Automatic Stepwise Correctness Assessment of Constructed Mathematical Response Schemes

[1]Nuru'l – 'Izzah Othman, [2]Zainab Abu Bakar, [3]Arsmah Ibrahim
[1,3]Centre of Mathematics Studies, [2]Centre of Computer Sciences Studies
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA
40450 Shah Alam, Selangor
Malaysia
[1]nurul@tmsk.uitm.edu.my, [2] zainab@tmsk.uitm.edu.my, [3]arsmah@tmsk.uitm.edu.my

ABSTRACT

Computer-aided free-response mathematics assessments are becoming increasingly important vehicles in supporting the notion of "learning through assessment". Hence, having a marking engine that is able to assess each step in a scheme of constructed responses (or working scheme) is a vital requirement for a mathematics computer-aided system that offers free-response assessments. Existing computational solutions for assessing the correctness of intermediate steps do so by limiting the responses to be entered stepwise, and not all at once, with the delivery of qualitative feedbacks to steps that are incorrect. In this paper, we propose our marking engine that is able to execute a stepwise assessment of responses in a working scheme that closely emulates the marking practices of a typical human. The marking engine delivers quantitative feedbacks in the form of numeric scores that indicate the degree of correctness of the responses in a working scheme and of the working scheme as well. The results of the reliability analysis on the automated correctness scores suggest that the scores are reliable indicators of responses correctness as well as working schemes correctness.

KEYWORDS

Mathematics computer-aided assessment, constructed responses, marking engine, stepwise correctness assessment, stepwise scoring

1 INTRODUCTION

Assessments can play a crucial role in achieving successful mathematics learning in terms of bridging the gap between the learning objectives that have been set and the actual realization of the objectives. "Assessment is the most powerful lever teachers have to influence the way students respond to courses and behave as learners" [1]. An assessment is regarded as "the primary driver of students' learning" [2] as it can facilitate in diagnosing learning difficulties and in pinpointing knowledge shortfalls [3]. Thus, computer-aided free-response mathematics assessments are becoming increasingly important vehicles in supporting this notion of "learning through assessment". A computer-aided free-response mathematics assessment refers to the type of assessment that requires students to construct their own solutions and enter those responses into a mathematics computer-aided assessment (MCAA) system. The correctness of the constructed responses entered by students is then automatically assessed by the automated marking component or the marking engine of the MCAA system.

Assessing the correctness of student-constructed responses is a central process in any free-response assessment thus making the marking engine a central component in any MCAA. Thus, research in the area of automatic assessment of free-response correctness has been carried out with much rigor and has resulted in many of the free-response MCAA systems that are available today. Our research interest in the area of free-response MCAA is centred on marking engine development which targets assessments that require the intermediate steps to be constructed and presented before coming to the final answer. Existing computational solutions for assessing the correctness of intermediate steps do so by limiting the responses to be entered stepwise, and not all at once, with the delivery of qualitative feedbacks to steps that are incorrect. Another approach is by breaking the problem into multiple parts. Hence, we confine our objective to achieving an automatic stepwise correctness checking and scoring of schemes of constructed responses, or working schemes in which the responses are entered all at once. By stepwise correctness checking and scoring we mean that the correctness of the responses in a working scheme is assessed step-by-step and a numeric score, $c \in [0,1]$, $c \in \Re$, is assigned to each step to indicate the degree of correctness of the constituting response. Based on the score of each step, an overall score that indicates the degree of correctness of the working scheme is calculated. The scores provide quantitative feedbacks that inform students to what extent their responses are correct.

In this paper, we describe the framework and the main capabilities of our marking engine in implementing stepwise correctness checking and scoring of schemes of constructed responses. The reliability of the correctness scores generated by the engine as indicators as response correctness is also presented. The responses are worked solutions to problems on solving linear algebraic equations in one variable. The SCCS abbreviation shall henceforth be used to refer to the phrase "stepwise correctness checking and scoring". Thus, our marking engine shall only be referred to as SCCS engine for easier reference. The construction of the computational methods for the SCCS engine is described elsewhere.

This paper is organised as follows. Section 2 describes some of the commendable solutions that are available for assessing equations correctness and the limitations that have been identified in existing solutions. Section 3 presents the framework of the SCCS engine which describes the approach taken by this research in solving the automatic stepwise correctness assessment problem. The section also describes the main capabilities of the SCCS engine which differentiates it from existing marking engines. Section 4 describes the analysis method and presents the results which are supplemented with some discussions on the results. This conclusion of the paper is in Section 5. In addition, the section describes the limitations of the approach that forms the basis of the SCCS engine and presents some future works in improving and enhancing the SCCS engine.

2 RELATED WORKS

Depending on the mathematical problem at hand, a constructed response henceforth shall only be referred to simply as response, can be a mathematical expression or an equation. Our intention currently is to provide computational solutions that will allow a marking engine to implement stepwise correctness checking and scoring of schemes of responses to problems on solving equations. To automate a marking process that emulates the customary practices of a typical human examiner is not a trivial task, especially in marking a working scheme to problems on solving equations. One of the challenges is how to manage the many variants of working schemes that can be formulated when solving an equation. These variants are the results of the multiple strategies that can be applied before the final answer can be reached.

MathXPert [4], Aplusix [5], [6] and T-Algebra [7] are some examples of MCAA systems that exhibit the capability of assessing the correctness of working schemes to problems in solving equations. These systems accomplish this task by allowing stepwise input of the intermediate steps and check the correctness after each entry. The error diagnosis of these systems depends on the way the expressions in the equations are manipulated and transformed. The manipulation and transformation of an expression in an equation are defined by a set of rewrite rules. These rules allow students to construct a response equation and enter the response in a stepwise manner as in a pen and paper situation. The correctness of the transformation is verified by determining the equivalence between two consecutive expressions [8]. These systems are commendable as they offer brilliant solutions for assessing the correctness of equations that are entered stepwise. However, these systems are designed to provide only qualitative feedbacks to steps that are incorrect with no consideration of a numeric score to indicate the degree of correctness of each step. In these systems only a single numeric score that indicates the overall correctness is awarded. The absence of a numeric score for each step, which serves as a quantitative feedback, may deprive students from knowing to what extent their answers are correct.

Judging by the published works in this area, it appears that many of the existing free-response MCAA systems tend to favour the method of establishing mathematical equivalence in assessing response correctness. The approach is also prevalent in MCAA systems that implement end-answer correctness assessment such as [9], [10], [11]. The marking engine of these systems is commonly underpinned by a computer algebra system (CAS). The employment of a CAS is due to the availability of a rich library of functions and enabling tools for performing symbolic manipulations. The symbolic computation capabilities allow the marking engine of CAS-based systems to establish the mathematical equivalence of the end-answer against a model solution in determining the correctness of the input.

3 THE SCCS ENGINE

Mathematical expressions and equations are mathematical notations with both content (and semantics) and structural (or syntactic) dimensions. The method of establishing mathematical equivalence, that is the content dimension, appears to be the widely used method for implementing automatic stepwise correctness checking. Nevertheless, we intend to forward our SCCS engine that is based on a simple approach of establishing the equivalence of the structural dimension. In this approach, a mathematical equation is treated as a literal string of symbols arranged in a linear structure.

3.1 The Framework

The SCCS engine is formally supported by Multiset Theory [12] and methods from the field of information retrieval. The construction of the computational techniques for the SCCS engine involves the following approaches. The correctness of a mathematical term in a response equation takes into account the correctness of its preceding "+" or "–" symbol. A term that is appended with a "+" or "–" symbol is referred to as a mathtoken. A formal definition of a mathtoken is formulated as follows. Let $\mathcal{A}$ be known as the mathematical alphabet (or vocabulary), which is the set of all mathematical symbols that encompasses $\varnothing$, numbers, variable symbols, operators and as such. Let the notation $\{u_i\}_{i=1}^{n}$ be used to denote a finite sequence of elements in a linear arrangement such that $\{u_i\}_{i=1}^{n} = u_1 u_2 ... u_n$, where $n \in \mathbf{Z}^+$.

Definition 1

A mathtoken $m = \{t_i\}_{i=1}^{k}, k \in \mathbf{Z}^+$ is a finite sequence of mathematical symbols in a linear arrangement such that $t \in \mathcal{A}$, where $t_1 \in \{+,-\}$ and the linear sequence $\{t_{i+1}\}_{i=1}^{k}$ represents a well-formed mathematical term.

In determining the correctness of a mathtoken, the naïve method of exact pattern matching has been adapted. The basis for applying exact pattern matching in determining the correctness of a mathtoken in a response equation is the premise that two structurally identical mathtokens are mathematically equivalent. In determining the correctness of the whole equation, the method of approximate string matching with *k*-mismatches has been applied. The implementation of an approximate matching between a response equation and the equation of an ideal solution requires an equation to be modeled as a multiset structure. In this paper, the notion of a mathstring is introduced.

Definition 2

A mathstring $M = [[m_i]]_{i=1}^{k}$, $k \in \mathbf{Z}^{+}$ is a collection of mathtokens where repetitions of mathtokens are allowed.

Hence, a mathstring is essentially a multiset of which the elements are mathtokens. The degree of correctness, d, of a response equation in comparison to its ideal solution is evaluated using our correctness measure which is formulated as:

$$\mathbf{C}(R,S) = \frac{\left|M_R \cap M_S\right|}{\max\left(\left|M_R\right|,\left|M_S\right|\right)} \qquad (1)$$

where M_R is the multiset of the mathstring associated with the response equation, R. M_S is the multiset of the mathstring associated with the ideal solution equation, S, and $\left|\cdot\right|$ is the multiset cardinality. The cardinality of the multisets intersection, $\left|M_R \cap M_S\right|$ is computed as

$$\left|M_R \cap M_S\right| = \sum_{i=1}^{p}\sum_{j=1}^{q} \mathcal{E}(r_i, s_j), \qquad (2)$$

where,

$$\mathcal{E}(r_i, s_j) = \begin{cases} 1 \text{ if } \forall\, a_{ik} \in r_i \text{ and } b_{ji} \in s_j,\ a_{ik} = b_{jt}, \\ 0 \text{ if } \exists\, a_{ik} \in r_i \text{ or } b_{ji} \in s_j,\ a_{ik} \neq b_{jt}, \end{cases} \qquad (3)$$

such that for any $1 \le h \le i$, $\mathcal{E}(r_h, s_j) \neq 1$. In other words, a mathtoken $s_j \in M_S$ has yet to be judged as a matching mathtoken to any mathtoken $r_h \in M_R$.

The accomplishment of an automatic SCCS process requires adapting the basic processes of textual information retrieval published in [13], [14]. The implementation of the SCCS process is illustrated in Figure 1. Figure 1 exemplifies the matching process of a collection of responses R consisting of m number of responses and a corresponding answer scheme S, consisting of n number of ideal solution equations.

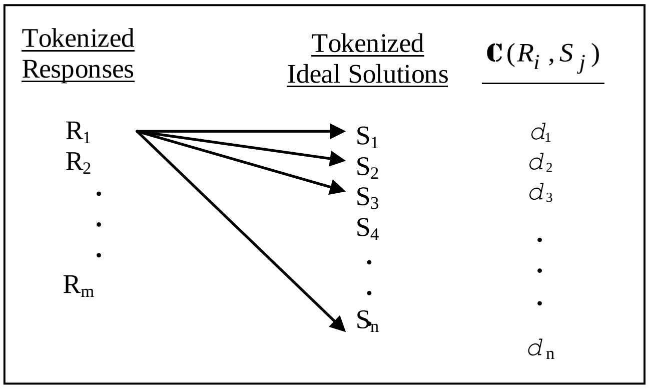


Figure 1. Stepwise Response-Solution Comparison and Evaluation Process

Note: All $R_i \in$ R are the student-constructed responses in a working scheme R, where $1 \le i \le m$. The score of each comparison is denoted by d_j for $1 \le j \le n$, where n is the number of equation in the solution scheme and $n > m$. If d_3 represents the degree of correctness of R_1, then the matching process for R_2 will start from S_4 and not from S_1.

For any $R_i \in$R the correctness score, c_i awarded is the maximum score of all the correctness scores, d_j, computed by the correctness measure which is given by

$$c_i = \max_{1 \le j \le n} (d_j), \text{ where } 1 \le i \le m. \qquad (4)$$

The correctness score of each step (or stepwise correctness score) which is basically the degree of correctness will inform students in which step a mistake has occurred, thus allowing them to trace the error. Given the correctness scores of each step, the average score is evaluated as follows:

$$\text{Average score} = \frac{\sum_{i=1}^{m} c_i}{D} \qquad (5)$$

where D is a value which is influenced by the number of invalid equations and equation duplicates present in a working scheme. Given the full mark, a value P, called the premark, is computed as

$$P = \text{Average score} * \text{full mark}. \qquad (6)$$

The assignment of an overall score or the final mark is restricted by the threshold value (to be defined by the teacher) and is computed as

$$\text{Overall score} = \begin{cases} \lfloor P \rfloor, \text{ if } \lfloor P \rfloor \le P < \lfloor P \rfloor + threshold \\ \lfloor P \rfloor + 0.5, \text{if } \lfloor P \rfloor + threshold \le P < \lceil P \rceil \\ \lceil P \rceil, \text{otherwise} \end{cases}. \quad (7)$$

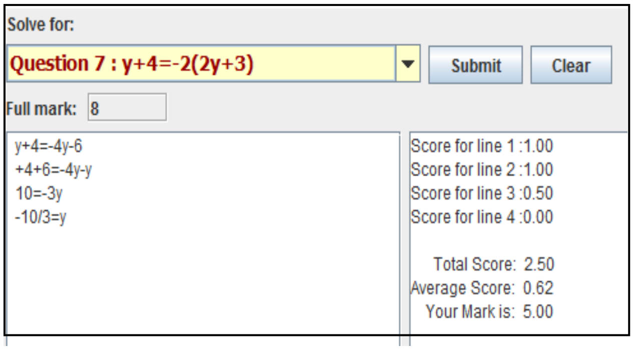
The overall score is presented as the mark for the working scheme that indicates the holistic correctness of the whole scheme. The threshold value may be interpreted as a restriction on the level of strictness in marking, or as the level of competency that is expected of a student. The construction of the SCCS computational techniques of the SCCS engine are discussed elsewhere.

3.2 SCCS Engine Capabilities

Despite our simple approach, our SCCS engine has succeeded in emulating much of the stepwise marking process normally performed by a human examiner with the execution of the following tasks:

i. Capable of allowing a working scheme of responses to be entered all at once and not via stepwise entry (Refer Figures 2 – 6).
ii. Implement a stepwise checking and scoring of mathematical responses that can provide quantitative feedbacks that inform students of the stepwise correctness of their responses and the holistic correctness of the working scheme. The assignment of stepwise correctness scores includes awarding a partial score for a partially correct response (Refer Figure 2). The assignment of overall scores for indicating holistic correctness includes awarding a partial overall score for a partially complete working scheme consisting of correct responses (Refer Figure 3).
iii. Able to identify duplicates in a working scheme (Refer Figures 3 and 4).
iv. Able to identify a response that is an invalid equation and to assign a zero correctness score to the response (Refer Figure 5).
v. Able to identify a complete working scheme (Figure 6) and an incomplete working scheme (Figures 2, 3, 4 and 5).
vi. Able to identify a response in standard final answer format and to discriminate between a response that truly forms the final answer (Figure 6) from one which is not (Figure 4) .

Samples of the results of the SCCS engine execution are shown in Figures 2 – 6.

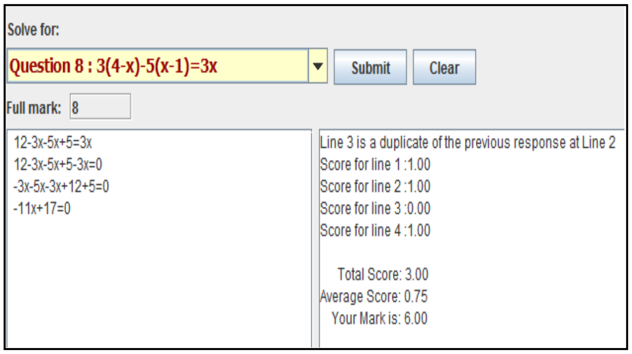


y+4=-4y-6
+4+6=-4y-y
10=-3y
-10/3=y

Score for line 1 :1.00
Score for line 2 :1.00
Score for line 3 :0.50
Score for line 4 :0.00

Total Score: 2.50
Average Score: 0.62
Your Mark is: 5.00

Figure 2. A sample of working scheme with a partially correct response in L3.

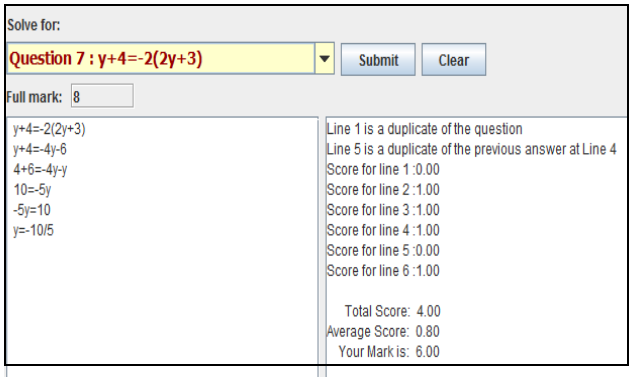


12-3x-5x+5=3x
12-3x-5x+5-3x=0
-3x-5x-3x+12+5=0
-11x+17=0

Line 3 is a duplicate of the previous response at Line 2
Score for line 1 :1.00
Score for line 2 :1.00
Score for line 3 :0.00
Score for line 4 :1.00

Total Score: 3.00
Average Score: 0.75
Your Mark is: 6.00

Figure 3. A sample of an incomplete working scheme with correct responses.



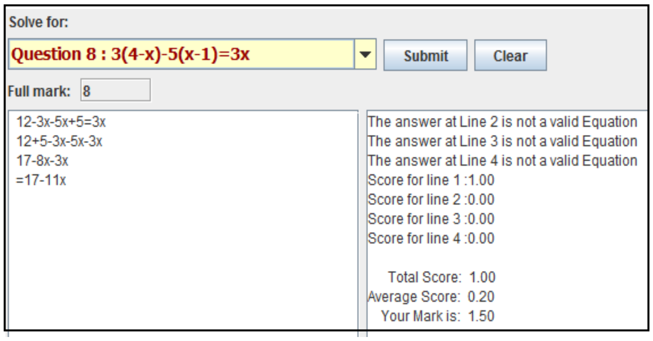
y+4=-2(2y+3)
y+4=-4y-6
4+6=-4y-y
10=-5y
-5y=10
y=-10/5

Line 1 is a duplicate of the question
Line 5 is a duplicate of the previous answer at Line 4
Score for line 1 :0.00
Score for line 2 :1.00
Score for line 3 :1.00
Score for line 4 :1.00
Score for line 5 :0.00
Score for line 6 :1.00

Total Score: 4.00
Average Score: 0.80
Your Mark is: 6.00

Figure 4. A sample of working scheme with duplicates.

Note: Notice that the engine is able to judge the response in L6 as not the final answer although it has the final answer format. Compare this working scheme with the scheme in Figure 6.



Solve for: Question 8 : 3(4-x)-5(x-1)=3x

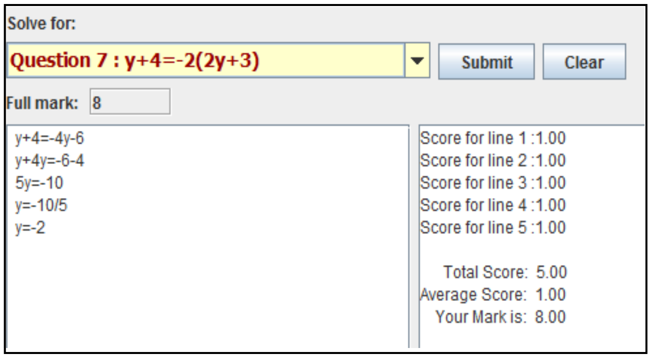
Full mark: 8

```
12-3x-5x+5=3x
12+5-3x-5x-3x
17-8x-3x
=17-11x
```

```
The answer at Line 2 is not a valid Equation
The answer at Line 3 is not a valid Equation
The answer at Line 4 is not a valid Equation
Score for line 1 :1.00
Score for line 2 :0.00
Score for line 3 :0.00
Score for line 4 :0.00

    Total Score:  1.00
Average Score:  0.20
   Your Mark is:  1.50
```

Figure 5. A sample of working scheme with invalid equations.



Solve for: Question 7 : y+4=-2(2y+3)

Full mark: 8

```
y+4=-4y-6
y+4y=-6-4
5y=-10
y=-10/5
y=-2
```

```
Score for line 1 :1.00
Score for line 2 :1.00
Score for line 3 :1.00
Score for line 4 :1.00
Score for line 5 :1.00

    Total Score:  5.00
Average Score:  1.00
   Your Mark is:  8.00
```

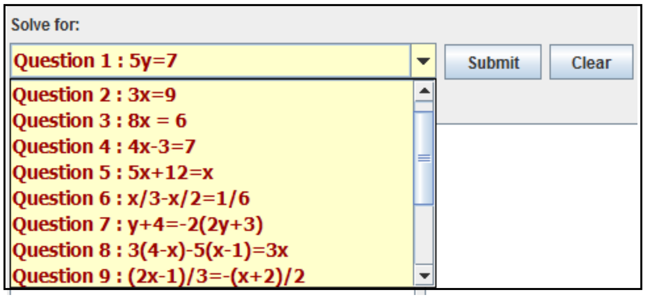
Figure 6. A sample of a totally correct scheme of responses.

4 RELIABILITY ANALYSES

The performance of the SCCS engine is established by analysing the reliability of the correctness scores generated by the engine.

4.1 Method

The reliability of the automated correctness scores as indicators of response correctness was established by determining the degree of agreement between the automated correctness scores and manual correctness scores obtained from manual marking process.



Question 1 : 5y=7
Question 2 : 3x=9
Question 3 : 8x = 6
Question 4 : 4x-3=7
Question 5 : 5x+12=x
Question 6 : x/3-x/2=1/6
Question 7 : y+4=-2(2y+3)
Question 8 : 3(4-x)-5(x-1)=3x
Question 9 : (2x-1)/3=-(x+2)/2

Figure 7. The set of questions used in the analysis.

Between 36 – 40 working schemes of responses for each the questions on solving linear algebraic questions in one variable shown in Figure 7 were collected using the SCCS engine prototype.

A total of 350 working schemes were collected with the number of responses totaling 1385. The correctness of the responses were automatically assessed by the marking engine and saved in the database of the prototype. The working schemes were then retrieved from the database, printed and distributed to the examiners to be manually marked. Following the recommendation in [15], three human examiners, each with a teaching experience of more than five years, were involved in the manual marking process. The 1385 automated stepwise scores (AS) of students' responses were compared against the average manual stepwise scores (AMS) of the examiners. 350 automated overall scores (AOS) were compared against the average manual overall scores (AMOS). The average manual stepwise scores and the average manual overall scores of the examiners are regarded as standards. In order to qualify the examiners' scores the merit of a standard, the degree of agreement between the examiners' scores for each question was first established.

The degree of agreement was first estimated using a correlation test. The Shapiro-Wilks results of the normality tests carried out on the scores prior to the tests on the degree of agreement showed that the distribution of each of the automated scores and the manual scores do not assume a normal distribution. Following the normality tests results, the Spearman's Rho Correlation Coefficient [16] was used. The test of correlation was carried out to provide a preliminary indicator of any degree of agreement where a non-positive correlation coefficient value will indicate whether or not agreement exists between the automated and manual scores. A positive coefficient indicates the strength of the correlation. Hence, some degree of agreement exists. A coefficient value of zero indicates no correlation, and hence no agreement exists between the (AS, AMS) and the (AOS, AMOS) sets of scores. The correlation was tested at a significance level of $\alpha = 0.01$ with the assumptions of the null and the alternative hypotheses stated as follows:

$$H_0 : \rho = 0.$$
$$H_1 : \rho \neq 0.$$

A value of $\rho = 0$ will indicate that there is no correlation between the (AS, AMS) or between the (AOS, AMOS) sets of data points, while $\rho \neq 0$ indicates otherwise and the null hypothesis will be rejected.

The degree of agreement of the (AS, AMS) and the (AOS, AMOS) sets of scores were determined using the Krippendorff's Alpha Reliability Index. Following the recommendations in [17] and [18], the standards of reliability for the Krippendorff's alpha, denoted here as α_K are stated as follows:

$$\alpha_K < 0.70 \quad : \text{Not Reliable}$$
$$0.70 \le \alpha_K < 0.80 \quad : \text{Reliable}$$
$$\alpha_K \ge 0.80 \quad : \text{Very Reliable}$$

The α_K computations were performed using the ReCal online software available at [19] which is explained in [20]. Finally, the absolute error between each (AS, AMS) data points and the (AOS, AMOS) data points was calculated to indicate the accuracy of the automated scoring.

4.2 Results and Discussions

The results of the analyses of the examiner stepwise scores and examiner overall scores using the Krippendorff's Alpha Reliability Index are recorded in Table 1.

Table 1. Examiners Analytical and Overall Scores Agreement Analyses

Qtn	No. of WS	No. of responses	ESS agreement		EOS agreement	
			α_K	Decision	α_K	Decision
1	40	84	0.912	Very reliable	1.000	Very reliable
2	40	113	0.979	Very reliable	0.959	Very reliable
3	40	122	0.966	Very reliable	0.978	Very reliable
4	40	165	0.971	Very reliable	0.992	Very reliable
5	40	175	0.984	Very reliable	0.987	Very reliable
6	36	164	0.920	Very reliable	0.923	Very reliable
7	39	188	0.957	Very reliable	0.997	Very reliable
8	39	189	0.985	Very reliable	0.987	Very reliable
9	36	185	0.924	Very reliable	0.996	Very reliable

Key: ESS: Examiner Stepwise Score EOS: Examiner Overall Score

The results in Table 1 show that the examiners correctness scores for each question display a high degree of agreement. The average reliability coefficient of the stepwise scores $\alpha_{K_{ave}} = 0.955$, while the average reliability coefficient for the overall scores $\alpha_{K_{ave}} = 0.980$. The $\alpha_{K_{ave}}$ values give a clear indication that the examiners' scores are reliable and hence can be used as true values to benchmark the scoring of responses and working schemes of this research.

The results of the agreement analysis are recorded in Table 2. They show that the automated correctness scores display a strong positive correlation with the manual correctness scores.

Table 2. Results of the Score Agreement Analysis

	(AS, AMS)	$(AOS, AMOS)$
Total number of WS	350	350
Total number of responses	1385	-
Correlation Analysis	0.966 - 0.996	0.977 - 0.999
	0.983	0.990
	Reject H_0	Reject H_0
Degree of Agreement Analysis	0.835 - 0.966	0.740 - 0.991
	0.899	0.893
	Very reliable	Very reliable
Average Absolute Error	0.02	0.1

Note: The maximum score for the stepwise score is 1.00, while the maximum score for the overall score follows the full mark allocated for a question which ranges from 2 to 8.

The average value of the correlation coefficient calculates to $\rho = 0.983$ for the (AS, AMS) datasets and $\rho = 0.990$ for (AOS, AMOS) datasets. The results in Table 2 also show that the automated correctness scores exhibit high degree of agreement with the manual correctness scores for the (AS, AMS) datasets as well as the (AOS, AMOS) datasets. The α_K values for the (AS, AMS) datasets range between 0.835 – 0.966 giving the average of $\alpha_{K_{ave}} = 0.899$. The α_K values for the (AOS, AMOS) datasets range between 0.740 – 0.991 which calculates to an average of $\alpha_{K_{ave}} = 0.893$. The average absolute error for the automated stepwise correctness scores is 0.02. For the automated overall correctness scores, the average absolute error is 0.1. These values simply mean that on an average the SCCS engine has the capability to award automated stepwise correctness scores that will differ by an error of ±0.02 from the stepwise correctness scores awarded by a human examiner, while the automated overall scores will differ by an error of ±0.1 from the manual overall correctness scores.

5 CONCLUSIONS

Given the results that have been obtained, we can conclude that the automated stepwise correctness scores and the automated overall scores are very reliable indicators of responses and of working schemes correctness. This conclusion permits us to further conclude that the SCCS process that has been based on a simple string matching approach is reliable. Furthermore, the correctness scores produced by the SCCS engine for the questions used in this research are comparable to the correctness scores awarded by human examiners. The SCCS engine is able to assess response correctness with an accuracy of ±0.02 and the correctness of a whole working scheme with an accuracy of ±0.1. Despite the promising results, there still exist some limitations exhibited by the SCCS engine due to the approach that has been taken. The most apparent is the requirement of sufficiently many variants of the ideal working schemes for a given question, which proved to be quite tedious to achieve. Hence, a plan to integrate rule-based techniques into our solution, but still maintaining the string matching approach in determining response correctness, will be considered in our future research. For our on-going works we are currently testing the engine SCCS capability on equations involving logarithms and exponents.

6 ACKNOWLEDGMENT

This work is financially supported by the RIF grant from the Universiti Teknologi MARA, Malaysia.

7 REFERENCES

1. Gibbs, G.: Using assessment strategically to change the way students learn. In: Brown, S., Glasner, A. (eds.) Assessment Matters in Higher Education: Choosing and Using Diverse Approaches, pp. 41 – 53. Buckingham, SHRE and Open University Press (1999).
2. Sangwin, C. J.: What is a mathematical question? In: Proc. 2007 1st JEM Workshop ePlusCalculus. Lisboa, Portugal (2007). http://web.mat.bham.ac.uk/C.J.Sangwin/Publications/2007JEMLisbonSangwin.pdf
3. Saleh Al-shomrani & Wang, P.: Building DMAD: A distributed mathematics assessment database for WME. In: Proc. 2005 IEEE Southeast Conference, pp. 630 – 635. 8 – 10 April 2005, Ft. Lauderdale, Florida.
4. Beeson, M.: MathXpert:Learning Mathematics in the 21st Century, vol. 9, no. 1 – 2, (2002). (Translated from: MathXpert: un logiciel pour aider les élèves à apprendre les mathématiques par l'action, Sciences et Techniques Educatives). http://www.michaelbeeson.com/research/papers/English-ste/English-ste.html.
5. Chaachoua, H., Nicaud, J. F., Bronner, A., and Bouhineau, D.: Aplusix, a learning environment for algebra, actual use and benefits. In: Proc. 2004 10th International Congress on Mathematical Education, Denmark (2004). http://hal.archives-ouvertes.fr/hal-00190393/en/.
6. Trgalová, J., Chaachoua, H,: Development of Aplusix software. In: Proc. 2008 11th International Congress on Mathematics Education (ICME 11), Monterrey, Mexico, 6-13, July 2008. http://tsg.icme11.org/tsg/show/23.
7. Prank, R., Lepp, M., Lepp, D., Vaiksaar, V., Tõnisson, E.: T-algebra - intelligent environment for expression manipulation exercises. In: Proc. 2008 11th International Congress on Mathematics Education (ICME 11), Monterrey, Mexico, 6-13, July 2008. http://tsg.icme11.org/document/get/253.
8. Issakova, M.: Comparison of student errors made during linear equation solving on paper and in interactive learning environment. In: Proc. 2005 7th International

Conference on Technology in Mathematics Teaching (ICTMT 7), vol. 1, pp. 250 – 258, Bristol, U.K (2005).

9. Sangwin, C. J.: Assessing Elementary Algebra with STACK. International Journal of Mathematical Education in Science and Technology, 38(8), pp. 987 - 1002 (December 2008).
10. Melis, E., Siekmann, J. H.: ACTIVEMATH: An Intelligent Tutoring System for mathematics. Lecture Notes in Computer Science, vol. 3070, pp. 91 – 101. Springer Berlin/Heidelberg (2004).
11. Sangwin, C. J.: Assessing higher mathematical skills using computer algebra marking through AIM. In: Proc. 2003 Engineering Mathematics and Applications Conference (EMAC), pp. 229 – 234.
12. Blizard, W. D.: Multiset Theory. Notre Dame Journal of Formal Logic, vol. 30, no. 1, pp. 36 – 66, (1989). http://projecteuclid.org/DPubS/Repository/1.0/Disseminate?view=body&id=pdf_1&handle=euclid.ndjfl/1093634995.
13. Salton, G., McGill, M. J.: Introduction to modern information retrieval. McGraw-Hill, New York (1983).
14. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley Longman Publishing Co. Inc. (1999)..
15. Livne, N. L., Livne, O. E., Wright, C. A.: Can automated scoring surpass hand grading of students' constructed responses and error patterns in mathematics? MERLOT Journal of Online Learning and Teaching, vol. 3, no. 3, pp. 295 – 306, (2007). http://jolt.merlot.org/vol3no3/livne.pdf.
16. Wackerly, D. D., Mendenhall, W., Scheaffer, R. L.: Mathematical Statistics with Applications (Seventh edition). Belmont, CA, USA: Brooks/Cole Cengage Learning (2008).
17. Lombard, M., Snyder-Duch, J., Bracken, C. C.: Practical resources for assessing and reporting intercoder reliability in content analysis research projects, (Online edition), (2010). http://matthewlombard.com /reliability/.
18. Krippendorff, K.: Content Analysis: An Introduction to Its Methodology. California: Thousand Oaks (2004).
19. Dfreelon.org.: ReCal for Ordinal, Interval, and Ratio Data (OIR) (Web service). http://dfreelon.org/utils/recalfront/recal-oir/.
20. Freelon, D. G.: ReCal OIR: Ordinary, Interval and Ratio Intercoder reliability as a Web service. International Journal of Internet Science, vol. 8, no. 1, pp. 10 – 16. ISSN: 1662-5544 (2013)..