High Speed High Stakes Scoring Rule

Assessing the performance of a new scoring rule for digital assessment

S. Klinkenberg

Weesperplein 4
1018 XA Amsterdam
s.klinkenberg@uva.nl

Abstract. In this paper we will present the results of a three year subsidized research project investigating the performance of a new scoring rule for digital assessment. The scoring rule incorporates response time and accuracy in an adaptive environment. The project aimed to assess the validity and reliability of the ability estimations generated with the new scoring rule. It was also assessed whether the scoring rule was vulnerable for individual differences. Results show a strong validity and reliability in several studies within different domains: e.g. math, statistics and chess. We found no individual differences in the performance of the HSHS scoring rule for risk taking behavior and performance anxiety, nor did we find any performance differences for gender.

Keywords. "computer adaptive testing", "speed accuracy trade-off", "scoring rule", "digital assessment", validity, reliability, CAT, DIF

1 Introduction

This paper covers the results of the project: "New scoring rule for digital assessment" performed for SURF, the national collaborative organization for ICT in Dutch higher education and research. The project was part of a nationwide tender called "Testing and Test-Driven Learning". The program stimulated institutions to cooperate in digital testing. It aimed to generate a positive impact of digital testing in terms of study success, lecturer workloads and test quality (SURF[1]). In the following sections we will describe the speed accuracy trade-off, guessing behavior in testing and how the high speed high stakes (HSHS) scoring rule could offer a solution for digital assessment.

[1] http://www.surf.nl/en/themes/learning-and-testing/digital-testing/testing-and-test-driven-learning-programme/index.html

1.1 Speed accuracy trade-off

One of the two classic problems in assessment is the trade-off between speed and accuracy (Wickelgren, 1977). The problem concerns for example the comparison of two respondents, of whom one answered more items correct and the other responded faster. The question is how speed should be balanced against accuracy. Much research has been done on this subject. Not only in the assessment domain but also within experimental psychology and psychonomics. The developed solutions within experimental psychology are based on mathematical decision models (Bogacz et al., 2006). One of the best known examples is the Ratcliff diffusion model for dichotomous decisions where a respondent decides when the evidence transcends a certain threshold (Ratcliff, 1978). This model describes the relationship between speed and accuracy very well (Ratcliff & Rouder, 1998), but is hard to apply in the context of testing and examination. The estimation of the model parameters of the diffusion model requires many observations of behavior within one person on identical items (Vandekerckhove & Ruerlincks, 2007). Within psychometrics, Van der Linden (2007) recently proposed a hierarchical model where speed and accuracy are modelled separately. Item and person parameter within this model are merged on a higher level. This approach is applicable to digital testing.

Both approaches share that the trade-off between speed and accuracy is left to the respondent and is modeled afterwards. Individual differences in chosen strategies will affect ability scores. It is conceivable that sequential answering of items will result in different results than selective answering, e.g. first answering easier items. Strategy choice can be reduced by imposing a scoring rule. The effectiveness of such a rule depends on the understanding of the rule by the participant.

1.2 Guessing behavior

The second classical problem consists of guessing behavior of respondents. Due to costs and psychometric problems with the scoring of open ended questions, multiple choice questions are still very popular, but they have their constraints. Simply stated, chance plays an important role in the test results, especially for respondents who score just below or above the caesura. To reduce the role of chance, test constructors must either increase the amount of items or the amount of answer options, which results in overly complex or long assessments. Some respondents may be better at guessing, excluding irrelevant alternatives or distributing the available time for all items in a test. This differentiation poses a threat to the unidimensionality of the test. Many solutions have been proposed in the history of psychometrics (Lord, 1974).

The most frequently used scoring rule that we know is the sum correct rule, which sums the amount of correctly answered items. The result of this rule is that every wrong and non answer has a negative effect on the final score. Students who are aware of this will benefit by always giving an answer, while students who are not will lose points when leaving an item unanswered. The probability of a correct answer is indeed 1/M, where M is the number of answer options. A scoring rule that corrects for the number of answer options is the following. Suppose we have an item with M an-

swer options. Respondents can choose to answer or skip an item. For skipping respondents gain no points, for answering correct respondents gain one point and for an incorrect answer they lose c points. Now there is a penalty for incorrect responses. When the penalty c is larger than 1/(M-1) it is unwise to make a random guess. Suppose there are M=4 answer options and the penalty c=1, then the expected value for a blind guess is negative -1/2. If c is equal to 1/(M-1), this scoring rule is known as correction for guessing (Holzinger, 1924; Thurstone, 1919; Lord, 1975). If respondents do not have a clue of the wright answer, then the expected value for guessing or skipping is equal. This scoring rule has been for many years implemented in important assessments in the US (Budescu and Bar-Hillel, 1993). The success of this rule has been debated. Burton (2005) and Lord (1975) are mostly positive but Budesco and Bar-Hillel (1993) have expressed concerns. The choice for c=1/(M-1) is a bit strange for when subjects blindly guess it does not matter if they guess or skip the item, but if they possess just a bit of (partial) knowledge, it is always better to guess. All in all it is always better to guess, though not all respondents understand that. The majority of honest students will skip all items that they are not sure of, resulting in systematic score reduction. This of course can be solved by choosing c>1/(M-1), but even then the drawback proposed by Budesco and Bar-Hillel remains that knowing how to use this rule implies an added skill in with respondents can systematically differ.

1.3 High speed high stakes

The scoring rules described in the previous section only provide a solution for guessing, but not for the speed accuracy trade-off problem. Van der Maas and Wagenmakers (2005) proposed a solution for the second problem. Their scoring rule consists of a per item time limit d, in their study on chess ability measuring 30 seconds. The accuracy (acc) was scored 0 for incorrect and 1 for correct answers. The score per item was equal to the remaining time multiplied by the accuracy: acc(d-RT). A wrong answer will result in no points while a fast correct response will yield more than a slow correct response. Using this scoring rule, Van der Maas and Wagenmakers (2005) managed to increase the validity of their test. Maris and Van der Maas (2010) proposed an important improvement to this rule. The described rule has the disadvantage that it can, as the earlier rules, provoke risk taking behavior. If the respondent recognizes that an item is too difficult, it is better to guess immediately, since a score higher than zero is then still probable. Maris and Van der Maas therefore propose to make the rule symmetric by transforming accuracy to -1 (incorrect) and 1 (correct). The same formula (d-RT) multiplied by acc*2-1 now makes fast guessing extremely risky. A fast wrong answer will result in a very negative score (fig. 1).

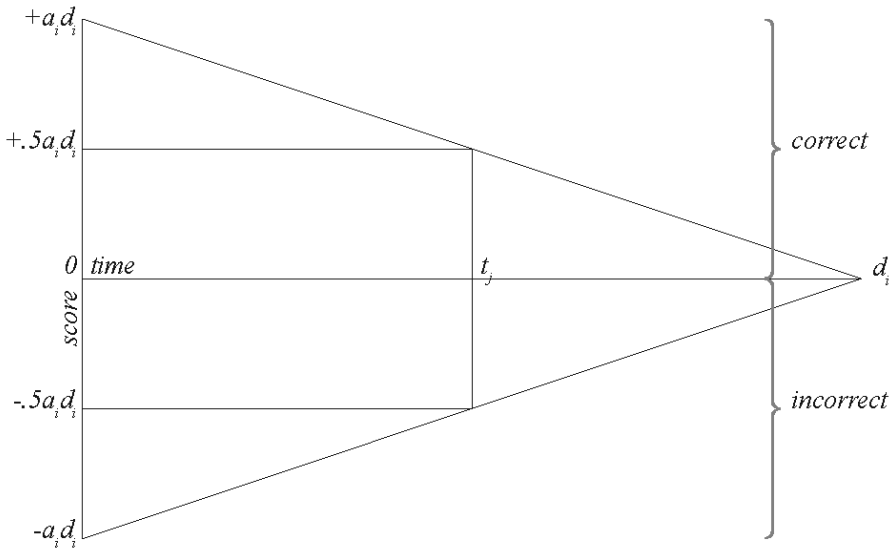

Figure 1: High speed high stakes scoring rule, where d is the time limit and a is a scaling factor.

The high speed high stakes scoring rule thereby offers a solution for guessing as well as the speed accuracy problem. Given that the certainty about an answer increases with time (an assumption in almost all decision theories), there is an optimal moment for actually responding. Interestingly, Maris and Van der Maas (2005) have proven, providing that the scoring rule is a sufficient statistic for measuring ability, that the model for the probability for answering correctly is identical to the most frequently used model in assessment, namely the two parameter logistic model (Van der Linden & Hamleton, 1997). The discrimination parameter is shown to be equal to the time limit d for the item. This elegant result offers many opportunities for, among others, adaptive testing with reaction times. Maris & van der Maas (2011) have to this end derived all relevant conditional probability distributions.

It is still relevant that respondents understand the rule. Through digital assessment this rule can easily be visualized. Figure 2 shows an implementation of this rule in the Math Garden. Respondents see their remaining time decreasing with the amount of available coins. The result of their response is shown by increasing or decreasing the total amount of coins, hereby shortening the feedback loop.

Figure 2: Math garden multiplication game.

In the following section we will concisely present the main results of the carried out research. We will attempt to answer the following research questions.

1. Does the use of the HSHS scoring rule result in an increased reliability and validity?
2. Are respondents able to find an optimal balance between speed and accuracy?
3. Are there individual differences in this ability?
4. Do these differences relate to background variables as experience, gender and ability?
5. Do respondents need to learn how to use the scoring rule or can it be applied easily?
6. Does the use of the scoring rule result in more accurate ability estimations in adaptive testing with easy items?

To answer these questions data has been analyzed from the Math Garden, from data gathered at the CORUS chess event 2008 and results from the "statistiekfabriek". A short description of these three sources is in order here. In 2007 the department of psychological methods from the University of Amsterdam initiated the development of the Math Garden, a computer adaptive practice and tracking system for math in which the HSHS scoring rule was implemented. Currently Math Garden is commercialized by Oefenweb and about 1100 schools in the Netherlands are subscribed to the service. The responses, about half a million per day, form the basis for the Math Garden data set. Parallel to this development at the CORUS chess tournament of 2008, a chess test was administered where the same scoring rule was used. National and in-

ternational chess players ranging from novices to grand masters participated in this event. Their responses form the chess test data set. Within SURF's nationwide tender "Testing and Test-Driven Learning" another project ran in which a statistics version based on the Math Garden was being developed, called "Statistiekfabriek". The results of that project form the basis for the statistics data set.

In the following section a brief overview of the main results from the different data sources will be presented. We have chosen not to include the result section as the theoretical introductions due to the resumptive nature of this paper. The descriptions can be found in the original works, though some are only available in Dutch.

2 Results

2.1 Validity

To get an indication of the convergent validity the scores had to be compared with an external measure. For the Math Garden data, scores could be compared to the national Dutch norm reverenced CITO scores. The chess players from that data set all had national or international chess ratings that could be used for comparison. Finally the "Statistiekfabriek" scores could be compared to different partial exams. Convergent validity criteria where thus available for all three data sets.

The correlations with the external measure proved significant for all three sources. For the Math Garden and the chess data sets, these where particularly high. Figure 3 shows the scatterplots for the Math Garden where the domains addition, subtraction, multiplication and division where correlated with the CITO scores (Klinkenberg, 2011). The correlations ranged from .78 to .84 all with $p<.05$. Regression lines are also plotted for each grade.

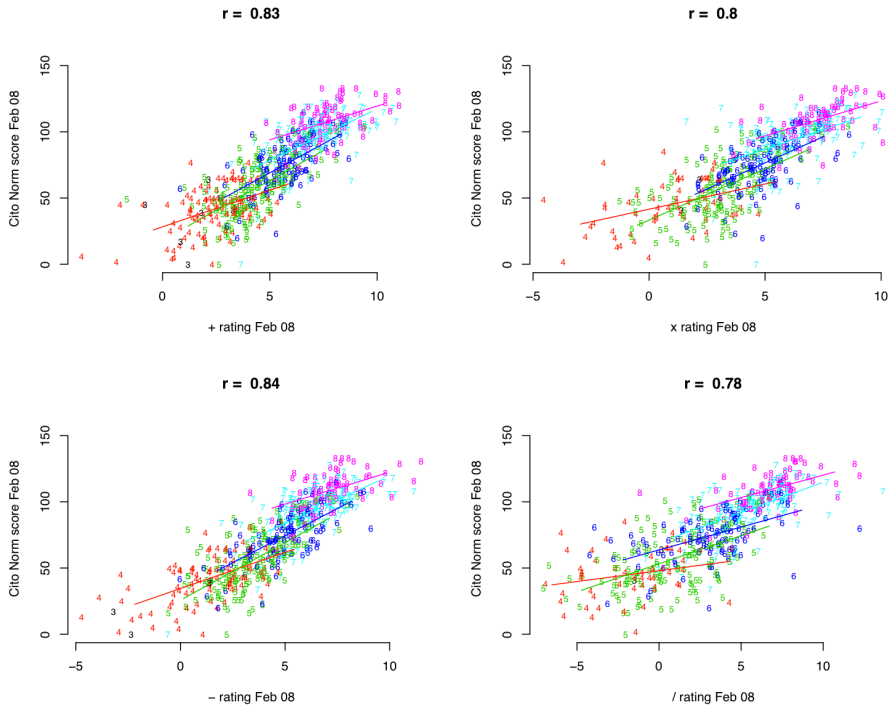

Figure 3: Correlations of HSHS rating with CITO scores.

The chess scores also showed, with regard to the known Elo ratings, high correlations (Table 1). It is striking that the correlation between the HSHS sum score is higher than the sum correct score with the FIDE rating but not with the tournament performance rating (TPR). Evidently ratings based on an adaptive procedure perform considerably better (Klinkenberg & Van der Maas, 2013).

Table 1: Correlations of test performance and with known chess ratings (FIDE) and tournament performance ratings (TPF). All *p*'s < .05.

Response time	Method	FIDE	TPR
Excluded	Som score	0.575	0.547
Included	HSHS test rating	0.808	0.777
	HSHS som score	0.617	0.525

The correlations between the statistics exams and the HSHS sum scores and the sum correct rule also indicate a better performance. The sum correct score correlated r=.39 while the HSHS sum score correlated r=.48 though these differences were not significant (Özen et. al., 2012).

A different way of looking at validity is to assess if ability increases over time. It could be expected that children in higher grades in the Math Garden would perform better at math. This has been analysed with the math data (Klinkenberg, 2011). It was

tested whether children in higher grades performed better on the four domains. Figure 4 shows the increase of rating for ascending grades.

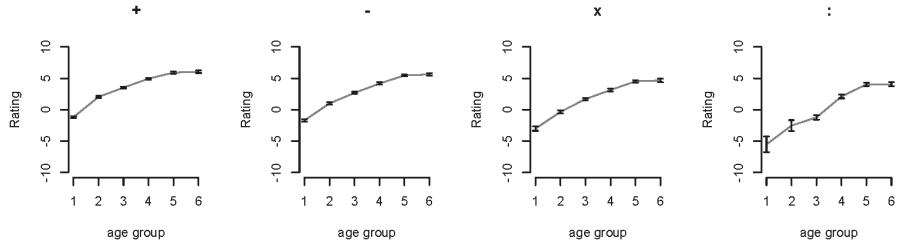

Figure 4: Rating per domain per group.

We see a clear increase in ability across ascending grades except for the last grade. This can be explained by the fact that in Dutch education no new knowledge is taught in the final grade.

Everything indicates that the HSHS score gives a valid indication of ability. In most cases the rule performs better than the sum correct, especially while testing adaptively.

2.2 Reliability

The reliability of the HSHS test scores r=.60 did not significantly differ from the reliability of the sum correct scores r=.57 in "statistiekfabriek" report 1 (Özen et. al., 2012). Here non adaptive tests where administered using both rules. Determining the reliability in an adaptive dynamic test is somewhat more challenging. Given the adaptive method, normal procedures do not apply. To still get an indication of reliability in the Math Garden, parallel test where used for the domain addition and multiplication. For these domains n+m and nxm could be paralleled with m+n and mxn, resulting in correlations of r=.74 for addition and r=.71 for multiplication ($p<.05$) (Klinkenberg, 2011).

A different approach is to look at test retest reliability. The difficulty of items at point x should than be equal to that at point x+y. This would indicate stability of item difficulty over time and therefore a reliable measurement tool. The dotted line in figure 5 shows that after two months, item ratings are stable and do not fluctuate much in the following months (Klinkenberg, 2011). These results show that both the HSHS test scores as the adaptive HSHS test rating are reliable indicators of ability.

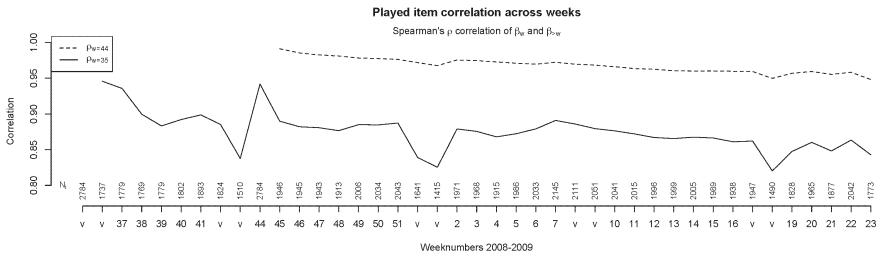

Figure 5: Stability of items ratings for initial ratings (solid line) and established ratings after 2 months (dotted line). The x-axis displays week numbers (v = vacation). Correlations are computed over active (played) items in each week (N_i = amount of administered items).

2.3 Optimal SAT

To determine the optimal balance between speed and accuracy given the HSHS scoring rule, we plotted the frequency distribution of correct and incorrect answers at response time intervals. Figure 6 superimposes the scoring rule on these distributions. Given the HSHS scoring rule one would expect, if the rule is understood correctly, that respondents would not guess quickly, but would respond as quickly as possible once they found the answer. The expected frequency distribution would in that case be skewed to the right. The expected distribution for incorrect answers is harder to predict. One would at least expect few fast responses but the frequency of slower responses could increase.

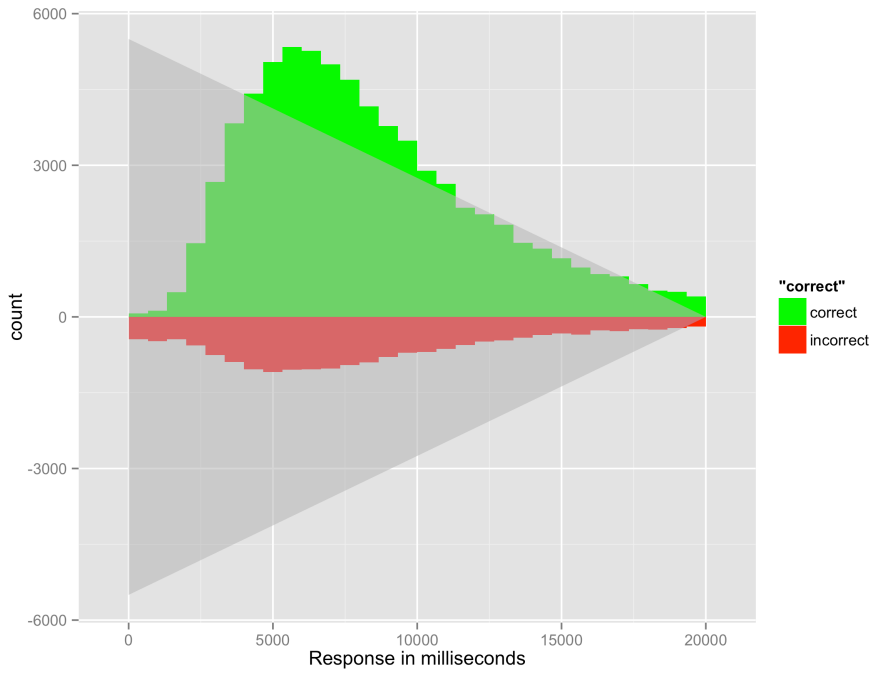

Figure 6: Frequency distributions of correct and incorrect responses with the HSHS scoring rule superimposed.

Figure 6 shows both distributions based on 100,000 responses of 65,000 unique players in the Math Garden. The distribution of correct answers looks as expected. Very fast responses are rare, given that respondents need a few seconds to read the question. For incorrect responses we see no massive guessing and a rather flat distribution. It is notable that in the first two seconds there is a relatively high frequency of incorrect answers. We presume this could be accounted for by a small group of risk takers. Further analysis would have to shed some light on this.

2.4 Individual differences

With the introduction of response times in the assessment the question immediately rises if respondents could feel pressured and therefore perform worse than without a time constrain. The susceptibility for this could be a determining factor in the ability estimation and therefore diminishing the effect of the HSHS scoring rule.

"Statistiekfabriek" report 2 (Barkhof et. al., 2013) suggests no relation between positive or negative performance anxiety and ability. Ability correlated r=.09, p=.142 with negative and r=.01, p=.85 with positive performance anxiety. The latter is noteworthy because positive performance anxiety was expected to enhance ability.

There also was no relation between risk orientation and the mean deviation with the expected score. Based on the available information of the item difficulty and the ability of respondents generated by to the computer adaptive algorithm, the expected

score could be inferred. It was expected that risk takers would deviate more from this expected score and that their variation would be larger than non risk takers. Neither showed in the data. The mean deviation correlated r=.01, p=.125 and the dispersion of the deviation correlated r=.02, p=.404 with risk orientation.

Risk orientation also did neither relate to response time nor to accuracy. The correlations were r=-.06, and r=.03 respectively (p>.05). Differences in the amount of risk orientation did not manifest in the speed of responding or the amount of correct answers. Nor was there a difference in ability between males and females.

In "statistiekfabriek" report 1 (Özen et. al., 2012) the effect of risk taking behavior was examined to assess if the items showed differential item functioning (DIF). Students were asked to indicate if they saw themselves as risk takers or not. Subsequently, it was investigated if items performed different for these groups. Only one out of twenty items showed DIF when the HSHS scoring rule was applied, in comparison to six out of twenty when applying the sum correct rule.

The above analyses imply that the HSHS scoring rule in these samples does not work differently for respondents with varying amounts of sensitivity to performance anxiety. Nor did the items perform different for risk takers and non risk takers.

In Jansen et. al. (2013) we were not able to show, while using the adaptive assessment procedure in the Math Garden with the HSHS scoring rule, that the perceived math anxiety decreased with ascending levels of administered item difficulty. Also the perceived math competence was not higher with easier items. Math performance did increase as the difficulty of administered items was lower. As mediation analysis showed, this appeared to be mediated by the amount of items played. Administering easy items increased the playing frequency which in turn resulted in higher math performance. In all situations there was no effect of grade nor where there any differences between boys and girls.

2.5 Learning the rule

Both in the Math Garden as in the "Statistiekfabriek" respondents were able to apply the HSHS scoring rule immediately. The short feedback loop ensures that respondents directly see what the result of their response is. Figure 2 showed how this was implemented in the Math Garden. In the "Statistiekfabriek" the available time is visualized by a descending counter and the total earned points is displayed as a numeric indicator (fig. 7).

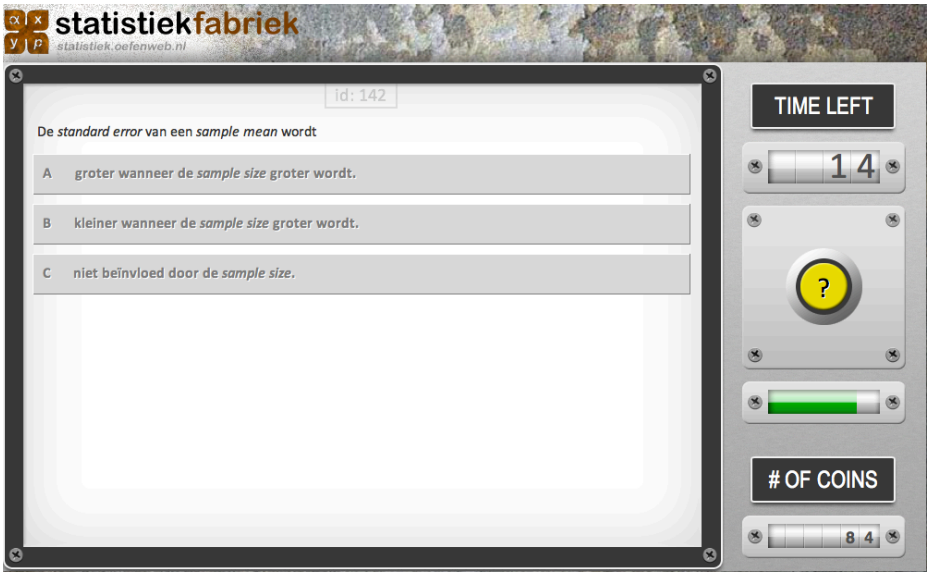

Figure 7: "Statistiekfabriek" descriptive statistics game. Remaining time is displayed in the upper right, total earned coins in the lower right corner.

Students were asked to indicate what they thought would result in the highest possible points per question. Most realized that speed and accuracy played the primary role but figure 8 also shows that a portion thought only speed was important and some did not have a clue and only a few thought only accuracy was involved.

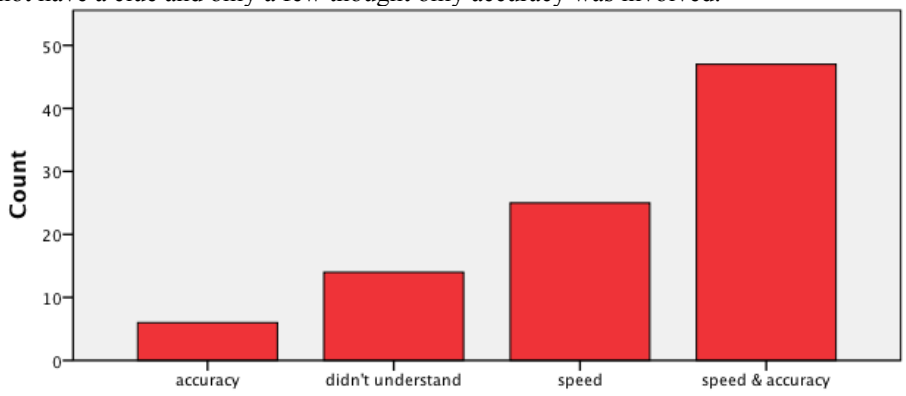

Figure 8: HSHS scoring rule assesment.

With the exception of one person, a limited amount of guessing took place while performing a statistics trial exam. Of the 20 items in the trial exam using the HSHS scoring rule the frequency distribution was skewed to the right, indicating that most students guessed less than eight items out of twenty. It appeared, however, that the appreciation of the HSHS scoring rule varied widely. Figure 9 shows these verdicts

on a Likert scale ranging from 1, strongly negative, to 7, strongly positive. Semi high stakes testing using the HSHS scoring rule appear to result in some mixed verdicts.

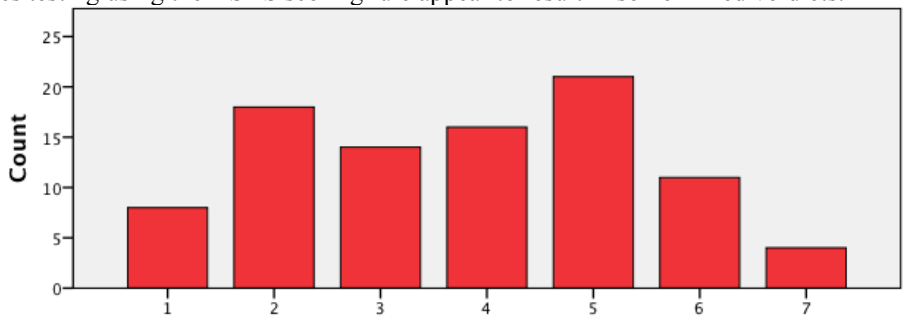

Figure 9: Positive vs negative verdict about the HSHS scoring rule.

2.6 Ability estimation

To assess if the use of the HSHS scoring rule in an adaptive setting results in accurate estimations of ability while administering easy items, it is necessary to perform some simulations. Figure 10 displays the simulations performed by Klinkenberg (2011). The left graph shows the mean deviation from the simulated 'real' ability for a weighted maximum likelihood method (Eggen et. al., 2006), an Elo procedure based only on accuracy (Elo + 1PL), and an Elo procedure based on the HSHS scoring rule (Elo + HSHS). The x-axis shows the item difficulty from hard (left) to easy (right). The graph on the right shows the standard error of these deviations for the same three procedures supplemented by the maximum information possible in a maximum likelihood setting. With easy items, the HSHS procedure exhibits the least amount of bias and the standard error is lower than when only accuracy is involved. The HSHS scoring rule is potentially better at estimating ability while administering easy items.

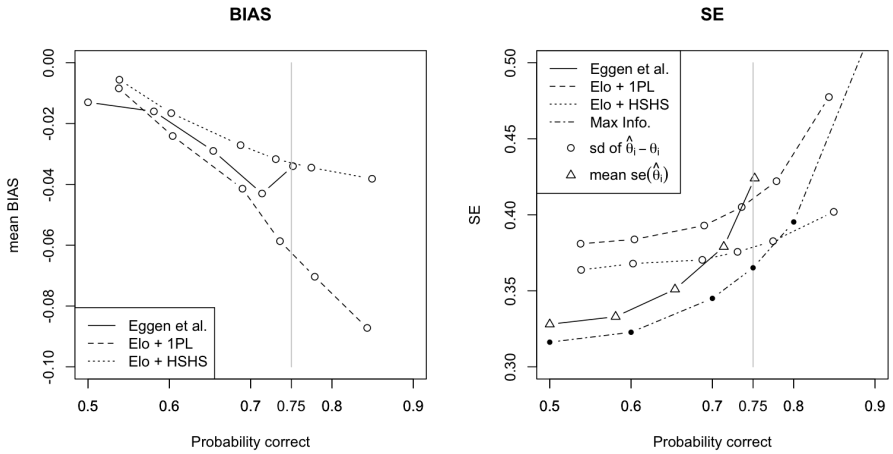

Figure 10: Bias and standard error for different computer adaptive methods at different values of the expected probability correct.

3 Conclusion

Applying the HSHS scoring rule to both a non-adaptive and an adaptive setting yields valid and reliable estimations of ability. Though the reliability does not seem to be better than other scoring rules, the validity is shown to be higher than when only accuracy is used, especially in an adaptive setting. In particular when administering easy items, the adaptive HSHS scoring rule performs better. Risk taking disposition does not seem to influence the performance of the HSHS scoring rule in terms of ability estimation, though respondents do indicate a wide range of positive and negative feelings towards the rule. While the frequency distributions of response times for correct and incorrect answers show that respondents comply with the aim of the scoring rule, a considerable amount of students also reported a wrong interpretation of the rule. Either the intuitive behavior does not coincide with the perceived rule or the "statistiekfabriek" implementation fundamentally differed from the Math Garden. This is partly the case as the "statistiekfabriek" incorporated a semi high stakes environment by applying the rule in a trial exam. Experiencing the rule for weeks on end in the Math Garden would result in a more elaborate understanding of the rule than applying is once in a trial exam. This would argue for getting students to familiarize with the rule before applying it in high stake assessments.

We are encouraged by the research findings in the ability of the HSHS scoring rule to produce valid and reliable estimations of ability, though we remain mindful of individual differences and of the perceived attitude towards the rule in high stakes testing. We are, nonetheless, confident these attitudes are less of an issue in low stakes testing. The HSHS scoring rule promises to bridge the gap between speed and accuracy and we think we are on track with this approach.

4 References

1. Barkhof, J., Bekker, T., Bersma, M., Groenendijk, E., Maza, S. (2013). Oefening Baart Kunst (unpublished research report). University of Amsterdam, Netherlands.
2. Bogacz, R., Brown E., Moehlis, J., Holmes, P., & Cohen, J. C. (2006). The Physics of Optimal Decision Making: A Formal Analysis of Models of Performance in Two-Alternative Forced-Choice Tasks. Psychological Review, 113(4), 700-765.
3. Budescu, D. and M. Bar-Hillel (1993). "To Guess or Not to Guess: A Decision-Theoretic View of Formula Scoring." Journal of Educational Measurement 30(4): 277-291.
4. Burton, R. F. (2005). "Multiple-choice and true/false tests: myths and misapprehensions." Assessment & Evaluation in Higher Education 30(1): 65-72.
5. Eggen, T. J. H. M. & Verschoor, A. J. (2006). Optimal Testing with Easy or Difficult Items in Computerized Adaptive Testing. Applied Psychological Measurement, v30 n5 p379-393.
6. Holzinger, K. J. (1924). "On Scoring Multiple Response Tests." Journal of Educational Measurement 15: 445-447.
7. Jansen, B.R.J., Louwerse, J., Straatemeier, M., Van der Ven, S.H.G., Klinkenberg, S., Van der Maas, H.L.J. (2013). The influence of practicing maths with a computer-adaptive program on math anxiety, perceived math competence, and math performance. Learning and Individual Differences, 24, 190–197.
8. Klinkenberg, S., Straatemeier, M., Van der Maas, H. L. J. (2011). Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. Comput. Educ., 57(2), 1813–1824.
9. Klinkenberg, S., Van der Maas, H.L.J. (2013). A dynamic paired comparison based computer adaptive testing method. Unpublished manuscript.
10. Lord, F. M. (1975). "Formula Scoring and Number-right Scoring." Journal of Educational Measurement 12(1): 7-11.
11. Maris, G. & Van der Maas, H. L. J. (2012). Speed-accuracy response models: scoring rules based on response time and accuracy. *Psychometrika, 77*(4), 615-633.
12. Özen, S., Pronk, A., Sanchez Maceiras, S., Stel, N., Van Wersch, T. (2012). De Invloed van de HSHS scoreregel op het Meten van Werkelijke Vaardigheid (Unpublished research report). University of Amsterdam, Netherlands.
13. Ratcliff, R. (1978). A theory of memory retrieval. Psychological Review, 85(2), 59-108.
14. Ratcliff, R. & Rouder, J. N. (1998). Modeling Response Times for Two-Choice Decisions. *Psychological Science,* 9(5), 347-356.
15. Thurstone, L. L. (1919). "A method for scoring tests." Psychological Bulletin 16: 235-240.
16. van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika, 72*, 287-308.
17. Van der Linden, W.J. & Hambleton, R.K. (Eds.) (1997). *Handbook of modern item response theory.* New York: Springer.
18. Van der Maas, H. L. J. & Wagenmakers, E.J. (2005). The Amsterdam Chess Test: a psychometric analysis of chess expertise. American Journal of Psychology, 118, 29-60.
19. Vandekerckhove, J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. *Psychonomic Bulletin & Review*, *14*, 1011-1026.
20. Wickelgren, W. A. (1977). Speed-accuracy tradeoff and information processing dynamics. Acta Psychologica, 41, 67-85.