evaluated parents' interpretations of measurement error information (error bars) provided in their child's score report. Specifically, we varied the details in the amount of explanatory information (i.e., footnotes) provided about measurement error across three conditions in a randomized online experiment and evaluated parents' understanding of score variability and precision through a series of comprehension questions.

## Literature Review

In the brief review below, we will first examine the types of information usually included in individual student score reports (ISRs) intended for parents. We will then synthesize some relevant literature that focuses on measurement error and how different non-technical audiences access and use that information to make decisions. Our goal throughout this review will be to consider how parents, as non-technical score report users, can understand information about measurement error when presented in their child's score report.

### Information presented in individual student score reports (ISRs) intended for parents

The overarching challenge in designing a score report that is capable of meaningfully conveying information is to make sure that the information is presented in ways that are appropriate to meeting the needs, pre-existing knowledge, and attitudes of the range of relevant stakeholders (Zapata-Rivera & Katz, 2014), in addition to accommodating the heterogeneity within each stakeholder group. The information presented in ISRs intended for parents has gradually evolved. Based on recommendations from numerous sources, ISRs designed for parents now tend to include various useful features such as comparisons to state and district average scores, sub-area performance, recommendations for next steps, and suggestions targeted at helping the student (e.g., AERA, APA & NCME Standards, 2014; Goodman & Hambleton, 2004; Hambleton & Zenisky, 2013; NEGP report, 1998; Ryan, 2006; Zenisky & Hambleton, 2012).

The inclusion of such features listed above has also been informed by research around the needs of diverse and underserved groups of parents (Kannan, Zapata-Rivera, & Leibowitz, 2018; Zapata-Rivera et al, 2014), and presented in a way that lends itself to easier interpretations and more appropriate inferences by parents. Particular attention is paid to scaffolding the technical language so that parents can understand the information presented more easily (see Kannan, Zapata-Rivera, & Leibowitz, 2018). Research has also focused on providing parents access to additional support resources such as sample questions at different levels of performance, suggested next steps for students performing at different levels, supplemental videos and guides to walk parents through the content provided in the reports, and links to school, district, and state requirements on websites (see Kannan, 2020; Zapata-Rivera, Vezzu, & Biggers, 2013; Zapata-Rivera et al, 2014).

However, one piece of information that has been extensively debated, based on its potential usefulness for non-technical audiences (especially parents), is the inclusion of information about measurement error in ISRs. In particular, the standards (AERA, APA & NCME Standards, 2014) and several researchers (e.g., Hattie, 2009; Zapata-Rivera, Zwick & Vezzu, 2016) have suggested that a description of the nature and precision of scale scores and what test results truly mean should be presented in an easily interpretable way in score reports for *all* stakeholders. On the other hand, some researchers (e.g., Rick, et al., 2016; Wainer, Hambleton, & Meara, 1999) have shown that parents tend to misinterpret or not value this information and have recommended that it is best to omit this information on

ISRs. It is, therefore, not clear from the research if the information about measurement error should be presented in ISRs for parents, and when presented, if it is either interpretable to or considered useful by parents.

In their survey of international score reports, Bradshaw and Wheater (2009), found that descriptive information (i.e., how and what results are presented) was easy to find on most reports. But, it was rare to find any information that was intended to explain reliability/error in the score reports they reviewed. Overall, these authors found little evidence in the literature that there have been any steps taken to explain or quantify error when reporting test results (to any stakeholder group). Moreover, there was no clear guidance in the literature as to how reliability of test score information should/could be reported in a meaningful way. Therefore, Bradshaw and Wheater suggest that there should be a balance between improving public understanding of results by explaining measurement error and improving public confidence in the system by not pointing out errors in a way that they are misunderstood.

In practice, when it comes to K-12 results reporting in the United States, there is quite a bit of variation across the states in the amount and nature of information about measurement error that is provided in ISRs. Several states do not provide information about measurement error in their score reports. For example, Faulkner-Bond et al. (2013) found that, of over 18 states and one consortium they reviewed, only two states provided information about measurement error on parent ISRs for their English Language Proficiency (ELP) assessments. More recently, Slater (2019) presented a review of ISRs for summative assessments from 47 states, and found that 27 of these states now reported measurement error information. While the practice of including error information in ISRs has increased, we have found that (Kannan, 2020) states either do not provide a clear explanatory text or provide a very succinct footnote that may not be comprehensible to parents.

## Interpretation of measurement error information by various audiences

Some previous studies have evaluated how well technical and non-technical audiences understand and make sense of information about measurement error. Ibrekk & Morgan (1987) found that graphical representations about measurement error can severely mislead some participants in estimating the mean, and that a self-reported "rusty" knowledge of statistics did not improve participant understanding. Belia, Fidler, Williams and Cummings (2005) explored researchers' understanding and use of standard error bars around two cell means, and demonstrated that presentation and interpretation of measurement error is challenging even for technical audiences such as researchers.

Correll and Gleicher (2014) focused on non-technical audiences and conducted a series of crowd-sourced experiments using the Amazon Mechanical Turk population. In their online study, they explored the interpretation of different visual representations (e.g., gradient plot, violin plot) of the margin of error around a mean including a bar chart with a 95% error bar – most commonly found in journal articles. Participants were presented with means (for one or two samples) and a postulated potential outcome and were asked to provide a rationale for the likelihood of said outcome when error around the mean(s) were presented. On a positive note, they found that even a non-technical (lay) audience is able to take into account information about measurement error and make nuanced inferences about potential outcomes. However, they found that participants were more likely to misinterpret error as contained within values represented by the bar for the 95% error bar representation than for other representations.

A few studies (e.g., Hopster-den Otter, et al., 2018; Zwick, Zapata-Rivera & Hegarty, 2014) have specifically focused on the presentation and interpretation of measurement error information to score report users, in particular, teachers. Zwick, et al. (2014) compared the effectiveness of four alternative graphical and verbal methods of representing measurement error in the comprehension of such information by teachers and university students. Although they did not find any statistically significant differences in comprehension across graphical representations, similar to Correll and Gleicher (2014), they found that participants who reported greater comfort with statistics preferred more informative displays that included variable-width error bars for scores.

Hopster-den Otter et al. (2018) also investigated teachers' understanding and preference of three alternate representations of measurement error (blur, color value, and error bar). They evaluated the extent to

which representations of measurement error in score reports influence teachers' decision making. Of the three representations they evaluated, they found that the error bar was the most preferred format among teachers. In addition, they also found that the position of a student's obtained score in relation to the cut score significantly impacted decisions, and that teachers significantly requested more information when the error bar straddled two levels.

Rick, et al. (2016) conducted focus groups with 11 middle-school parents to understand their needs for results from summative assessments and to evaluate their understanding of and preferences of various alternative representations created for each score report element. These researchers found that parents did not prefer representations that included error bars. Particularly, parents' comments revealed both dislike of the presentation (e.g., it looks like a 'star wars fighter') as well as an underlying misunderstanding of measurement error (e.g., "why don't you tell me where the good test is, and my child can sit there and take that one") as described in the hypothetical reports' footnote. Moreover, when asked to rank-order the importance of the various elements presented in the ISR, parents in this study ranked the 'error bars' as the least important. Therefore, these researchers caution against the presentation of error bars in ISRs developed for parents. However, it should be noted that evaluation of measurement error presentations was not the primary focus of this study, and parents were not presented with alternative scenarios where the error bar straddles two performance levels.

Overall, prior research suggests that both semi-technical and non-technical audiences have misconceptions about representations of measurement error. For example, one common misconception is that scores are perfectly precise and there is no need for error information. Conversely, another common misunderstanding is that test scores are imprecise because an error bar implies scoring errors. Educational stakeholders (particularly teachers) also demonstrate inconsistent interpretations of graphical representation on score reports when there is varying degrees of explanatory text included with the graphic (Zwick, Zapata-Rivera & Hegarty, 2014).

Parents and teachers both play a vital role in the decision-making process for students in K-12 contexts. While teachers may have a variety of goals in understanding student performance and supporting student needs (both at the individual and group levels), it should be noted that parents' main goal here is to be able to understand their own child's results and to ensure that their child is provided with the opportunities and/or support relative to their performance. It is, therefore, critical for score reports to support the decisions made by these important stakeholders (i.e., parents and teachers) and better enable them to draw correct inferences about what their students know and can do. Particularly for parents, score reports are the first, and perhaps only, point of interaction with the assessment, its purpose, and the decisions made as a result of their child's performance on this assessment.

Therefore, in order for parents to be able to engage in an informed communication with other stakeholders and actively participate in the decision-making process related to their child's academic performance and needs, it is critical that they are able to understand and use the information presented in their child's test score report. Studies that focus on parents as a stakeholder group or recipients of score reports are very limited. Moreover, there have been no known studies to investigate the extent to which parents understand different representations of measurement error, and if such information is even desired by parents – this was, therefore, the main motivation behind the current study.

## Current study

Score reports should be designed in a way that works well for the intended audience so that they can understand and use the score reports in a meaningful way. And, as already reiterated, the correct interpretation and use of test performance results by all stakeholders is integral to the validity arguments surrounding the test (Kane, 2006; 2013; Tannenbaum, 2019). Although ISRs designed for parents now tend to include various pieces of information, including measurement error, studies evaluating the interpretation and use of this information by parents are minimal. Results from our previous investigation (Kannan, Zapata-Rivera, & Leibowitz, 2018) with parents from diverse subgroups (disaggregated by education level and language proficiency) suggested that parents across all subgroups particularly struggled with the comprehension of information presented about measurement error in a hypothetical score report. Therefore, in this follow-up study, we used a between-subjects experimental design to evaluate how increasing the amount of explanatory

information provided around measurement error is helpful to all parents. However, since the focus of this study was not exclusively on parents from underserved groups, we did not explicitly recruit parents from various disaggregated and underserved subgroups in this study.

## Methods

### Research Questions

We evaluated the following two research questions in this study (each research question is identified with a brief word or phrase in parentheses to make it easier for readers to remember the focus of each question):

**RQ1 (comprehension):** Does providing more information about measurement error lead to *increased understanding* for parents?

**RQ2 (preference):** Do parents *prefer more or less information* about the measurement error around their child's score?

### Study Design

We used a 3 x 2 mixed design with 3 between-subjects conditions and 2 within-subject scenarios (see study design in Table 1); the two within-subject scenarios, however, were considerably different (as described below) with different comprehension questions, such that this was not a univariate or repeated measures design.

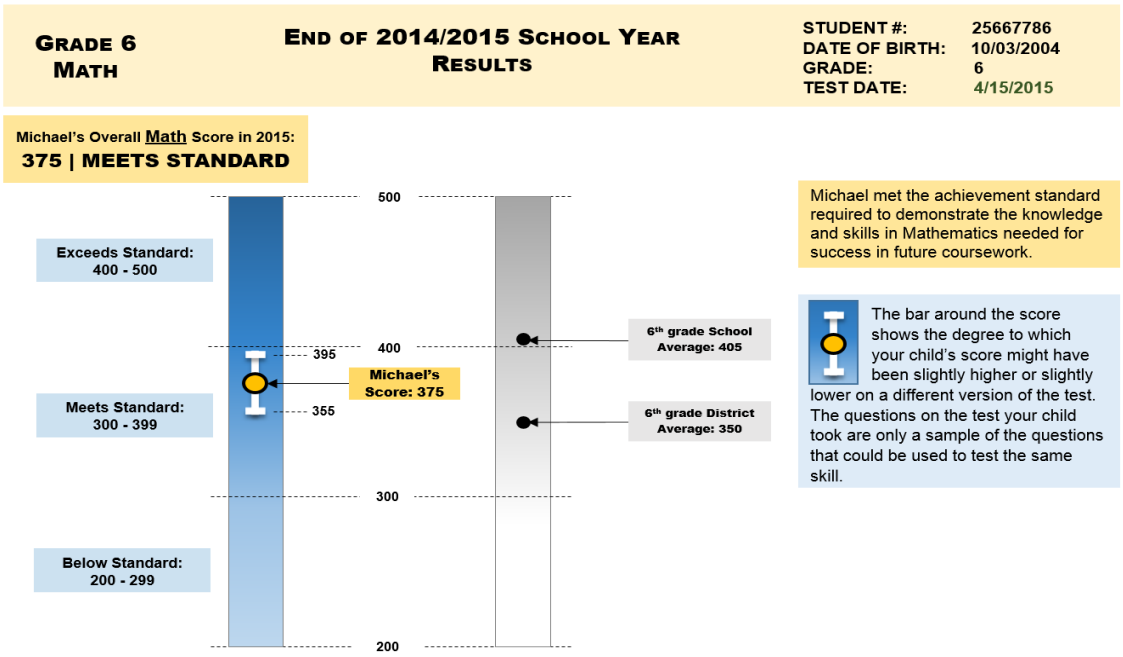
Previous research (e.g., Zwick, Zapata-Rivera & Hegarty, 2014; Correll & Gleicher, 2014) has found that some participants preferred more informative displays of measurement error. Therefore, in the three between-subjects conditions, we included different amounts of information about measurement error written for a non-technical audience. One hundred and ninety six parents of middle school children were randomly assigned to three conditions in an online experiment: (i) a condition where no error bar or explanatory information was presented; (ii) a condition where an error bar was presented with a brief (or standard) footnote (see Figure 1) which only includes information regarding variability due to the sampling of questions in different test forms; (iii) a condition where an error bar was presented with a more detailed footnote[1] (see Figure 2) that fully described the various factors, including testing occasion, that could affect a child's score on any given test administration, and spells out the actual range provided by a 95% confidence interval – we did not, however, describe how these error bars were computed (or the confidence level) to the participants in any of our study conditions.

**Table 1.** Research Design

| **Within-subject scenarios** | **Between-subject conditions** | | |
|---|---|---|---|
| | No error bar or footnote presented | Error with standard footnote | Error with detailed footnote |
| Scenario 1 (child meets standards) | *N = 69* | *N = 62* | *N = 65* |
| Scenario 2 (child is just below standards) | *N = 69* | *N = 62* | *N = 65* |

[1] Please note that the score report mockups used in this study were based on a hypothetical student's performance on a hypothetical assessment. The footnotes presented here do not represent operational score reports designed for any current standardized assessments. The intention of the various footnotes used in this study were to evaluate the extent to which parents understand explanations about the various sources of measurement error – these footnotes do not reflect actual computation of reliability statistics for any given assessment. The detailed footnote used in this study should not be used for operational score reports without consultation with respective program psychometricians. We recommend that practitioners should consult with their program psychometricians before constructing footnotes for operational testing programs, so that the footnote appropriately reflects the methods used to compute reliability estimates for their respective program.

**Figure 1.** Snapshot of the first within-subject scenario (where the child met standards and was clearly within that performance level) for the 'standard footnote' condition.




We designed two alternative (within-subject) scenarios of a hypothetical student performance (with minor variations across our three study conditions) that included select score report elements. Participants completed an online survey where they each reviewed two different scenarios of a hypothetical student's performance. In the first scenario, the hypothetical student met standards and was clearly within that performance level. In the second scenario, the hypothetical student placed just below standards, but the error bar (when presented) straddled two performance levels. Figure 1 shows a snapshot of the first within-subject scenario (where the child met standards and was clearly within that performance level) for the 'standard footnote' condition. Figure 2 shows a snapshot of the second within-subject scenario (where the child placed just below standards) for the 'detailed footnote' condition, but the error bar straddles two performance levels.

Each snapshot described a hypothetical student's observed score on a standardized end-of-year mathematics assessment displayed on a score range broken into three performance levels. Each snapshot also verbally described the student's performance level classification and provided some normative comparisons (i.e., school and district averages). We also provided some introductory text that described the scenario before participants reviewed the snapshots and answered corresponding comprehension questions.

## Participants and Procedures

The participants (i.e., parents) for our study came from two separate sources. Ninety-six of these parents were recruited using various traditional recruitment methods. An online post inviting parents of middle school students was posted in a company's internal website with locations across the country – this method accounted for about 40 participants in our sample. In addition, parents were recruited by contacting the national parent teacher organization who helped with posting flyers in several schools around the country – 56 of our participants came from this method of recruitment. In addition, we recruited another 100 participants for our study through crowdsourcing, using Amazon Mechanical Turk. We opened an initial HIT (Human Intelligence Task) where 1000 Turk workers were asked to complete a brief intake form. This intake form included a total of 18 questions, including questions about number of children and their grades – the actual purpose of the study and our selection criteria were not revealed in this HIT. This was done in order to mask the actual purpose of the study so that participants would not intentionally appear to be within our target population.

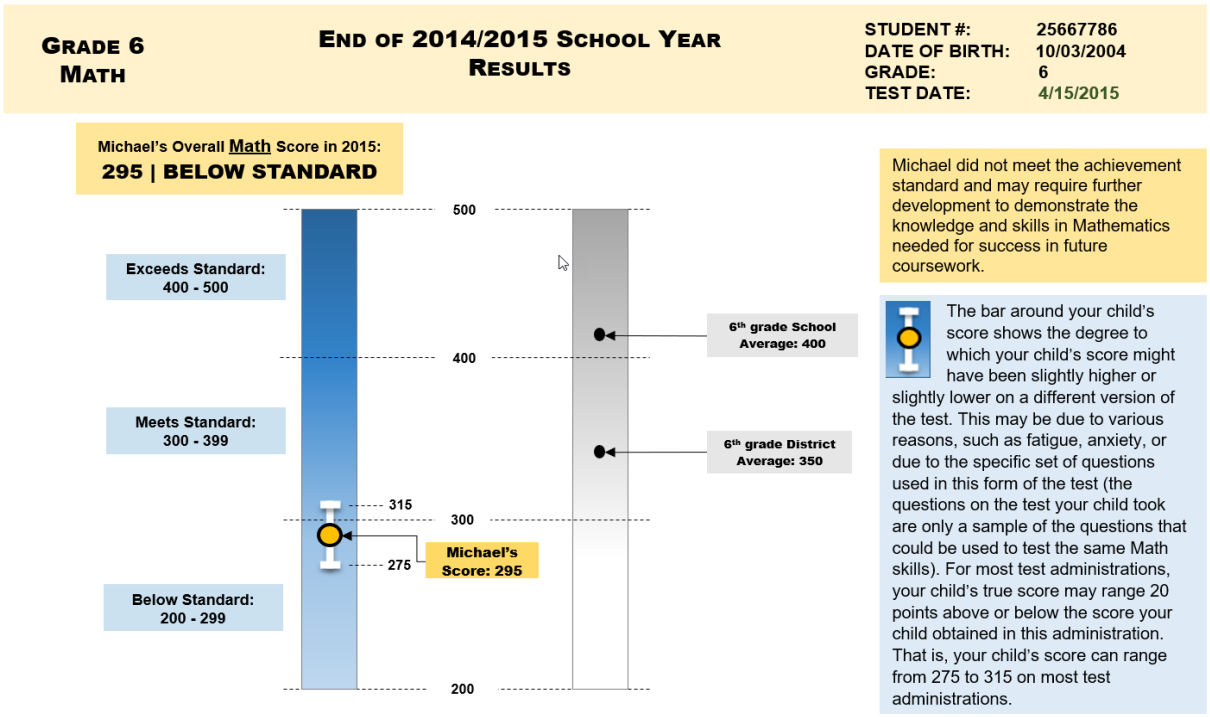
Based on participant responses to the 18 intake questions, we selected 100 parents of middle school children making sure that the selected sample represented a diversity of states, gender, and educational attainment. All 196 participants were compensated for their time. All data were collected between May and September 2016.

## Instruments

As described above, all participants completed a brief intake form prior to starting the survey. Participants were then provided with a brief overview and context for the hypothetical score report indicating that results are presented for a hypothetical student on a made-up standardized state assessment that is administered once at the end of the school year. Participants were asked to view the information presented in each hypothetical scenario (described above – see Figures 1 and 2), and answer a set of comprehension questions.[2]

For each hypothetical scenario, participants were first asked to describe the score report snapshot in their own words. This exercise was introduced so that they would become acquainted with the information presented in the snapshot. Following the initial presentation and description of the scenario snapshot, all participants then responded to selected-response comprehension questions; nine of these questions followed the first scenario, and eight followed the second scenario. Of the 16 selected-response questions, nine questions (i.e., 5 for the first scenario, and 4 for the second scenario) asked them to report factual pieces of information presented in the report snapshots (e.g. "What was Michael's score?"; "Which performance level was Michael classified into?").

**Figure 2.** Snapshot of the second within-subject scenario (where the child placed just below standards) for the '**detailed footnote**'[3] condition.




[2] All questions and instruments used in this study can be made available to the readers upon request.

[3] The simplification "...may range 20 points above and below…" is not completely accurate, and does not include the possibility of scores beyond the 95% confidence interval. However, in order to help parents from various educational backgrounds understand the concept of score variability, we choose to use this simplified language so that it could be understood by most parents.

For each scenario, we also included a few questions (mostly 'true/false' statements) that required participants to make some inference about measurement error (e.g. "No matter when Michael takes this test, he would always obtain a score of 295."); there were three such questions requiring an inference about score precision/variability for scenario one and four such questions for scenario two. Following each of these seven inferential questions, participants were asked to provide a brief justification (citing, if they chose, any piece of information from the report) for their choice of response. We subsequently coded these open-ended justifications for demonstrated understanding of measurement error, as described in the analyses section below.

Once participants answered the comprehension questions based on both scenarios, we showed all participants (across all three conditions) all three different representations for scenario two, and asked them to then respond to the following question: "Which of these three representations would you prefer to be included in your child's score report, and why?" Participants selected one of the three images as their choice, and provided a justification for their choice.

Finally, all participants completed an exit questionnaire that included questions about their age, gender, ethnicity, state and school district, education level, familiarity with statistical terms, educational exposure to statistics (i.e., if they have ever taken a statistics course, and at what level), and language fluency, among other things. Select variables from this exit survey are included in this paper to demonstrate the comparability of participants (based on demographics) across the three randomly assigned between-subjects conditions.

## Analyses

As described above, in this study we collected both quantitative and qualitative data. Participant responses to the closed ended comprehension questions were scored based on a key. In order to answer RQ1 (comprehension), we performed one-way between-subjects analysis of variance (ANOVA) to compare the comprehension of parents in the three study conditions. In order to answer RQ2 (preference), we used descriptive statistics to compare the proportions of parents across the three groups who preferred each image. Open-ended justifications provided for the inferential questions and preference of image were scored by two raters. The final scored categories were triangulated with quantitative results to further understand parent comprehension and preference. The specific details about how the scoring categories were developed and how interrater reliability was established for the open-ended responses is described in the rest of this section.

Two raters reviewed a sample training set of 20 randomly selected responses (per question) to develop a coding scheme for the open-ended prompts. After coding and reviewing 20 responses per question, we decided to code participant responses to the seven inferential questions into the same 3 categories (see Table 1). These 3 categories reflected differential levels of their displayed comprehension of measurement error concepts. Similarly, we used 20 randomly selected responses to come up with the coding categories for the preference question (see Table 8).

This coding scheme was then used to recode the 20 training responses and 50 additional responses per question which constituted the anchor set. In addition, each rater independently scored half of the remaining 126 responses for each comprehension question [i.e., 63 responses per rater]; in other words, each rater scored 133 (63, independent set + 50, anchor set + 20, training set) responses for each question. It should be noted that the responses were randomized during coding, and neither rater had any knowledge of the between-subjects group assignment, or of the recruitment method (Amazon Turk or traditional recruitment) of the participants while assigning their responses to these various coding categories. After the first round of independent coding, we compared the consistency with which the two raters had assigned responses to categories for the 70 responses (per question) that were double-coded. Interrater reliability (Cohen's Kappa) was computed for each question at the end of the first round of independent coding. The agreement at the end of the first round was low to moderate, with Kappa ranging from $\kappa = 0.41$ to $\kappa = 0.71$ across questions.

Based on the discrepant ratings at the end of the first round, the following criteria was used to flag responses for review. First, we flagged for review any responses where the two raters were more than 1 category apart in scores (e.g., where rater 1 had coded the responses in bucket '1', while rater two had coded the responses in bucket '3'). These discrepant responses

were flagged because they reflected very different understanding of the rubric between the two raters (of the 196 responses coded for each comprehension question, the number flagged for this type of discrepancy ranged from 2 to 6 per question). Second, we flagged discrepancies between the raters' assigned category which reflected some type of systematic deviation (e.g., where rater 1 had consistently coded responses in bucket '1' while rater 2 had coded the same responses in bucket '2'). These systematic deviations (sometimes more than one kind per question) constituted anywhere from 5 to 15 responses for each question.

As we reviewed these flagged discrepant responses, we had the opportunity to understand the reasoning behind the possible differential assignment of categories, and came to a consensus on some coding rules. Subsequently, we went back and recoded the training and anchor responses (all 70 common responses per question). In addition, we also used the revised coding rule to revisit the items coded independently by each rater [i.e., 63 responses per question] to evaluate if any codes needed to be changed for these items. The resulting interrater reliability from the second round of ratings ranged from $\varkappa = 0.86$ to $\varkappa = 1.00$.

**Table 2.** Scoring categories used for the open-ended justifications provided to the seven inferential questions with example responses at each score level.

| | **Category buckets** | | |
|---|---|---|---|
| **Category description** | **(1)** Incorrect understanding of score variability and / or misunderstood the question | **(2)** Reflects an understanding that test scores are not deterministic, but does not provide a clear explanation as to why that might be | **(3)** Reflects a clear understanding of score variability and measurement error |
| | **Example justifications provided for responses scored in each category** | | |
| **Example responses** | *"As long at the test was testing the same skills taught in school the test score should always be the same"* | *"The test makes students demonstrate a certain set of knowledge and skills; some variance may be expected but not that much."* | *"In the blue box on the right side it says: "… your child's true score may range 20 points above or below the score your child obtained…" That means 375-20= 355 - still meets standards; 375+20=395, still meets standards."* |
| | *"The score is an average over 2015. He could continue to take the test and either exceed or go below which will change the average score for the year. "* | *"Just because he did well this time though, doesn't mean he will always do well. I wouldn't think he would forget what he had already learned but if the way the test questions are asked changes, he may struggle more to meet standards."* | *"I believe students can perform differently based on the time of day, the day of week, the month of year, etc. I have personally experienced this myself when taking standardized testing. I often see my children come home with a "disappointing" grade even though they knew the subject. They say they were tired, uninterested, or other that I believe had an impact on their score that specific day."* |

One item had lower reliability (i.e., $\varkappa = 0.86$) when compared to the rest of the items (with kappa ranging from $\varkappa = 0.94$ to $\varkappa = 1.00$). Therefore, we briefly reviewed this one item and recoded this item a third time. The interrater reliability for this item improved to $\varkappa = 0.96$; overall, interrater agreement on the final set of independent coding ranged from $\varkappa = 0.94$ to $\varkappa = 1.00$. As a rule of thumb, Kappa values of 0.60 or higher are considered acceptable (Landis & Koch, 1977); therefore, it can be concluded that the interrater agreement among the two raters for the open-ended responses was more than satisfactory.

At the end of this process, there were 13 responses (across the questions, i.e., out of the 490 dual-coded responses) where the raters still disagreed. For the current paper, these responses have been coded as missing from the subsequent analyses, where we computed the frequency of participant responses in each coding category (i.e., results presented in Table 8 and Figure 4).

## Results

### Participant profile and characteristics

All 196 participants were parents of middle-school children, and had at least one child between grades 4 and 8 who attended a public school in the United States. Overall, participants in our study came from 40 states. Parents in the '*no footnote*' condition ($N = 69$) came from 26 different states, with about 32% of these parents recruited from New Jersey, and another 10% and 9% from Pennsylvania and Texas, respectively. Parents in the '*standard footnote*' condition ($N = 62$) came from 21 different states, with about 24% of these parents recruited from New Jersey, and another 11% and 10% from Texas and Alabama, respectively. Parents in the '*detailed footnote*' condition ($N = 65$) came from 24 different states, with about 20% of these parents recruited from New Jersey, and another 14% each from Pennsylvania and Texas.

The average age of parents across all three between-subjects conditions was about 40 years ('*no footnote*' condition: $M$=39.5, $SD$=6.8; '*standard footnote*' condition: $M$=39.9, $SD$=6.4; and '*detailed footnote*' condition: $M$=40.8, $SD$=5.9). Between 60% and 66% of the participants in all three conditions were female (or mothers); between 27% and 32% were male; the rest preferred not to report their gender. On average, participants across all three study conditions had about three children ('*no footnote*' condition: $M$=2.7, $SD$=1.0; '*standard footnote*' condition: $M$=2.6, $SD$=1.5; and '*detailed footnote*' condition: $M$=2.5, $SD$=1.1), with maximums ranging from 5 to 10 children across the conditions. The proportion of participants' who classified themselves into various ethnic groups is presented in Table 3. It can be seen from Table 3 that about 74% to 78% of the participants across all three conditions identified themselves as 'White (non-Hispanic)'.

**Table 3.** Participant ethnicity for parents across all three between-subject experimental condition.

| | Study condition | | | | | |
|---|---|---|---|---|---|---|
| | No Footnote | | Standard Footnote | | Detailed Footnote | |
| Ethnicity | N | % | N | % | N | % |
| Asian or Asian American | 7 | 10.1% | 3 | 4.8% | 5 | 7.7% |
| Black or African American | 2 | 2.9% | 6 | 9.7% | 5 | 7.7% |
| Hispanic or LatinX | 6 | 8.7% | 6 | 9.7% | 7 | 10.8% |
| White (Non-Hispanic) | 54 | 78.3% | 47 | 75.8% | 48 | 73.9% |
| Total N per condition | 69 | | 62 | | 65 | |

*Note:* % = The proportion of individuals of each ethnicity in each experimental condition.

**Table 4.** Educational attainment for parents across all three between-subjects experimental condition.

| | Study condition | | | | | |
|---|---|---|---|---|---|---|
| | No Footnote | | Standard Footnote | | Detailed Footnote | |
| Educational Attainment | N | % | N | % | N | % |
| Less than Bachelor's or no college | 27 | 39.1% | 26 | 41.9% | 24 | 36.9% |
| Bachelor's and additional credits | 24 | 34.8% | 19 | 30.7% | 24 | 36.9% |
| Master's Plus (including Doctoral) | 18 | 26.1% | 17 | 27.4% | 17 | 26.2% |
| Total N per condition | 69 | | 62 | | 65 | |

*Note:* % = The proportion of individuals in each experimental condition by level of educational attainment.

**Table 5.** Number and proportion of statistics courses taken by parents across all three between-subjects experimental conditions.

| | Study condition | | | | | |
|---|---|---|---|---|---|---|
| | No Footnote | | Standard Footnote | | Detailed Footnote | |
| Statistics Course Experience | N | % | N | % | N | % |
| High School-level | 9 | 24.3% | 6 | 17.1% | 15 | 35.7% |
| Undergraduate-level | 31 | 83.8% | 25 | 71.4% | 29 | 69.0% |
| Graduate-level | 8 | 21.6% | 11 | 31.4% | 12 | 28.6% |
| Minored in Statistics at Undergraduate Level | 1 | 2.7% | 0 | 0.0% | 0 | 0.0% |
| Majored in Statistics at Undergraduate Level | 0 | 0.0% | 1 | 2.9% | 0 | 0.0% |
| Parents (in each condition) who had taken at least one statistics course | 37 (54%) | | 36 (57%) | | 42 (65%) | |
| Parents (in each condition) who had not taken any courses in statistics | 32 (46%) | | 27 (44%) | | 23 (35%) | |

*Note:* % = The proportion of Statistics Course Experience among responses in each experimental condition. No Footnote (N=69); Standard Footnote (N=62); Detailed Footnote (N=65).

Table 4 presents participants' highest education level across the three conditions. It can be seen from Table 4 that the participants across the three conditions were more or less similar in their educational level; about 26% to 27% of participants reported that they have an advanced degree (Masters or higher) in all three study

conditions, while between 37% and 42% reported that they had less than a college degree. In addition, between 95% and 100% of the participants across all three study conditions reported that they were either native English speakers or had a very advanced level of English fluency.

In order to get an understanding of parents' level of familiarity with statistical terms and concepts, we asked them if they had ever taken a statistics course (see Table 5). Between 36 and 42 parents in each study condition reported having taken a statistics course at some level. We also asked them at what level (high-school, undergraduate or graduate level) they had taken these statistics course(s). Parents could choose multiple categories if they applied. It can be seen from Table 5 that, of those that had taken a statistics course, the majority of parents in all three study conditions reported having taken an undergraduate-level course in statistics.

In addition, we also asked parents to indicate their self-reported level of familiarity with a number of statistical terms and concepts they are likely to encounter while trying to read and understand their child's score report. These results are presented in Table 6. As would be expected, the results show that, across the three study conditions, parents typically reported higher levels of familiarity with the measures of central tendency (i.e., mean, median, and mode) and percentile rank, and relatively lower levels of familiarity with the concepts of standard error, quartile, sub-score, error band, and reliability.

The one exception to this general trend was noted where parents from the '*detailed footnote*' condition reported somewhat higher levels of familiarity with the concept of reliability. It is not clear if this is because these questions were part of an exit questionnaire that participants completed after reviewing the various study scenarios (since parents in this condition saw the most detailed explanation about measurement error), or if this was because there were a marginally higher proportion of parents from the '*detailed footnote*' condition who reported having taken at least one statistics course (see Table 5).

However, this one difference notwithstanding, the review of participant responses to these demographic variables clearly demonstrates that parents who were randomly assigned to the three study conditions were generally similar in their background and levels of understanding of information typically presented in score reports. Therefore, we can reasonably attribute the differences in comprehension observed between groups in this study to the experimental manipulation, and not due to any inherent differences in background or knowledge among parents who were assigned to each study condition.

**Table 6.** Average familiarity with statistical terms for parents across all three experimental conditions.

| | Study condition | | |
|---|---|---|---|
| Statistical terms | No Footnote | Standard Footnote | Detailed Footnote |
| Mean | 3.5 (0.6) | 3.5 (0.6) | 3.5 (0.8) |
| Median | 3.5 (0.7) | 3.6 (0.6) | 3.6 (0.7) |
| Mode | 3.3 (0.8) | 3.2 (0.8) | 3.3 (0.8) |
| Standard error | 2.8 (0.8) | 2.8 (0.9) | 2.8 (1.0) |
| Percentile rank | 3.5 (0.8) | 3.5 (0.7) | 3.6 (0.6) |
| Quartile | 2.8 (1.0) | 2.8 (0.9) | 2.7 (1.1) |
| Sub-score | 2.4 (1.0) | 2.5 (0.9) | 2.3 (0.9) |
| Error band | 2.3 (1.1) | 2.3 (0.9) | 2.2 (0.9) |
| Reliability | 2.9 (0.9) | 2.9 (0.8) | 3.2 (0.9) |

*Note:* Standard deviations are reported in parentheses; *Participants responded on 4-point scale with 1 = not at all familiar to 4 = very familiar.*

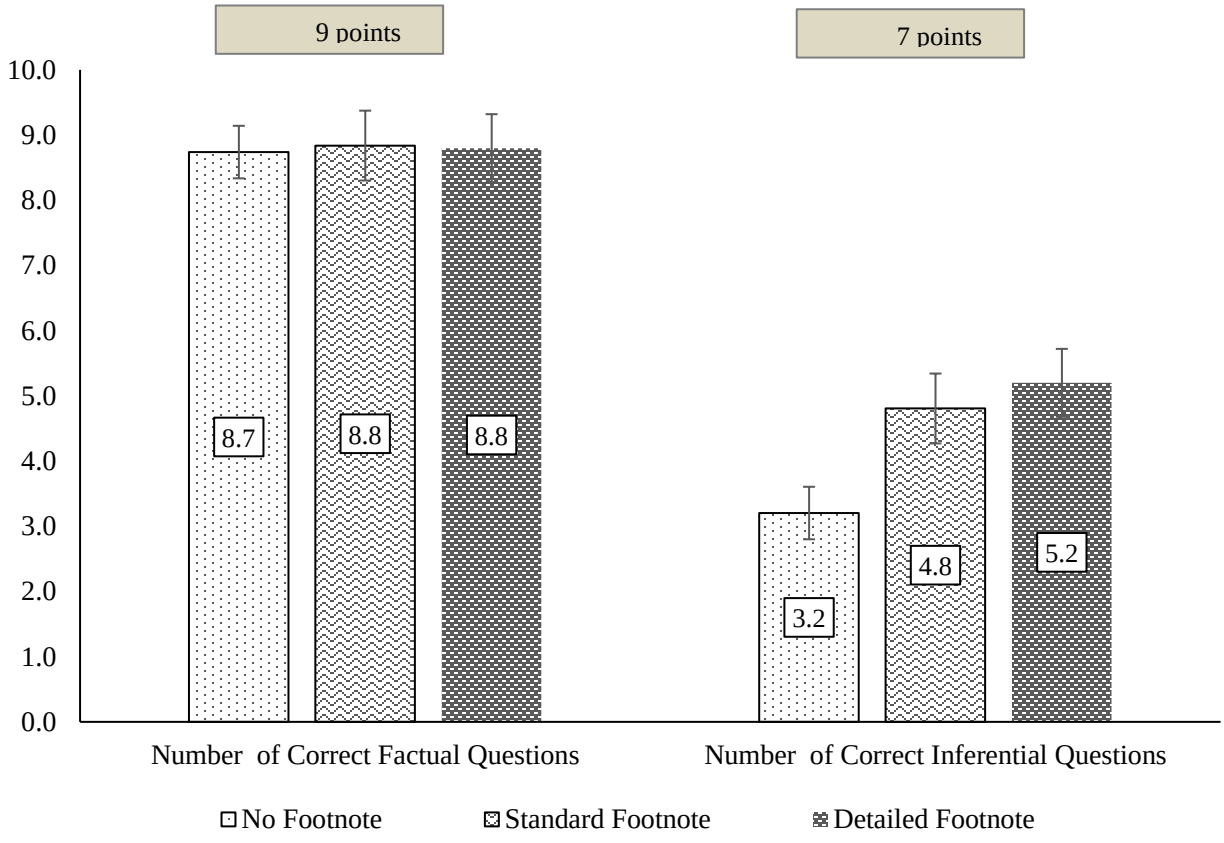
## Main Research Questions

Results as they pertain to the main research questions evaluated in this study are discussed in this section.

***RQ1 (comprehension): Does providing more information about measurement error lead to increased understanding for parents?*** In order to answer RQ1 (comprehension), we performed two one-way ANOVAs to compare the comprehension of parents across the three between-subjects study conditions on the following two dependent variables: (i) mean scores for the factual questions, and (ii) mean scores for the inferential questions. We anticipated that parents who were randomly assigned to the three study conditions, on average, should not significantly differ in their comprehension of factual pieces of information presented in reports. However, since we manipulated the amount of information we provided about measurement error, we anticipated that parents across the three conditions will significantly differ in their comprehension of this information. That is, parents who were randomly assigned to conditions where more information about measurement error was provided should have higher comprehension scores on the inferential questions (i.e., questions pertaining to comprehension of measurement error).

Although we used a 2 x 3 mixed factorial design, since the comprehension questions across the two scenarios were not identical it was not prudent to run a univariate two-way ANOVA on these results. We treated these dependent variables as unique and performed two univariate one-way analyses. We understand that treating these dependent variables in separate analyses increases the chance of Type-I errors. However, we found that the pattern of differences

**Figure 3.** Parents' responses to comprehension questions – showing mean answers correct within each experimental condition and 95% confidence interval.




***Note:*** No Footnote (N=69); Standard Footnote (N=62); Detailed Footnote (N=65); The error bars show the 95% confidence interval around the group means.

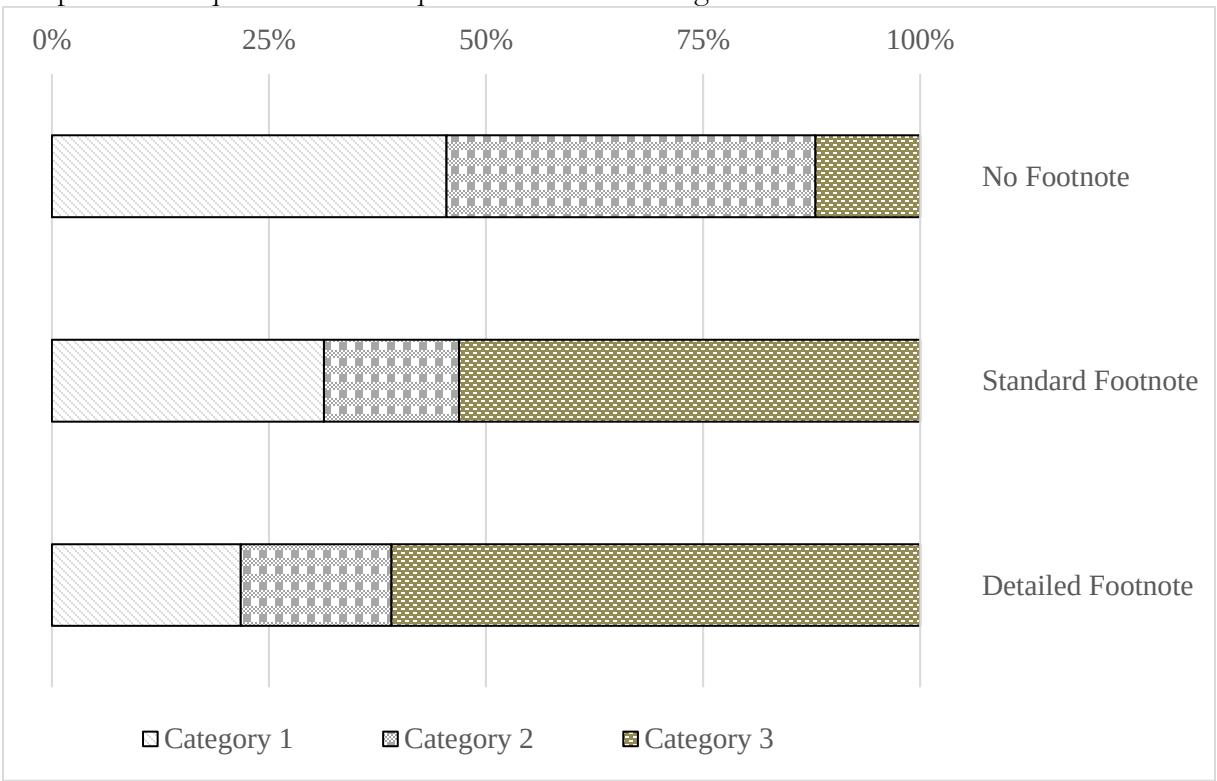
(including homogenous subsets) across the two scenarios were identical when we analyzed the data for each scenario individually. Therefore, in order to take another approach to minimize Type-I errors, we combined across scenarios and report results separately for the items types (i.e., factual vs. inferential): (i) mean scores for the factual questions (9 questions aggregated across the two scenarios), and (ii) mean scores for the inferential questions (7 questions aggregated across the two scenarios).

Results indicate that the three groups were not significantly different in their comprehension of the factual questions, $F$ (2, 193) = 0.6, $p$ = 0.5. Parents across all three conditions missed less than one item on average on these questions with comparable mean scores for all three groups on the 9 factual questions (see Figure 3). The mean scores for the three groups were as follows: "*no footnote*" ($M$ = 8.7, $SD$ = 0.6), "*standard footnote*" ($M$ = 8.8, $SD$ = 0.4), and "*detailed footnote*" ($M$ = 8.8, $SD$ = 0.5).

However, as expected, the mean comprehension scores for the inferential questions differed significantly across the three groups (see Figure 3), $F$ (2, 193) = 19.5, $p < 0.001$. In order to find the pattern of differences in comprehension across the three conditions, we performed a post-hoc Tukey HSD test. Results from the Tukey test showed that the "*no footnote*" condition demonstrated significantly lower comprehension ($M$ = 3.2, $SD$ = 1.7) than both the "*standard footnote*" ($M$ = 4.8, $SD$ = 2.1) and "*detailed footnote*" ($M$ = 5.2, $SD$ = 2.1) conditions. However, the "*standard footnote*" and "*detailed footnote*" conditions formed a homogenous subset and were not significantly different from each other.

Parents' open-ended justifications for the inferential questions help us further understand the pattern of results obtained for the quantitative data presented above. The coding categories used to code responses to all 7 inferential questions are presented in Table 2 along with some examples that illustrate representative responses that were coded within each category. Figure

**Figure 4.** Proportion of parents whose justifications were scored in the different coding categories across all seven comprehension questions that required an understanding of measurement error.




*Note.* **See Table 2 for an explanation of the 3 coding categories**; A total of 15 responses (out of 1372 (196*7) total responses) are excluded from this summary; of which the two raters disagreed on the categorization for 13 responses and 2 responses were left blank by the participants.

4 summarizes the proportions of parent responses across the 7 inferential comprehension questions that were coded into each of the 3 categories for parents from each between-subjects condition.

Results presented in Figure 4 show some very clear patterns in terms of parents' understanding of measurement error in each condition. It can be seen from Figure 4 that across the 7 questions, parent responses in the "*no footnote*" condition were more likely to be scored in category (1) which reflects a misunderstanding about measurement error (see Table 2 for a description of the coding categories), and may be attributed to a lack of clear communication in the score report.

Moreover, it is clear from Figure 4 that parent responses in the "*detailed footnote*" condition were most likely to be coded in category (3) which reflects a clear understanding about the concept of score variability and measurement error (see Table 2 for a description of the coding categories). Across the 7 questions, about 61% of the parent responses in the "*detailed footnote*" condition were coded in category (3); these proportions were about 53% for parents in the "*standard footnote*" condition and about 12% for parents in the "*no footnote*" condition. Moreover, only 22% of all parent responses in the "*detailed footnote*" condition were categorized as incorrect understanding when compared to about 45% of the responses for the parents in the "*no footnote*" condition that were coded in this category.

***RQ2 (preference): Do parents prefer more or less information about the measurement error around their child's score?*** In addition to parents' comprehension of the information presented, we were also interested in their preferences. Not only did we want to know (for all 196 parents in our study) if they cared about information about measurement error (or would rather not like to get such information), but we also wanted to know if they preferred more information describing the measurement error around their child's score rather than a standard statement. Therefore, as described previously, at the end of the scenario-based questions, we presented all three representations to all parents and asked them which one they preferred.

Parent preferences of the three types of images are presented in Table 7. It can be seen from Table 7 that, irrespective of the representation they saw during the rest of the study, between 58% and 80% of parents across all three study conditions preferred the image with the most information (i.e., the "*detailed footnote*" representation).

Participants were also asked to provide a brief justification (see Table 8) for their preference of the image they chose. It can be seen from Table 8 that parents who chose the "*detailed footnote*" representation either specifically mention the additional information about factors that can affect their child's performance as provided in this image (41 of the 131 parents who chose this image), or they indicated that they just preferred more information (80 of the 131 parents who chose this image). As one parent put it: "*I like the most detail. It helps me to interpret the meaning of the score and acknowledge there could be variability in the score received each time the test is taken.*"

As can be seen from Table 7, 25% of the parents in the "*no footnote*" condition and 29% of the parents in the "*standard footnote*" condition preferred the "*standard footnote*" image. From a review of their justifications, we found that there is almost a "Goldilocks" principle to their justification – they thought the information presented in this image was "just right" – not too much, but still enough to understand that their child's score is not unchangeable and that on any given test administration their child's score could vary slightly based on the test form used. Finally, from a review of the justifications provided by parents who preferred the "*no footnote*" representation (i.e., 13% to 16% across conditions in Table 7), we found that these parents just did not want the additional information (see Table 8) and several of these parents indicated that they do not want to know about the "*what-ifs*", especially when their child cannot possibly retake this test.

## Discussion

Parents are one of the most important stakeholder groups in the K-12 assessment context. With the increasing focus on "valid and reliable assessments" of "challenging academic content standards" (ESSA, Pub.L. 114-95, Sec. 1111), there has been a systemic change in K-12 curriculum, standards, and assessments in the last couple of decades. With these systemic changes, the following statement from the NEGP (1998) report rings more true today than ever before: 'If parents are well informed and made a part of the improvement efforts from the beginning, they are more likely to be the catalyst needed for change – they are more likely to support their school's goals and demand the

**Table 7.** Parents' preference for the amount of measurement error information to include in their child's score report across all three between-subjects experimental condition

| | Study condition | | | | | | Total per image |
|---|---|---|---|---|---|---|---|
| | No Footnote | | Standard Footnote | | Detailed Footnote | | |
| Image preferred | N | % | N | % | N | % | |
| Image with no error bar or footnote | 9 | 13.0% | 8 | 12.9% | 10 | 15.9% | 27 |
| Image with error bar and a standard footnote | 17 | 24.6% | 18 | 29.0% | 3 | 4.6% | 38 |
| Image with error bar and a detailed footnote | 43 | 62.3% | 36 | 58.1% | 52 | 80.0% | 131 |
| Total per condition | 69 | | 62 | | 65 | | 196 |

*Note:* % = The proportion of parents who preferred each image in each experimental condition; No Footnote (N=69); Standard Footnote (N=62); Detailed Footnote (N=65).

**Table 8.** Justifications provided by parents for their preference of one of three images.

| Justifications provided | Image with "no footnote" | Image with "standard footnote" | Image with "detailed footnote" | Total per justification |
|---|---|---|---|---|
| Specifically mentions additional information about factors that can affect their child's performance on one assessment *(Provided by parents who chose the image with "detailed footnote")* | | | 41 | 41 |
| Just mentions that C provides the most amount of information / the most thorough explanation *(Provided by parents who chose the image with "detailed footnote")* | | | 80 | 80 |
| Mentions that range was provided, but does not clearly explain the differences between the three images or the reason for their choice *(Provided by parents who chose the image(s) with "standard footnote" and "detailed footnote")* | | 13 | 10 | 23 |
| Enough information, but not too much – Goldilocks *(Provided by parents who chose the image with "standard footnote")* | | 25 | | 25 |
| Simple and easy to read (other two have unwanted or otherwise confusing information) *(Provided by parents who chose the image with "no footnote")* | 27 | | | 27 |
| Total per image | 27 | 38 | 131 | 196 |

instructional changes necessary to meet these goals' (NEGP, 1998, p.2).

However, parents, in general, do not understand how the policy-based improvement efforts for standards affect their children and schools. Moreover, it should be reiterated that parents are mostly interested in understanding their own child's performance; the policy implications, at large, are not of particular concern to most parents. Score reports are one of the most important, if not the single point of interaction where parents learn about the assessment, its purpose, their child's performance on this assessment, the impact of this performance on their child's academic development, and potentially their future success in college and careers. Particularly when a number of high-stakes placement decisions are made based on standardized assessment scores in practice, a systematic effort should be made to provide relevant information (vis-à-vis measurement error) to parents via score reports that are accurate, yet comprehensible to the average parent population.

Yet, although a lot of attention has been directed to the creation of technically sound assessments, there has been considerably little focus on how results from these assessments are reported to various stakeholders, particularly parents. Our previous study (Kannan, Zapata-Rivera, & Leibowitz, 2018) evaluated the extent to which parents from diverse subgroups (disaggregated by education level and language proficiency) understand and interpret the information presented in a hypothetical student score report. Our results from that study suggested that parents, across all subgroups, struggled with the comprehension of information presented about measurement error. Therefore, in this follow-up study, we wanted to use a between-subjects experimental design to evaluate parents' comprehension of measurement error information. We anticipate that the results from this study will contribute to the growing body of literature on score report design and development (e.g., Hambleton & Zenisky, 2013; Zapata-Rivera, 2011) that is critical to the underlying validity arguments around a test score's interpretation and use. By determining how parents understand and make sense of the measurement error around a hypothetical student's performance, we hope to inform the development of audience-specific score reports that are designed to facilitate parents' comprehension and usability of said reports.

For the general population of parents in this study [note that the sample in this study is predominantly 'White (Non-Hispanic)'], our results suggest that across all the study conditions parents were able to more or less accurately glean the factual information presented in their child's score report. In addition, when presented with information about measurement error, parents make a concerted effort to understand this information, and presenting more information had a positive effect on their understanding of measurement error. In addition, across all study conditions, this group of parents tended to prefer more information about measurement error and preferred the image that provided the most information (i.e., the "*detailed footnote*" representation).

Finally, though parents in the "*no footnote*" condition were more likely to misinterpret information about measurement error (since appropriate information was not presented in the score report to support accurate interpretations), it should also be noted that some parents in the "*no footnote*" condition (in varying frequencies across the 7 questions) presented clear enough justifications that warranted assigning them into the score category (3). We interpret these results in a positive light to show that parents are not only trying to understand the information presented about measurement error (as evidenced by the higher comprehension scores for the "*standard footnote*" and the "*detailed footnote*" conditions), but may also make an effort to parse the information by themselves to try and understand their child's performance on standardized assessments.

## Practical implications

Overall, from these results, we conclude that it may be important to consider providing detailed and clearly comprehensible information about measurement error in individual student score reports (ISRs) intended for parents. In addition, caveats about appropriate use should also be included in the score report so that parents understand how to use the results in the right manner. In addition, concerted efforts should be taken by schools and districts to communicate the purpose of standardized testing to parents. Assessment literacy support materials designed for parents and the general public should not only include information about the test's content, format, and purpose, but also be

specifically designed to provide parents with a layman's overview of concepts such as scale scores, percentiles, cut scores, performance levels, measurement error, etc., to help facilitate accurate interpretations of their child's performance.

As one example, five years ago the National Council on Measurement in Education (NCME) kickstarted an initiative to develop modules on 'Assessment Literacy' for the consumption by the general public (Weiss, et al., 2016). A number of these modules have now been made available on YouTube for the general public (e.g., '*What is Educational Measurement: Anatomy of Measurement*' at present publicly available online at https://www.youtube.com/watch?v=A7XWAsPwbqY). Several states and assessment programs have also taken the initiative to develop their own assessment literacy materials for parents (see Kannan, 2020). It would be useful to include links to access such resources directly in parent ISRs. Inclusion of such resources would be valuable in helping parents better understand the information presented in their child's score report, and also empower them with the right kind of knowledge in discussing their child's performance with their teachers.

## Limitations and future directions

Despite the large sample size used in this study, we have to point out that we have barely grazed the topic of parent *interpretation* and *use* of information provided in score reports. There are a number of limitations to this study that we should acknowledge, and we would also like to use this opportunity to provide some directions for follow-up investigations that may evaluate parents as consumers of score reports more holistically. In this study, we only focused on parent interpretation of one piece of information (i.e., measurement error) provided in their child's K-12 score report. Future studies should aim to evaluate parent comprehension and preference of various score report elements (e.g., sub-scores, growth / performance over time) to better understand the needs and pre-existing knowledge of parents thereby informing the design of score reports that are catered to this stakeholder group. In addition, rather than use hypothetical examples as used in this study, future studies should also try to use operational score reports that include footnotes constructed for operational use by program psychometricians to specifically evaluate parent interpretation and use of this information.

Moreover, since the purpose of this study was to evaluate if parents are able to understand information about measurement error, we decided to include a holistic definition of measurement error in the '*detailed footnote*' condition that describes factors underlying both test-retest and alternate-form reliability. However, measurement error due to testing condition was not highlighted in this study, and not included in the '*standard footnote*' condition, since it is not practical to test students across multiple occasions in operational large-scale summative testing conditions. Nevertheless, to evaluate parents' comprehension of a complete representation of various sources of error, future studies should try to explicitly evaluate parents' understanding of error from various sources in addition to evaluating the extent to which parents understand 'measurement error' as computed for a specific operational assessment.

In addition, future studies may also investigate variations in the terminology and how measurement error is represented in the footnote for parents. As we have already pointed out, using the term "error" may be distracting for some parents, and it is possible that the use of this terminology leads parents to misinterpret measurement error as representing scoring errors. Therefore, alternative language where these bars are presented as precision bars, rather than error bars, and footnote language around score precision rather than measurement error may be used to evaluate if alternative language helps parents better understand concepts around score precision and measurement error.

Most importantly, we would like to reiterate that parents are a particularly heterogeneous stakeholder group who come from various ethnic backgrounds with varying educational, socio-economic, and language proficiency levels (Kannan, Zapata-Rivera, & Leibowitz, 2018). However, since we were interested in the comprehension of measurement error by parents as a broader stakeholder group, we did not explicitly recruit to ensure diversity and representation of underserved groups in this study. As can be seen from Table 3, the group of parents who participated in this study, though drawn from several states across the country, were fairly non-diverse in background and mostly White (Non-Hispanic). Therefore, the findings from this study should be interpreted with caution and may not directly generalize to other parent subgroups, and particularly to parents from underserved groups. For example, our previous study (see Kannan, Zapata-Rivera, &

Leibowitz, 2018) shows how parents with certain background characteristics (i.e., those who had less than a college degree and were non-native speakers of English) struggled with the comprehension of all types of information presented in the hypothetical score report which also includes factual pieces of information presented in score reports. Therefore, we highly recommend that future studies should focus on the interpretation and use of score reports and support resources by parents from various diverse subgroups, with varied educational and linguistic backgrounds, paying particular attention to parents from underserved groups.

Previous research with teachers (Hopster-den Otter et al., 2018) has shown that, when measurement error information is presented, the location of a student's score in relation to a cut score had significant implications for stakeholder interpretation and the subsequent decision-making. Although we evaluated a scenario in this study where the hypothetical student was placed just below standards, with the error bar overlapping two performance levels, we did not systematically evaluate differences in the degree to which parents' value information about measurement error in the two scenarios in our study. This is because the focus of our study was not on the use and decision-making from scores. Nevertheless, it would be valuable to learn about how parents intend to use information about measurement error in their subsequent conversations with teachers under various scenarios, so that caveats about appropriate use may also be included in the score report. Future studies should not only evaluate the degree to which parents understand information about measurement error, but also the degree to which they would value such information under various alternative scenarios.

Finally, though designed as an experimental study, this study was conducted as an online survey study. Therefore, it was impossible for us to directly interact with parents who did not understand the concept of score variability to seek further clarifications about the nature and context of their misunderstandings. We also had to limit the total number of questions (factual and inferential) we could ask parents in this online setting, which may have resulted in a ceiling effect in comprehension scores, and a lack of power in detecting differences between study conditions. Future cognitive laboratory studies should specifically focus on improving the presentation of information in score reports by trying to elicit responses from parents in a face-to-face setting or virtual cog lab setting. These studies should focus on evaluating specific alternative representations and alternative terminology with a variety of comprehension questions to evaluate parent understanding of the information presented in ISRs. In addition, score reports and support resources from state operational testing programs should be used to help parents better understand the information presented in their own child's ISR. Moreover, it would be crucial for these studies to include parents from underserved subgroups to inform the design of score reports and other supplementary information for all parents.

## Conclusions

In order to be useful, ISRs should be designed in a manner such that they are able to serve as the mediators of communication between parents and teachers. Parents must be able to participate or engage with their child as well as communicate with school personnel about target key areas for student growth. Therefore, it is imperative that the information provided in ISRs are clear and interpretable to parents. However, in light of the continuing controversy about presenting information about measurement error in ISRs for parents, where some authors (e.g., Zapata-Rivera, Zwick & Vezzu, 2016), who have studied the use of short video tutorials on measurement error for teachers, suggest providing similar information to other stakeholders, while others (e.g., Rick, et al., 2016; Wainer, Hambleton & Meara, 1999) explicitly caution against providing measurement error information to parents, we hope that the results from this study provides some guidance to developers of ISRs intended for parents.

Our results indicate that carefully crafted information about measurement error can be included in ISRs designed for parents with positive results. The value of adding more explanatory information than typically contained in a standard footnote (included in some state reports) is somewhat more uncertain based on these results. Though the results from this study did not point to any significant differences in comprehension based on the standard and detailed footnote representations, parents' responses to the preference question shows that parents appreciate and desire more information when it comes to understanding their child's performance on tests, and that they are willing to make a concerted effort to

interpret this information. Though these results should definitely be evaluated with parents from underserved subgroups in future studies, the results from our study at least provide preliminary insights into parents' comprehension of and preferences for such information. We hope that these results, and any additional results from follow-up investigations, prove useful to states and testing programs in designing ISRs for parents.

## References

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

Barber, B. L., Paris, S. G., Evans, M., & Gadsden, V. L. (1992). Policies for reporting test results to parents. *Educational Measurement: Issues and Practice, 11*(1), 15–20.

Belia, S., Fidler., Williams, J., and Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods, 10*(4), 389–396.

Bradshaw, J., & Wheater, R. (2009). International survey of results reporting. Coventry, UK: Office of Qualifications and Examinations Regulation.

Correll, M., Gleicher, M. (2014). Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error. *IEEE Transactions on Visualization and Computer Graphics, 20(6),* 2142–2151.

Every Student Succeeds Act of 2015 (ESSA), P.L. 114–95, 20 U.S.C. § 114 stat. 1177 (2015–2016).

Faulkner-Bond, M., Shin, M., Wang, X., & Zenisky, A.L. (2013, April). *Score Reports for English Proficiency Assessments: Current Practices and Future Directions*. Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), San Francisco, CA.

Goodman, D. P. & Hambleton, R. K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17*(2), 145–220..

Hambleton, R. & Zenisky, A. (2013). Reporting test scores in more meaningful ways: A research-based approach to score report design. *APA handbook of testing and assessment in psychology*. Washington, DC: APA.

Hattie, J. (2009, April). *Visibly learning from reports: The validity of score reports*. Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), San Diego, CA.

Hopster-den Otter, D., Muilenburg, S. N., Wools, S., Veldkamp, B. P., & Eggen, T. J. H. M. (2018). Comparing the influence of various measurement error presentations in test score reports on educational decision-making. *Assessment in Education: Principles, Policy & Practice, 26*(2), 123–142.

Ibrekk, H. & Morgan, M.G. (1987). Graphical communication of uncertain quantities to nontechnical people. *Risk Analysis, 7*(4), 519–529.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4 ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kane, M. (2013). Validating the interpretation and use of test scores. *Journal of Educational Measurement, 50*(1), 1–73.

Kannan, P., Zapata-Rivera, D. & Leibowitz, E. A. (2018). The Interpretation of Score Reports by Diverse Subgroups of Parents. *Educational Assessment, 23*(3), 173–194. https://doi.org/10.1080/10627197.2018.1477584

Kannan, P., (2020). *Supporting the interpretation of summative assessment results for parents from underserved groups*. Final report submitted to an ETS testing program. Princeton, NJ: Educational Testing Service.

Landis, J. R., Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics 33*(1), 159–174.

National Education Goals Panel, NEGP (1998). *Talking about tests: An idea book for state leaders*. Washington, DC: U.S. Government Printing Office. Retrieved from http://govinfo.library.unt.edu/negp/reports/98talking.PDF.

Rick, F., Slater, S., Kannan, P., Sireci, S., Zenisky, A., & Dickey, J. (2016). *Parent's perspectives on summative test score reports* (Center for Educational Assessment Research Report No. 937). Amherst, MA: Center for Educational Assessment, University of Massachusetts.

Ryan, J. M. (2006). *Practices, issues, and trends in student test score reporting*. In S. M. Downing, & T. M. Haladyna (Eds.), Handbook of test development. (pp. 677–710). Mahwah, NJ: Lawrence Erlbaum Associates.

Slater, S. (2019, April). *How states are communicating student growth to parents*. Paper presented at the meeting of the National Council on Measurement in Education, Toronto, Canada.

Tannenbaum, R.J. (2019), Validity Aspects of Score Reporting (pp. 9–18). In D. Zapata-Rivera (Ed.), Score reporting Research and Applications. Routledge, NY.

Wainer, H., Hambleton, R. K., & Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement, 36*(4), 301–335.

Weiss, L., Walker, C. M., Rorick, B., Vasquez-Colina, M. D., & Morris, J. (2016). *Multiple perspectives for promoting assessment literacy for parents.* Coordinated session at the annual meeting of the National Council on Measurement in Education (NCME), Washington, DC.

Zapata-Rivera, D. (2011). Designing and evaluating score reports for particular audiences. In D. Zapata-Rivera & R. Zwick (Eds.), *Test score reporting: Perspectives from the ETS score reporting conference* (Research Report No. RR–11–45, pp. 32–61). Princeton, NJ: Educational Testing Service.

Zapata-Rivera, D., & Katz, I. (2014). Keeping your audience in mind: Applying audience analysis to the design of score reports. *Assessment in Education: Principles, Policy & Practice, 21*(4), 442–463.

Zapata-Rivera, D., Vezzu M., & K. Biggers. (2013, April). *Supporting Teacher Communication with Parents and Students Using Score Reports.* Paper presented at the annual meeting of the American Educational Research Association (AERA), San Francisco, CA.

Zapata-Rivera, D. Vezzu, M., Nabors Olah, L, Leusner, D., Biggers, K., and Bertling, M. (2014, April). *Designing and Evaluating Score Reports for Parents who are English Language Learners.* Paper presented at the annual meeting of the American Educational Research Association (AERA), Philadelphia, PA.

Zapata-Rivera, D., Zwick, R., & Vezzu, M., (2016). Exploring the Effectiveness of a Measurement Error Tutorial in Helping Teachers Understand Score Report Results. *Educational Assessment, 21*(3), 215–229. http://dx.doi.org/10.1080/10627197.2016.1202110

Zenisky, A. L., & Hambleton, R. K. (2012). Developing test score reports that work: The process and best practices for effective communication. *Educational Measurement: Issues and Practice, 3*1(2), 21–26.

Zwick, R. Zapata-Rivera, D & Hegarty, M. (2014). Comparing Graphical and Verbal Representations of Measurement Error in Test Score Reports, *Educational Assessment, 19*(2), 116–138. https://doi.org/10.1080/10627197.2014.903653

**Citation**:

Kannan, P., Zapata-Rivera, D., & Bryant, A.D. (2021). Evaluating parent comprehension of measurement error information presented in score reports. *Practical Assessment, Research & Evaluation*, 26(12). Available online: https://scholarworks.umass.edu/pare/vol26/iss1/11/

## Corresponding Author

Priya Kannan
Educational Testing Service

email: pkannan [at] ets.org