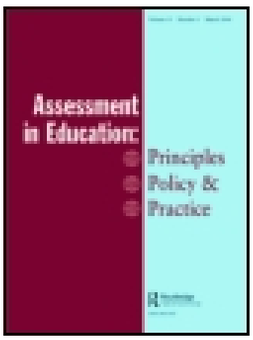


Routledge
Taylor & Francis Group

Assessment in Education: Principles, Policy & Practice

ISSN: 0969-594X (Print) 1465-329X (Online) Journal homepage: http://www.tandfonline.com/loi/caie20

Comparing the influence of various measurement error presentations in test score reports on educational decision-making

Dorien Hopster-den Otter, Selia N. Muilenburg, Saskia Wools, Bernard P. Veldkamp & Theo J. H. M. Eggen

To cite this article: Dorien Hopster-den Otter, Selia N. Muilenburg, Saskia Wools, Bernard P. Veldkamp & Theo J. H. M. Eggen (2018): Comparing the influence of various measurement error presentations in test score reports on educational decision-making, Assessment in Education: Principles, Policy & Practice, DOI: 10.1080/0969594X.2018.1447908

To link to this article: https://doi.org/10.1080/0969594X.2018.1447908

Published online: 13 Mar 2018.

ASSESSMENT IN EDUCATION: PRINCIPLES, POLICY & PRACTICE, 2018
https://doi.org/10.1080/0969594X.2018.1447908

Routledge
Taylor & Francis Group

Comparing the influence of various measurement error presentations in test score reports on educational decision-making

Dorien Hopster-den Otter[a,b], Selia N. Muilenburg[a], Saskia Wools[b], Bernard P. Veldkamp[a] and Theo J. H. M. Eggen[a,c]

[a]Faculty of Behavioral, Management and Social Sciences (BMS), University of Twente, Enschede, The Netherlands; [b]CitoLab, Cito Institute for Educational Measurement, Arnhem, The Netherlands; [c]Psychometric Research Centre, Cito Institute for Educational Measurement, Arnhem, The Netherlands

ABSTRACT

This study investigated (1) the extent to which presentations of measurement error in score reports influence teachers' decisions and (2) teachers' preferences in relation to these presentations. Three presentation formats of measurement error (blur, colour value and error bar) were compared to a presentation format that omitted measurement error. The results from a factorial survey analysis showed that the position of a score in relation to a cut-off score impacted most significantly on decisions. Moreover, the teachers (N = 337) indicated the need for additional information significantly more often when the score reports included an error bar compared to when they omitted measurement error. The error bar was also the most preferred presentation format. The results were supported in think-aloud protocols and focus groups, although several interpretation problems and misconceptions of measurement error were identified.

ARTICLE HISTORY
Received 21 August 2017
Accepted 20 February 2018

KEYWORDS
Assessment; test score reports; measurement error; educational decision-making; preferences

In education, decision-making is an everyday activity. For example, teachers make decisions about the next steps in instruction, the placement of students into different instruction groups or the need to provide a student with additional support. Since these decisions may have serious consequences for teaching and learning, they need to be informed by high-quality evidence (Brookhart & Nitko, 2008).

Several data sources can be used to inform decisions, such as student observations, oral questions, students' work, parental reports and test scores (Brookhart & Nitko, 2008; Mandinach, 2012). Due to a careful construction process, test scores are often considered a valuable source. In general, these scores are regarded as very reliable and non-biased (Shepard, 2006); however, they are also subject to a certain amount of measurement error (Gardner, 2013).

CONTACT Dorien Hopster-den Otter d.denotter@utwente.nl

Measurement error (ME) can be conceptualised as the difference between a student's actual or obtained score and the theoretical true score counterpart (Gardner, 2013). Feldt and Brennan (1989) list four categories of ME: (a) inherent variation in human performance, (b) variations in the environment within which the measurements are obtained, (c) variations in the evaluation of responses and (d) variation arising from the selection of the test items asked. In practice, different measures are used to quantify ME, including the standard error of measurement, the standard error of estimation and the test information function, depending on the measurement model being used.

As some degree of ME is common to all tests, corroboration between the test score and additional data sources is often recommended (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 2014). This recommendation is even more important if the test score contains a relatively large ME or if it, along with its ME, is positioned around the cut-off score of high-stakes decisions that cannot easily be reversed (AERA, APA, & NCME, 2014). High-stakes decisions may trigger major consequences for students, for example, students might not be assigned to an appropriate instruction group and, thus, might not get the instruction they need (e.g. Goodman & Hambleton, 2004; Newton, 2005; Phelps, Zenisky, Hambleton, & Sireci, 2010). When combined with other data sources that are potentially more authentic, such as student observations or other test scores, a more accurate picture of the student can be obtained, and decisions can be better informed (Brookhart & Nitko, 2008; Mandinach, 2012).

The extent to which ME around test scores influences teachers' educational decisions is hitherto unknown. This influence, however, determines the usefulness of displaying ME. On one hand, confusion around the concept of ME could result misinformed decisions with adverse consequences for students. Several studies indicate some misunderstanding by teachers around the interpretation of ME visualisations (e.g. Impara, Divine, Bruce, Liverman, & Gay, 1991; Zwick, Zapata-Rivera, & Hegarty, 2014). Considering the possible consequences of misinformed decisions, test designers would avoid the presentation of ME in score reports (Bradshaw & Wheater, 2009; Epp & Bull, 2015) or would place this information in the technical manuals (Phelps et al., 2010). On the other hand, the lack of ME reporting could be a serious problem as teachers would interpret test scores more accurately than they might be. Therefore, test publishers would have a duty to provide teachers with error information that would allow them to make valid inferences based on test results (AERA, APA, & NCME, 2014). Although teachers may not have a full understanding of the nature of ME, the presentation of ME could lead to greater awareness about the imprecision around test scores compared to a score report that omits ME. This awareness could stimulate teachers to gather additional information about a student's ability. Decision-making based upon multiple sources of information, after taken the validity and accuracy of this additional information into account as well, can result in more informed decisions.

This study investigated the extent to which various presentation formats of error information influence teachers' decisions within the context of primary education. Specifically, we examined the extent to which the ME presentation formats result in the need to gather additional information to enable decision-making regarding students, for example, from other information sources. The need for additional information was defined as an indication for awareness of ME. Furthermore, we investigated teachers' own perspectives on the presentation of ME. We asked teachers about their preference levels for each presentation

format since several studies have suggested that user preference and performance did not always coincide (e.g. Wainer, Hambleton, & Meara, 1999; Zwick et al., 2014). Teachers' decisions and preferences were examined in the context of a familiar type of action: the assignment of students into instruction groups. Two research questions were formulated:

(1) To what extent do various ME presentation formats result in teachers' need for additional information compared to a presentation format that omit ME?
(2) Which of the various presentation formats do teachers prefer?

The presentation of measurement error

The presentation of error information has received growing attention across a range of disciplines outside the field of education (e.g. Brodlie, Osoria, & Lopes, 2012; Kinkeldey, MacEachren, Riveiro, & Schiewe, 2015), resulting in the development of many potential visual tools for presenting ME. To help designers choose a presentation format, various taxonomies have been proposed (e.g. Gershon, 1998; Pang, Wittenbrink, & Lodha, 1997), and several review studies have been conducted (e.g. Epp & Bull, 2015; Kinkeldey, MacEachren, & Schiewe, 2014; Kinkeldey et al., 2015; MacEachren et al., 2005). These studies conclude that the presentation format could make a difference for user decision-making and understanding of the concept. Based on these studies, three promising formats presenting ME will be further explored: blur, colour value and error bar. Figure 1 presents these formats as well as a presentation format that omits ME.

Blur can be defined as changes in the clarity or fuzziness of objects (Epp & Bull, 2015). The technique provides a general overview of uncertainty without quantifying exact values. It seems to be a promising and widely used tool for presenting error information because users intuitively associate blur with uncertainty (e.g. Johnson & Sanderson, 2003; MacEachren et al., 2012).

Colour value is a naturally orderable presentation format that can be defined according to changes from light to dark (Epp & Bull, 2015). Lighter values are associated with higher uncertainty, while darker values correspond to lower uncertainty (e.g. Kinkeldey et al., 2014; Leitner & Buttenfield, 2000). Colour value is used as a categorical presentation format containing a number of discrete value levels.

Error bars are additional graphic objects in the visualisation (Gershon, 1998). Because of the numerical and continuous representation, it is a suitable technique for presenting quantitative data (Brodlie et al., 2012; Wainer, 1995). Several studies have however concluded

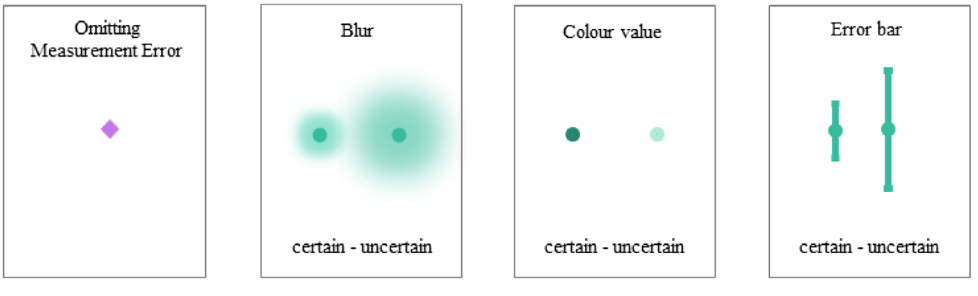


Figure 1. ME presentation formats compared to the presentation format omitting ME.

that error bars dominate the certainty scores because the greatest visual emphasis is on the long bars, that present the most uncertainty (e.g. Sanyal, Zhang, Bhattacharya, Amburn, & Moorhead, 2009). In addition to the amount of uncertainty, the length of the bar is influenced by the type of confidence interval (e.g. 68, 90 or 95%) that is represented. A sufficient level of statistical literacy is required to accurately interpret the length of the bar (Hullman, Rhodes, Rodriguez, & Shah, 2011; Zwick et al., 2014). Nevertheless, it is a commonly used technique for visualising ME within educational contexts (e.g. Phelps et al., 2010).

Method

A mixed-methods design was used to examine teachers' decisions and preferences regarding the various presentation formats (blur, colour value, error bar and omitting ME). Quantitative data were collected by means of a factorial survey. Qualitative data were collected by means of think-aloud protocols and focus groups to verify our findings, and to obtain a deeper analysis of the quantitative results.

Design of the visualisations

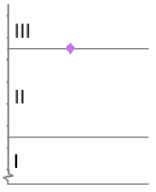
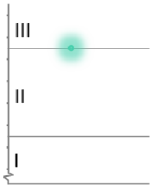
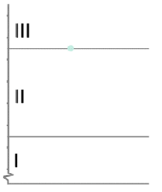
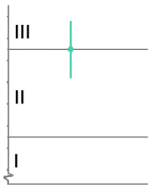
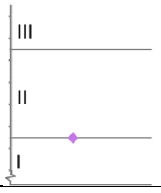
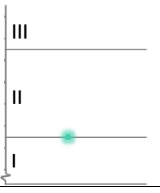
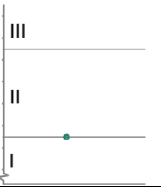
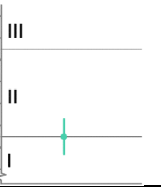
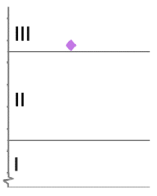
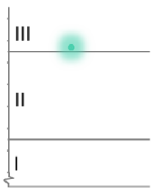
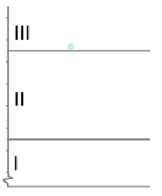
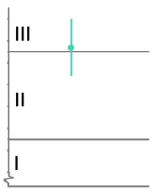
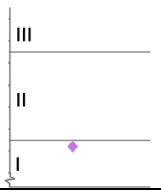
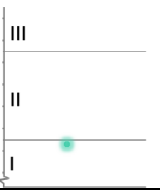
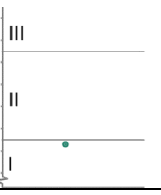
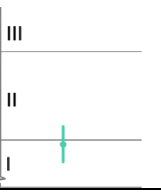
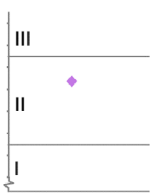
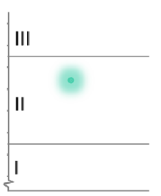
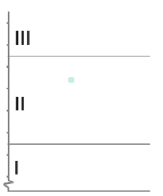
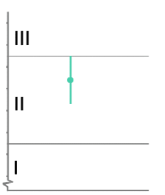
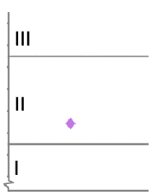
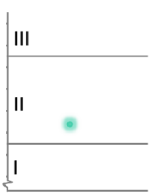
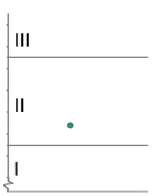
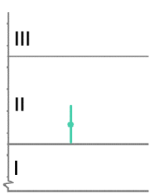
Real student data from a standardised test were used to develop the test score reports. This test is used at 85% of Dutch primary schools and covers various domains of mathematics (e.g. counting and comparing numbers and addition and subtraction sums). The data are usually gathered every six months to monitor student performance and to develop a group action plan for the next six months.

For this test, ME is commonly determined by calculating the standard error of the ability estimate using the one-parameter logistic model of item response theory. To simulate a real decision-making process, this calculation was also used in the current study. This resulted in a 68% confidence interval consisting of one standard error above and one standard error below the ability estimate or score. Due to the use of actual data, the confidence interval for the higher score points was smaller than for lower score points. However, since this occurred due to the use of actual data, it was not altered.

Because blur, colour value and error bar were considered promising presentation formats in the literature regarding the presentation of ME, we incorporated these formats into this study. This resulted into the comparison of two categorical (blur and colour value) and a numerical presentation format (error bar) with a presentation format omitting ME. Each of these presentation formats is associated with a certain amount of ME information. For example, the error bar is an exact and continuous presentation of the ME values, while blur and colour value provide only a global indication of the amount of ME. We investigated how these characteristics influence teachers' decisions and preferences.

In order to obtain a valid representation of the influence of the ME presentation formats, other essential ME characteristics that could influence teachers' decisions were investigated. Six educational measurement specialists were interviewed to indicate other essential ME characteristics that could influence teachers' decisions. Based on their input, two other characteristics were added: the position of the error in relation to the cut-off score (i.e. the cut-off score is outside, within or exactly in the middle of the confidence interval) and the size of the error (i.e. large or small). This resulted in four presentations x three positions x two sizes = 24 visualisations for each respondent (see Table 1).

Table 1. Test score report visualisations.

Position and size	Presentation format			
	Omitting ME	Blur	Colour value	Error bar
Exactly in the middle Large				
Small				
Within Large				
Small				
Outside Large				
Small				

Respondents

Data on 487 pre-service and in-service teachers of Dutch primary education were collected, after contacting pre-service teachers from all 44 Dutch colleges as well as in-service teachers by email, Facebook and LinkedIn. From these 487 teachers, 150 did not complete the survey, which means that the responses of 337 teachers (N_{male} = 40; N_{female} = 297) were

used for analysis. The male to female ratio is typical for the Dutch primary school teacher population (www.onderwijscijfers.nl).

The teaching experience of the teachers varied: 77.7% of them were pre-service teachers in the last year, 6.8% taught less than 5 years, 5.0% taught 5–10 years and 10.4% taught more than 10 years. Furthermore, 82.2% of the teachers did not take a course on testing during their study, and 86.9% of them indicated that they had little or no statistical experience. Think-aloud protocols were conducted with a typical selection of 14 teachers, and 8 focus groups were held ($N = 35$) with an average of 4 teachers per focus group.

Instruments and procedure

Survey

To study teachers' decision-making processes, a factorial survey with true-to-life cases was developed (Auspurg & Hinz, 2014). In this survey, all the respondents were presented with all 24 visualisations (Table 1). We started with the six visualisations omitting ME, which is the usual format presented to Dutch teachers. As we started with these formats, the respondents' answers were not influenced by the ME visualisations. Following this, the 18 visualisations containing blur, colour value and error bar were shown in random order as set by the online survey programme. Teachers were shown a single visualisation on the screen, which showed a score report of one student. For every visualisation, the respondents were asked to judge a familiar type of educational action: the assignment of students to an instruction group for tailored instruction.

In the Netherlands, the assignment of students is often done by dividing them into three instruction groups: (I) an extended instruction group, (II) a basic instruction group and (III) a shortened instruction group. The 25% lowest scoring students are usually assigned to the extended instruction group for which teachers provide additional instruction using concrete learning materials. The 25% highest scoring students are commonly assigned to the shortened instruction group, in which they receive brief instruction and in-depth exercises. The remaining 50% of the students are assigned to the basic instruction group, in which teachers provide regular instruction.

In this study, respondents were asked to specify which instruction groups (I, II, III) they would assign the student to or to indicate that they needed additional information about the student to make this decision. The need for additional information was defined as an indication for awareness of ME. It suggested the desire to gather multiple sources of information before making a decision, since the single test score contains some uncertainty.

Alongside the investigation of respondents' decisions regarding the 24 visualisations, the survey consisted of 6 items on the respondents' background and three questions about respondents' preferences (Appendix A). With regard to their background, the respondents were asked questions about their gender, their level of educational attainment, their years of experience teaching primary education, the name of the high school of teacher training, their courses on testing and how much experience they have with statistics (i.e. no, little (followed one course), medium (followed multiple courses), large (work activities)). With regard to the preference questions, respondents were asked to rank the four presentation formats from most preferred to least preferred. Furthermore, they were asked to indicate the extent to which the various presentation formats influenced their decisions as well as the extent to which the error presentations influenced their confidence in their decisions.

The survey was pretested, with 22 test experts completing the survey and indicating whether some questions were unclear. Subsequently, we pretested the survey with two teachers. Both pretests resulted in some minor adaptations, like changing the score point of no measurement error into a purple rhombus for a clear distinction with the colour value presentation. During the pretest and data collection, the survey was completed online by the respondents.

Think-aloud protocols and focus groups

Think-aloud protocols were used to obtain insight into the cognitive processes underlying the teachers' decision-making processes (Bannert & Mengelkamp, 2008). The respondents were asked to verbalise their thoughts (i.e. think-aloud) while responding to items in the survey. The researcher was not allowed to request explanations because this could interfere with the respondents' cognitive processes.

After filling out the survey, focus groups were held to verify and clarify the findings of the survey. The respondents were asked to indicate their decision for a varying selection of four visualisations and to explain their choice. Furthermore, we investigated their interpretation of the score report and their comprehension of the ME concept, an explanation of their preferences and their perspective regarding the usefulness of visualising ME. The design of the focus group method included the characteristics of a group interview as well as a group discussion (Newby, 2010). The researcher fulfilled the role of moderator.

The think-aloud protocol and focus group were pretested with three teachers, resulting in some points of attention. The verbalisation of the think-aloud protocols and the focus group discussion were tape recorded.

Data analysis

Teachers' decisions

To test whether the presentation formats resulted in a significantly greater need for additional information, the respondents' answers to the 24 score reports were recoded into dichotomous variables. Score '0' indicated the assignment of a student to groups I, II or III. Score '1' indicated the need for additional information.

After performing frequency analyses, we conducted a Generalised Linear Mixed Model (GLMM) using the lme4 package for R (Bates et al., 2017). This model provides a method for analysing a dichotomous dependent variable in hierarchically structured data, which means a dependent variable containing precisely two distinct values and a data-set that is organised at more than one level. In this study, the teachers' decisions were defined as the dichotomous dependent variable, containing a score '0' (i.e. no need for additional information: assignment of student to group I, II or III) or '1' (i.e. need for additional information). The data were hierarchically structured, given the 24 cases of data nested within each respondent. This data structure resulted in a random intercept for persons. The independent variables were the teachers' background variables and the visualisation characteristics' position, size and format.

We started with a simple random intercept model containing a fixed intercept and a random intercept for persons. The independent variables were then added successively, and the fit of the new model was compared to the previous one. As the previous model was nested in the new one, a likelihood ratio (LR) test was used to test the improvement in goodness of fit. The resulting test statistic is χ^2 distributed, with the number of free

parameters of the alternative model minus the number of free parameters in the null model as the degrees of freedom. Furthermore, the AIC, BIC and -LL indices were used, with lower values indicating a better fit.

In addition to the survey data, think-aloud transcriptions were divided into 24 units belonging to the 24 visualisations. For each unit, we identified factors that the teachers kept in mind during the decision-making and the categories of misconceptions emerging for certain presentation formats. The final coding scheme was used to double-code 10% of the transcriptions. An inter-rater reliability analysis was subsequently performed to determine the consistency between the two raters, which was found to be substantial (Cohen's $\kappa = .738$). The transcriptions of the focus groups regarding the respondents' decisions and their interpretation of the score report and ME concept were classified and summarised.

Teachers' preferences

To analyse the respondents' preferences regarding the presentation formats, the respondents' answers to the second part of the survey were analysed using frequency analysis. Furthermore, the discussion in the focus group around the preferences and usefulness of ME were summarised. In this article, the results are illustrated by examples translated from Dutch.

Results

Teachers' decisions

Frequency analysis showed that the error bar format most often resulted in the need for additional information (see Table 2). The blur and colour value formats both resulted in less need for additional information compared to the omitting ME format. According to the respondents' think-aloud protocols, additional information for all instruction groups comprised information about previous test scores, scores of peers, sub-scores of the corresponding test, working attitude, student age and student anxiety.

Table 3 presents an overview of the comparisons between the estimated models of the GLMM analysis. Model 1a included all background characteristics. Only statistical experience had a significant effect ($F(3, 8075) = 4.84$, $p = .002$) on the decisions. Respondents who rated themselves as having a great deal of statistical experience requested additional information more often than respondents who rated themselves as having no ($B = -1.53$, $SE = .58$, $p = .009$), little ($B = -1.64$, $SE = .57$, $p = .004$) or quite a lot of ($B = -2.17$, $SE = .59$,

Table 2. Percentage of respondents ($N = 337$) needing additional information for each visualisation.

	Presentation format			
Position and size	Omitting ME	Blur	Colour value	Error bar
Exactly in the middle				
Large	69.4	69.4	67.4	66.8
Small	63.8	67.1	60.5	71.5
Within				
Large	46.0	47.2	51.0	53.4
Small	51.6	43.9	32.3	54.6
Outside				
Large	13.1	16.3	21.4	18.7
Small	18.1	14.5	9.8	20.8
Total	43.7	43.1	40.4	47.6

Table 3. Overview of the estimated models.

Model	Nested model	Effects		AIC	BIC	-LL	#p	LR test	
		Fixed	Random over persons					df	χ^2
0		Intercept	Intercept	9543.1	9557.1	−4769.6	2		
1a	0	+ sex, level of education, teacher experience, education in tests, and statistical experience	″	9547.8	9645.7	−4759.9	14	12	19.37
1b	0	+ statistical experience	″	9533.8	9568.8	−4761.9	5	3	15.32*
2	1b	+ format, position, size	″	7331.8	7408.8	−3654.9	11	6	2214*
3a	2	+size × format, size × position, position × format	″	7287.0	7441.0	−3621.5	22	11	66.76*
3b	2	+ size × format	″	7281.2	7379.2	−3626.6	14	3	56.55*

*$p < .01$.

$p < .001$) statistical experience. No additional statistical differences were found between respondents who rated themselves as having no, little or quite a lot statistical experience. Because of the significant improvement of Model 0 (see Model 1b), we decided to retain this background variable in subsequent analyses.

In Model 2, the role of the presentation format, position and size on the respondents' decisions was examined by adding these as fixed effects. The model improved significantly as all effects were reported as significant at $p \leq .001$.

Based on the results of Model 2 and the visualisation of the presentation formats, we investigated the interaction effects between the presentation format, position and size in Model 3a. We hypothesised at least an interaction between presentation format and size because the colour value, blur and error bar formats differ in size from each other. The model improved significantly as the interaction between presentation format and size was significant at $p < .001$. Based on this result, we removed the other interactions and maintained only the interaction between format and size in Model 3b. Model 3b resulted in the best fitting model, showing significant fixed effects for statistical experience, presentation format, position, size and size × presentation format interaction. We found random intercepts for persons (*Variance = 4.30; SD* = 2.07; $p < .001$). The results of Model 3b are presented in Table 4 and discussed in detail below.

Presentation format

Model 3b showed a significant main effect for presentation format ($F(3, 8075) = 13.628$, $p < .001$). The error bar presentation format resulted in the most need for additional information. In order to yield interpretable odds ratios, the fixed effect must first be exponentiated.

Table 4. Estimates of unstandardised (*B*) and standardised (*β*) effects on teachers' decisions in Model 3b.

	B	*SE*	*β*	*p*
Intercept	.26	.75	.00	.725
Background characteristics				
Statistical experience[a] (reference: A great deal (work activities))				
No	−2.70	.77	−2.68	<.001
Little (one course)	−2.81	.77	−2.82	<.001
Quite a lot (more courses)	−3.32	.83	−2.04	<.001
Visualisation characteristics				
Format[b] (reference: Omitting ME)				
Blur	−.21	.12	−.18	.091
Colour value	−.82	.13	−.72	<.001
Error bar	.35	.12	.30	.005
Position[c] (reference: Outside)				
Exactly in the middle	3.70	.10	3.51	<.001
Within	2.42	.09	2.30	<.001
Size[d] (reference: Small)				
Large	−.13	.12	−.13	.287
Interaction				
Size × Format[e]				
Large × Blur	.33	.17	.22	.063
Large × Colour value	1.11	.18	.74	<.001
Large × Error bar	−.08	.17	−.05	.661

[a] $F(3, 8075) = 2.09$, $p = .098$.
[b] $F(3, 8075) = 13.628$, $p < .001$.
[c] $F(2, 8075) = 707.304$, $p < .001$.
[d] $F(1, 8075) = 10.134$, $p = .001$.
[e] $F(3, 8075) = 19.023$, $p < .001$.

Thus, the estimated odds that additional information will be chosen for the error bar format above the omitting ME format is exp(0.35) = 1.42 times.

No significant difference was found between the blur and omitting ME formats ($B = -.21$, $SE = .12$, $p = .091$). According to the think-aloud protocols, it seems that respondents' interpretation of blur was too literal, as opposed to the actual meaning, as blur reflected a categorical value. For example, respondent 3 interpreted the outline of blur as a 68% confidence interval: 'Okay, I see a dot and a blur around it, indicating that this student scored around 37. However, (…) it could also be a score of 35 or even 40'. As a small blur presentation was smaller than the real 68% confidence interval, the blur presentation suggested a smaller ME size than it actually was.

Colour value resulted in significantly less need for additional information ($B = -.82$, $SE = .13$, $p < .001$). The estimated odds that additional information would be chosen for colour value above omitting ME were exp(−0.82) = 0.44 times. The think-aloud protocols illustrated that the lighter and darker values were not as associative as they should have been. Respondent 11 thought: 'This is certain, right? No, this is uncertain?' Respondent 13 illustrated that even a change in the meaning, such as a light colour value format, was associated with certainty: 'This student has a certain score at the border of group III. Therefore, I would assign her to group III purely because it is a certain score'.

To sum up, different ME presentation formats did not always result in a need for additional information. The respondents chose to seek additional information significantly more often only for the error bar. The blur format did not change the decisions, probably because of its literal interpretation. Colour value resulted in significantly less need for additional information, probably due to the confusing association of the values.

Position

Model 3b yielded a significant main effect for position, $F(2, 8075) = 707.304$, $p < .001$. Given the standardised effects in Table 4, the decisions were most impacted by the position of a score in relation to a cut-off score. The estimated odds that additional information would be chosen for a cut-off score exactly in the middle of the error above a cut-off score outside the error was exp(3.70) = 40.45 times. The odds for a cut-off score within the error above a cut-off score outside the error was exp(2.42) = 11.25 times. This meant that the respondents would more often request additional information in the event of a cut-off score exactly in the middle of the error, followed by a cut-off score within and outside the error. This corresponds to the idea that it would be more difficult to assign a student to a group when the test score approaches the cut-off score.

The think-aloud protocols corroborated this result. For example, respondent 7 argued about a cut-off score exactly in the middle of the error: 'This student scored exactly between groups I and II, so I would like to have additional information about the extent to which she is able to perform in group II'. This was in contrast with a cut-off position outside the error: 'This student is in the second group, so I would assign her to the second group'.

Size

Model 3b showed a significant main effect for size on decision ($F(1, 8075) = 10.134$, $p = .001$). However, the size estimates were no longer significant in Model 3b ($B = -0.13$, $SE = .12$, $p = .287$) as a result of the significant interaction effect between size and presentation format ($F(3, 8075) = 19.023$, $p < .001$). This indicated that size had different effects on respondents'

decisions, depending on which presentation format was shown. Decomposition of the interaction revealed that a dark colour value format – which is a small error size – resulted in more assignments of students to groups ($B = 1.11$, $SE = .18$, $p < .001$).

The think-aloud protocols showed that respondents considered a dark colour value as a very certain format. Respondent 14, for example, argued: 'This is a certain score in group I. Therefore, I would assign this student to group I'. By contrast, the other presentation formats were seen as less certain because the small error size was even larger than the colour value point: 'This [error bar] is a little uncertain; group II, I, III? Let me think. Perhaps additional information because it is a little vague' (respondent 12). Furthermore, some respondents confused the colour value format by regarding it as the smallest error size in terms of blur: 'I would assign him to the second group since there is no variation around the score' (respondent 13).

To sum up, there was an interaction effect between size and presentation format. Colour value resulted in significantly less need for additional information because a dark colour value was interpreted as very certain. As a result, the respondents tended to assign many students when a dark colour value was presented; however, this format also included a small error size.

Conceptions and misconceptions of ME

Although the error bar resulted in an increased need for additional information, the respondents varied in their understanding of ME. Several reasons explaining the cause of ME were given. All focus groups indicated the cause of ME as a variation in human performance and environment, such as influences relating to the well-being of the student or the location. In three out of eight focus groups, a respondent indicated the cause of ME as a variation caused by the selection of test items: 'I think it is about the selection of items. The more items you select, the more precise the results are – like the more research you do, the more solid your research is. I think that's what it means'.

However, there were also several misconceptions around the ME concept (see Table 5). Respondents in three focus groups attributed the cause of ME, for example, to the difference between the test score and the perspective of the teacher. During the think-aloud, respondent 6 thought that the error indicated the uncertainty of the students themselves, like their test anxiety. Furthermore, respondents from three focus groups had no idea about the causes of ME.

Teachers' preferences

Frequency analyses showed that the error bar was the most preferred presentation format for ME ($M = 3.06$, $SD = 1.09$), followed by blur ($M = 2.63$, $SD = 0.93$), colour value ($M = 2.58$, $SD = 0.97$) and omitting ME ($M = 1.72$, $SD = 1.05$). The respondents in the focus groups preferred the error bar due to the exact presentation of the numerical ME values, while the blur and colour values were categorical presentations. The clear borders of the error bar were most highly appreciated, although the vague borders of the blur format led the respondents to consider ME more often. Furthermore, both the higher values of the error bar and blur were seen as associated with greater uncertainty, while the association of a lighter colour value with greater uncertainty was lacking. One disadvantage of the error bar according to the respondents was the extensive length of some bars, resulting in less

Table 5. Misconceptions about (the Cause of) ME among Focus Groups (N = 8).

Misconception	Number of focus groups	Illustrative example
Difference between test score and other test scores	5	*'Perhaps contradicting scores on tests from teaching methods compared to standardised tests – so the standardised test has a very high score, and the score on the teaching method test is very low'*
Difference between test score and teacher's perspective	3	*'Maybe a teacher adds the uncertainty. He or she is not sure about the test score, or the score is not in line with his or her perspective on the student'*
Difference between process and outcome	2	*'The student performed the calculations well but gave the wrong answer. So you do not know what exactly went wrong'*
Response pattern of the student	1	*'Is it about duration: how much time does a child need to answer the question? (…) Or is it about changing the filled-in response often?'*
An invalid item	1	*'You have a contextual item in a math test, but the student cannot read, which resulted in a wrong answer, even though he is a good math student. (…) Then the question is not certain'*

confidence in the decision. Blur was less preferred because the width of the blur format had no meaning. A disadvantage of colour value was the limited possibility to convert the presented coloured format in a black and white score report when printed at home or at school, as it will become poorly readable. The reasons for omitting ME were the prevention of confusion and the availability of sufficient test information, including other test scores. A disadvantage of omitting ME was the lack of insight into the reliability of a test.

The results of the survey showed that the respondents believed that the different ME presentations affected their educational decisions. On a five-point Likert scale ranging from never to always, error bars influenced sometimes ($M = 3.11$), followed by blur ($M = 3.02$) and colour value ($M = 2.72$). During a think-aloud protocol, a respondent said:

> I notice that I've often indicated the need for additional information. This is quite logical for me because I now realise that the test scores are not as exact as they seem, and you still want to make good decisions for your students. (Respondent 3)

The presentation of ME did not affect confidence in educational decisions for 51.4% of the respondents, and 30.8% of them indicated that the presentation of ME had a positive impact on their confidence. For example,

> Anyhow, a lot more confidence: because now you do justice to the students. If you did not use this information, if you did not have the representation of uncertainty, you would make your decision based on an exact presented score. And now, you have included the influencing factors on the test score. (Respondent 3)

The remaining 17.8% indicated less confidence in their decisions, as indicated by respondent 10: 'Yes, you will still doubt because you see that there maybe uncertainties. I think this gives less confidence. I would like to indicate the need for additional information more often'.

Conclusion and discussion

This study set out to determine teachers' decisions and preferences regarding various ME presentations. Quantitative and qualitative data were collected by means of a factorial survey, think-aloud protocols and focus groups.

The results showed that ME presentations influence teachers' educational decisions compared to presentations that omit ME. The error bar format resulted in significantly greater need for additional information about a student. The colour value format resulted in significantly less need for additional information, while the blur format did not differ significantly from the omitting ME format. Furthermore, the results showed that the position of a score in relation to a cut-off score had the most impact on the decisions. The size was influenced by the format and had no independent effect in this study.

Moreover, the error bar was found to be the most preferred format because of its exact presentation of the numerical ME values. The desirability of this advantage can be questioned because ME is not exact. It is an estimated value and can be visualised by a 68% confidence interval as well as by 95 and 99% confidence intervals. By contrast, the vague borders of blur ensured that there would be no exact interpretation, resulting in further thoughts about the ME concept. However, the respondents interpreted blur as a numerical variable, while this study presented blur as a categorical variable. Colour value and omitting ME were the least preferred presentation formats.

As every study is accompanied by some measurement error, we should draw the conclusions of this study carefully. The first limitation is that 150 respondents did not complete the

survey. Since we do not know whether they differed (e.g. regarding their statistical experience) from those who completed the experiment, we urge caution in the interpretation of the results of the study. Secondly, the frequency of the decision regarding the need for additional information can be underestimated as the think-aloud protocols showed that the respondents assigned students to one of the groups but in fact wanted to gather additional information. For example, respondent 9 said: 'I would assign this student to group III and observe the progress. It is difficult to know that with one test score'. Thus, the assignment of students to an instruction group was less a matter-of-fact decision for teachers than we assumed. In addition, the think-aloud protocols provided insight into the teachers' cognitive processes; however, it seems that the teachers did not repeat their reasoning for each visualisation. Although the investigator encouraged the respondents to continue thinking aloud, and visualisations were shown in random order, the results might be an underestimation of the number of times the teachers really looked at the format, position and size when taking a decision. Finally, the context of the current study may have influenced the results obtained. Although we chose a common type of educational decision and a commonly used test, other types of decisions and tests may result in teachers wanting more or less additional information. For example, the presentation of real test results resulted into visualisations in which a small error size is always accompanied with lower scores compared to a large error size with higher scores. The results of the focus groups and think-aloud protocol, however, did not give reason to think that the height of student's score is confounded with the error size.

The results and limitations point to some suggestions for future research. First, this study indicates the fruitfulness and necessity of evaluating score reports with the intended audience so that they can be interpreted and used in a valid way. Therefore, based on this study's results, we suggest an investigation into whether a combination of blur and error bar functions can be a suitable presentation of ME in test score reports. Advantages relating to the error bar include the numerical presentation, the positive influence on decisions and teachers' preferences. Those relating to blur are the natural association with uncertainty and the avoidance of exact interpretation. A combination of blur and error bar is known as a gradient plot (see Correll & Gleicher, 2014) and consists of an error bar with blurred ends. As Correll and Gleicher recommend this gradient plot for indicating uncertainty among general audiences, it is interesting to examine the extent to which this presentation is deemed suitable for presenting ME in test score reports. Second, it seems worthwhile to examine the influence of other design factors on teachers' understanding and use, which were less relevant for the currently used context of test scores. For example, the width of the *Y*-as, the number of cut-off scores and the visualisation of previous scores can change the way in which teachers make their decisions. Moreover, it would be interesting to study the influence of ME on other kinds of educational decisions such as planning regarding the next steps in instruction. Third, despite the potential impact of ME presentation, the teachers demonstrated several new misconceptions about the concept itself. Future research is needed into teachers' understanding and misconceptions of ME and effective ways to reduce misconceptions. The study of Zapata-Rivera, Zwick, and Vezzu (2016) is a useful contribution to this area. It developed an ME tutorial to help teachers understand score report results. Future research should investigate the long-term effects of such tutorials on teachers' interpretation and use of test scores.

The findings of this study enhance our understanding of the usefulness of displaying ME. The results can be used in the design of new test score reports. Practical implications would

include the use of ME in order to make teachers more aware of the imprecision around scores as well as fostering the use of multiple sources for taking educational decisions, such as other test scores, observations and students' work. In deciding on the use of alternative sources, it is important to consider the psychometric characteristics that are inherent to specific data sources. The findings also imply the need for a clear explanation of the ME-concept as several misconceptions of teachers were identified. This way, carefully designed test score reports could lead to a better understanding by teachers, thereby improving the quality of their educational decisions.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Dorien Hopster-den Otter is a PhD candidate at the Research Centre for Examinations and Certification (RCEC), a collaboration between Cito and the University of Twente. Her research interests include formative assessments, psychometrics and score report development. She is currently working on projects relating to the quality of educational tests and the development of an educational master's programme for test development.

Selia N. Muilenburg used to work as a test developer at Cito and is now a PhD candidate at the University of Twente. Her research interests include working with data in educational settings, professional learning communities and professional development of teachers and school leaders. She is currently working on a project related to the role of school leadership in realising sustainable school improvement.

Saskia Wools is an educational researcher and manager of CitoLab, Cito's research and development department. Her research interests include the validity and validation of educational assessments. She is currently working on innovative projects involving formative assessments, educational technology and assessment quality.

Theo J. H. M. Eggen is a senior research scientist at the Psychometric Research Centre of Cito and a professor of psychometrics at the University of Twente in the Netherlands. He is director of the Research Centre for Examinations and Certification (RCEC). His research interests include the quality of educational testing and computerised (adaptive) testing. He is also working on projects like PISA (international educational survey) and on a variety of projects on educational assessment.

Bernard P. Veldkamp is head of the Department of Research Methodology, Measurement and Data Analysis and the scientific director of the Research Centre for Examination and Certification (RCEC). His research interests include measurement optimisation and behavioural data science. His interests focus on computerised assessment. Current projects are about automated test assembly where uncertainty in the item parameters due to measurement error is taken into account.

ORCID

Dorien Hopster-den Otter http://orcid.org/0000-0003-4009-2045

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Auspurg, K., & Hinz, T. (2014). *Factorial survey experiments. Applications for the Social Sciences* (Vol. 175). Thousand Oaks, CA: Sage Publications.

Bannert, M., & Mengelkamp, C. (2008). Assessment of metacognitive skills by means of instruction to think aloud and reflect when prompted. Does the verbalisation method affect learning? *Metacognition and Learning, 3*, 39–58. doi:10.1007/s11409-007-9009-6

Bates, D., Maechler, M., Bolkers, B., Walker, S., Christensen, R. H. B., Singman, H., … Green, P. (2017). *The lme4 package*. Retrieved from http://r-forge.r-project.org/projects/lme4/

Bradshaw, J., & Wheater, R. (2009). *National foundation for educational research: International survey of results reporting (OFQUAL 10/4705)*. London: Office of Qualifications and Examinations.

Brodlie, K. W., Osoria, R. A., & Lopes, A. (2012). A review of uncertainty in data visualization. In J. Dill, R. Earnshaw, D. Kasik, J. Vince, & P. C. Wong (Eds.), *Expanding the frontiers of visual analytics and visualization* (pp. 81–109). London: Springer.

Brookhart, S. M., & Nitko, A. J. (2008). *Assessment and grading in classrooms*. Upper Saddle River, NJ: Pearson Education.

Correll, M., & Gleicher, M. (2014). Error bars considered harmful: Exploring alternate encodings for mean and error. *IEEE Transactions on Visualization and Computer Graphics, 20*, 2142–2151. doi:10.1109/TVCG.2014.2346298

Epp, C. D., & Bull, S. (2015). Uncertainty representation in visualizations of learning analytics for learners: Current approaches and opportunities. *IEEE Transactions on Learning Technologies, 8*, 242–260.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York, NY: American Council on Education and Macmillan.

Gardner, J. (2013). The public understanding of error in educational assessment. *Oxford Review of Education, 39*, 72–92. doi:10.1080/03054985.2012.760290

Gershon, N. (1998). Visualization of an imperfect world. *IEEE Computer Graphics and Applications, 18*, 43–45. doi:10.1109/38.689662

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*, 145–220. doi:10.1207/s15324818ame1702

Hullman, J., Rhodes, R., Rodriguez, F., & Shah, P. (2011). *Research on graph comprehension and data interpretation: Implications for score reporting (ETS RR-11-45)*. Paper presented at the ETS Score Reporting conference, Princeton, NJ.

Impara, J. C., Divine, K. P., Bruce, F. A., Liverman, M. R., & Gay, A. (1991). Does interpretive test score information help teachers? *Educational Measurement: Issues and Practice, 10*(4), 16–18.

Johnson, C. R., & Sanderson, A. R. (2003). A next step: Visualizing errors and uncertainty. *IEEE Computer Graphics and Applications, 23*(5), 6–10.

Kinkeldey, C., MacEachren, A. M., Riveiro, M., & Schiewe, J. (2015). Evaluating the effect of visually represented geodata uncertainty on decision-making: Systematic review, lessons learned, and recommendations. *Cartography and Geographic Information Science, 44*, 1–21. doi:10.1080/15230406.2015.1089792

Kinkeldey, C., MacEachren, A. M., & Schiewe, J. (2014). How to assess visual communication of uncertainty? A systematic review of geospatial uncertainty visualisation user studies. *The Cartographic Journal, 51*, 372–386. doi:10.1179/1743277414Y.0000000099

Leitner, M., & Buttenfield, B. P. (2000). Guidelines for the display of attribute certainty guidelines for the display of attribute certainty. *Cartography and Geographic Information Science, 27*, 3–14. doi:10.1559/152304000783548037

MacEachren, A. M., Robinson, A., Hopper, S., Gardner, S., Murray, R., Gahegan, M., & Hetzler, E. (2005). Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science, 32*, 139–160. doi:10.1559/1523040054738936

MacEachren, A. M., Roth, R. E., O'Brien, J., Li, B., Swingley, D., & Gahegan, M. (2012). Visual semiotics & uncertainty visualization: An empirical study. *IEEE Transactions on Visualization & Computer Graphics, 18*, 2496–2505.

Mandinach, E. B. (2012). A perfect time for data use: Using data-driven decision making to inform practice. *Educational Psychologist, 47*(2), 71–85. doi:10.1080/00461520.2012.667064

Newby, P. (2010). *Research methods for education*. Harlow: Pearson Education Limited.

Newton, P. E. (2005). The public understanding of measurement inaccuracy. *British Educational Research Journal, 31*, 419–442. doi:10.1080/01411920500148648

Pang, A. T., Wittenbrink, C. M., & Lodha, S. K. (1997). Approaches to uncertainty visualization. *The Visual Computer, 13*, 370–390. doi:10.1007/s003710050111

Phelps, R. P., Zenisky, A., Hambleton, R. K., & Sireci, S. G. (2010). *On the reporting of measurement uncertainty and reliability for U.S. educational and licensure tests* (OFQUAL 10/4759). London: Office of Qualifications and Examinations.

Sanyal, J., Zhang, S., Bhattacharya, G., Amburn, P., & Moorhead, R. J. (2009). A user study to compare four uncertainty visualization methods for 1D and 2D datasets. *IEEE Transactions on Visualization and Computer Graphics, 15*, 1209–1218. doi:10.1109/TVCG.2009.114

Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational measurment* (4th ed., pp. 623–646). Westport: American Council on Education and Praeger.

Wainer, H. (1995). *Depicting error* (Technical Report No 95-2). Princeton, NJ: Educational Testing Service.

Wainer, H., Hambleton, R. K., & Meara, K. (1999). Alternative displays for communicating NAEP results: A redesign and validity study. *Journal of Educational Measurement, 36*, 301–335. doi:10.1111/j.1745-3984.1999.tb00559.x

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*, 215–229. doi:10.1080/10627197.2016.1202110

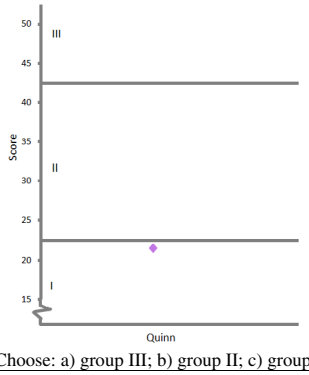
Zwick, R., Zapata-Rivera, D., & Hegarty, M. (2014). Comparing graphical and verbal representations of measurement error in test score reports. *Educational Assessment, 19*, 116–138. doi:10.1080/10627197.2014.903653

Appendix A.

Table A1. Survey.

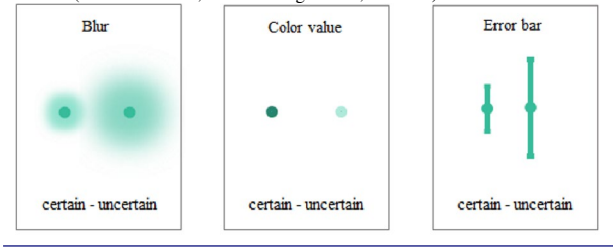
Respondents background
1. What is your sex?
a) Male; b) Female
2. What is your highest level of educational attainment?
a) Higher general secondary education; b) Pre-university education; c)Vocational education; d) Higher education; e) University
3. At which high school do you attend the teacher training? (Only for final-year pre-service teachers)
4. Did you take a course on testing during the teacher training?
a) No; b) Yes, namely….
5. How much experience do you have with statistics? Chose the most appropriate answer.
a) I have no experience with statistics; b) I have little experience with statistics (e.g. one course during secondary education); c) I have quite a lot of experience with statistics (e.g. more courses); d) I have a great deal of experience with statistics (e.g. more courses and own work activities).
6. How many years' experience do you have in primary education? (only for in-service teachers)
a) Less than 5 years; b) 5 to 10 years; c) More than 10 years

Score reports omitting ME
On the next page, you will see the test score of a group of students on the national mathematics test. The score reports will be used to create a group action plan for the next semester. The group action plan will consist of three groups:
Group I: extended instruction [consisting of students who get additional instruction]
Group II: basic instruction [consisting of students who require the regular amount of instruction]
Group III: shortened instruction [consisting of students who only need brief instruction]
Please assign each student to a group (I, II or III) or indicate that additional information (from other tests, method assignments, etc.) would be needed to make this decision.
NB. The number of students per group may not be the same. For example, you may also assign all students to Group III. Base your choice on the corresponding score report, and do not look at the reality of the group action plan.



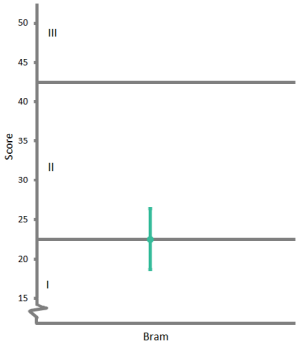
Choose: a) group III; b) group II; c) group I; d) I need additional information about the student to make this decision.

Score reports with ME
On the next page, you will see the score reports of the other students for the national mathematics test. For each score report, we added information about the certainty regarding the test score as a good estimation of the students' mathematics skills, other factors reaming equal. The figure on this page shows three examples of test score reports.
Indicate the group (I, II or III) to which you would assign each student, or indicate that additional information would be needed (from other tests, method assignments, et cetera) in order to make this decision.



(*Continued*)

Table A1. *(Continued)*



Choose: a) group III; b) group II; c) group I; d) I need additional information about the student to make this decision.

Preference

Finally, we look forward to your experience and preferrences for these presentations.

1.Which of the presentations do you prefer? Order them according to: 1 = most preferred; 4 = least preferred.

a) Presentation A. Error bar; b) Presentation B. Blur; c) Presentation C. Colour value; d) Presentation D. Omitting ME

2. To what extent did the presentation of uncertainty affect your decision compared to the presentation omitting uncertainty?

a) Error bar: never-rarely-sometimes-very often- always; b) Blur: never-rarely-sometimes-very often- always; c) Colour value: never-rarely-sometimes-very often- always

3. To what extent did the presentation of uncertainty affect your confidence regarding your decision compared to the presentation omitting uncertainty?

a) Less confidence 1-2-3-4-5 More confidence

4.Comments section