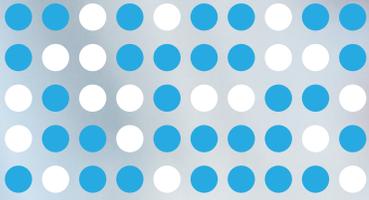


ASAP



PRICING STUDY Machine Scoring of Student Essays

Barry Topol, John
Olson, and Ed
Roeber

Assessment
Solutions Group

Published by:

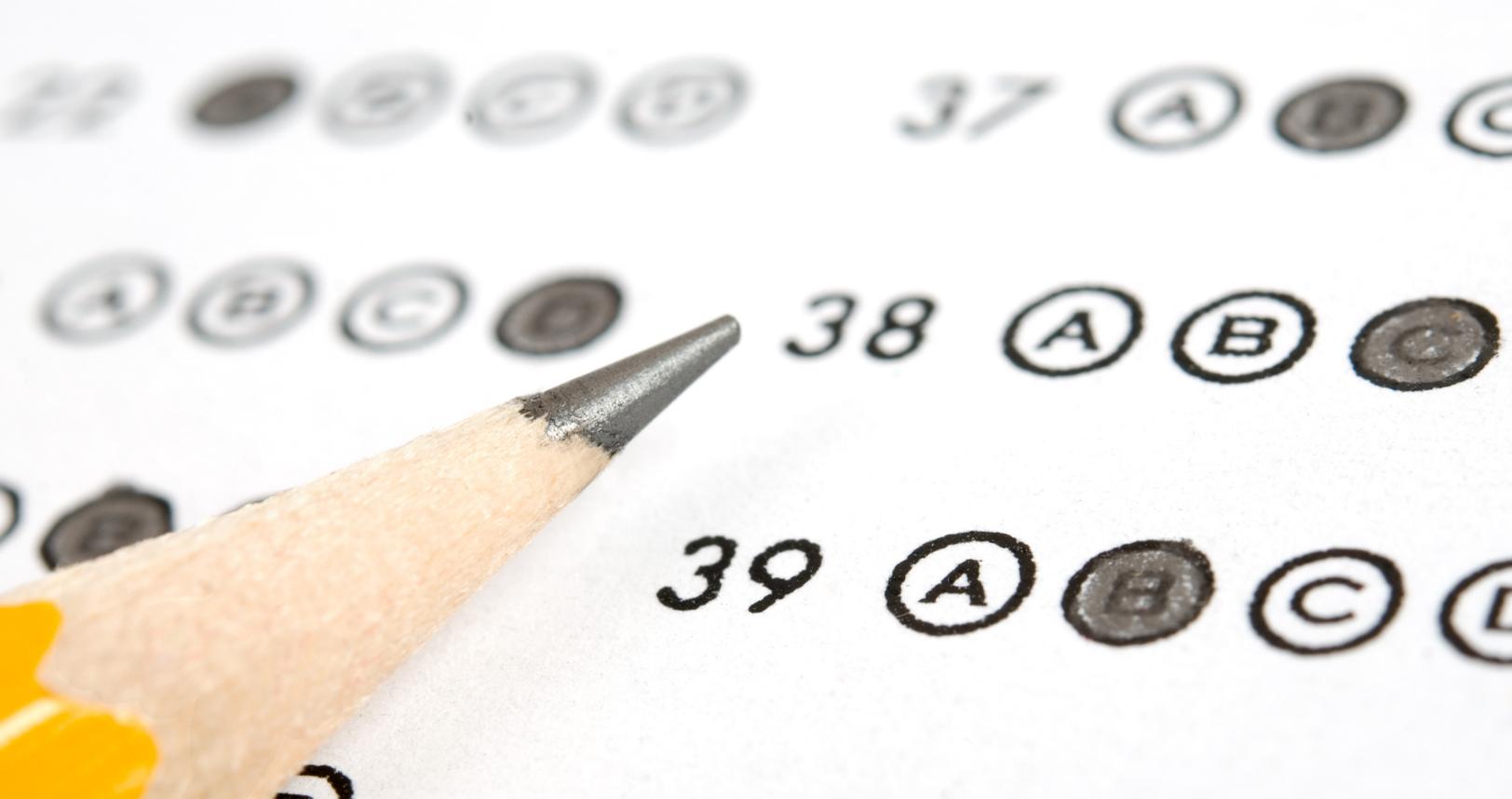


February
2014

TABLE OF CONTENTS

I. Introduction.....	1
II. Purpose Of The Study.....	3
III. Methodology / Process.....	5
IV. Key Cost Elements In Implementing Machine Scoring.....	8
V. Key Assumptions And Inputs For Inclusion In RFPs (Vendor Provided Information)...	19
VI. Additional Information For Inclusion In RFPs (Non-Vendor Provided Information)....	25
VII. Pricing Data.....	28
VIII. Pricing Expectations.....	33
IX. Final Conclusions And Recommendations.....	36
References	38
Acknowledgements	38
Appendix A.....	39
Appendix B	40





EXECUTIVE SUMMARY

Education experts agree that the next generation of assessments (such as those being developed by the Partnership for the Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (SBAC) in response to the new Common Core State Standards (CCSS)) need to do a better job of measuring deeper learning to determine if students are acquiring those skills critical to success in the 21st century. Existing assessments tend to emphasize “bubble in” multiple choice type questions because they are easier, more timely and cheaper to score. However, multiple choice questions do not provide as good a measure of critical thinking skills as performance type questions, in which students are asked to read a passage or passages and present an argument based on synthesizing the information they have read. The answers to these performance type questions tend to be scored by humans, which is a time intensive and expensive process. While some discussion about finding ways to increase the amount of money spent on state assessment systems overall has begun, at least for the near future, states only appear to be able to spend roughly what they spend today for new summative assessments. Therefore, the question is, can the next generation of assessments be designed to better measure student critical thinking skills while costing roughly the same amount as states spend today (about \$25 per student)?

The authors have previously published papers providing guidance to procurers of assessments on how to develop a higher quality assessment—one with more performance type questions, 50 percent fewer multiple choice type questions, and competitive in price with current assessments. Strategies for doing so include forming a state assessment consortium to take advantages of economies of scale in test development and other functions, using teachers to score open-ended questions and using technology in the delivery and scoring of assessments. This paper focuses on one aspect of the strategies, machine scoring of open-ended questions.

In a higher quality assessment, human scoring of student essays can comprise over 60 percent of the total cost of the assessment. Automating this scoring process holds tremendous promise in making higher quality assessments affordable. To determine the feasibility of using machine scoring engines to score open-ended responses, the Hewlett Foundation sponsored the Automated Student Assessment Prize (ASAP) contest(s) in 2012-2013. The results were that existing machine scoring engines are generally capable of producing scores similar, in the aggregate, to the scores of human raters for long form English Language Arts (ELA) essays (defined as essays with more than 150 word responses). Machine scoring engines were not as reliable as humans in scoring short form ELA essay responses (essays of less than 150 words).

The purposes of this paper are to determine the relative cost of machine scoring engines versus human scoring for long form ELA essays and to provide purchasers of machine scoring services guidance in structuring their requests for proposals (RFPs).

The authors gathered data from the major machine scoring vendors via interviews and a request for information (RFI), as well as examined responses to a recently issued RFP from the state of Michigan to implement the SBAC assessment in the state, including machine scoring of student essays. Using the information gathered, the authors present the key fixed and variable cost elements related to machine scoring and provide a range of the potential costs of machine scoring services relative to human scoring for long form ELA essays.

Machine scoring of student essays proved to be significantly less expensive than human scoring. Depending on the set of assumptions used, machine scoring of long form ELA essays can be as low as 20 percent to 50 percent of the cost of human scoring given favorable conditions and major volumes of student responses (5 million). In the case of a significantly sized mini-consortium of states (1.5 million to 3 million students), the cost of machine scoring was estimated at 25 percent to 55 percent of the cost of human scoring. A large state (750,000 to 1.5 million students) could see costs for machine scoring of essays that is 25 percent to 60 percent of human scoring costs, and a smaller state (up to 500,000 students) could see costs for machine scoring of 30 percent to 80 percent of human scoring. While it is still too early to tell if machine scoring engines will be capable of scoring the type of performance items being developed by PARCC and SBAC, assessment costs can be significantly reduced if the engines can score these items.

Because certain types of items that require different types of essay responses can be efficiently scored with existing machine scoring technology, while others cannot, it is recommended that procurers of machine scoring services work jointly with the vendor community to develop the type of items that can both assess students' deeper learning skills and be efficiently scored by current vendor machine scoring engines. This is an area where a partnership between those designing new assessments and the vendor community holds great promise.



INTRODUCTION



It is widely accepted that the summative assessments currently used to evaluate student achievement in the United States need to be significantly improved. The majority of the assessments used in states today consist largely of multiple choice or “bubble-in” questions, which do a poor job of measuring the deeper thinking skills students need to master in order to be successful in college and careers. Since 2001, the increased demands brought about by No Child Left Behind (NCLB) and tight state budgets have led most states away from performance assessments that measure critical thinking skills by asking students to perform complex tasks or projects, and more toward multiple choice assessments that primarily measure student retention of high- and low-level facts. In part, the multiple choice format has become so pervasive because it is relatively inexpensive to administer and easy to score. Any attempts to identify more effective alternatives must compete with those cost and operational considerations.

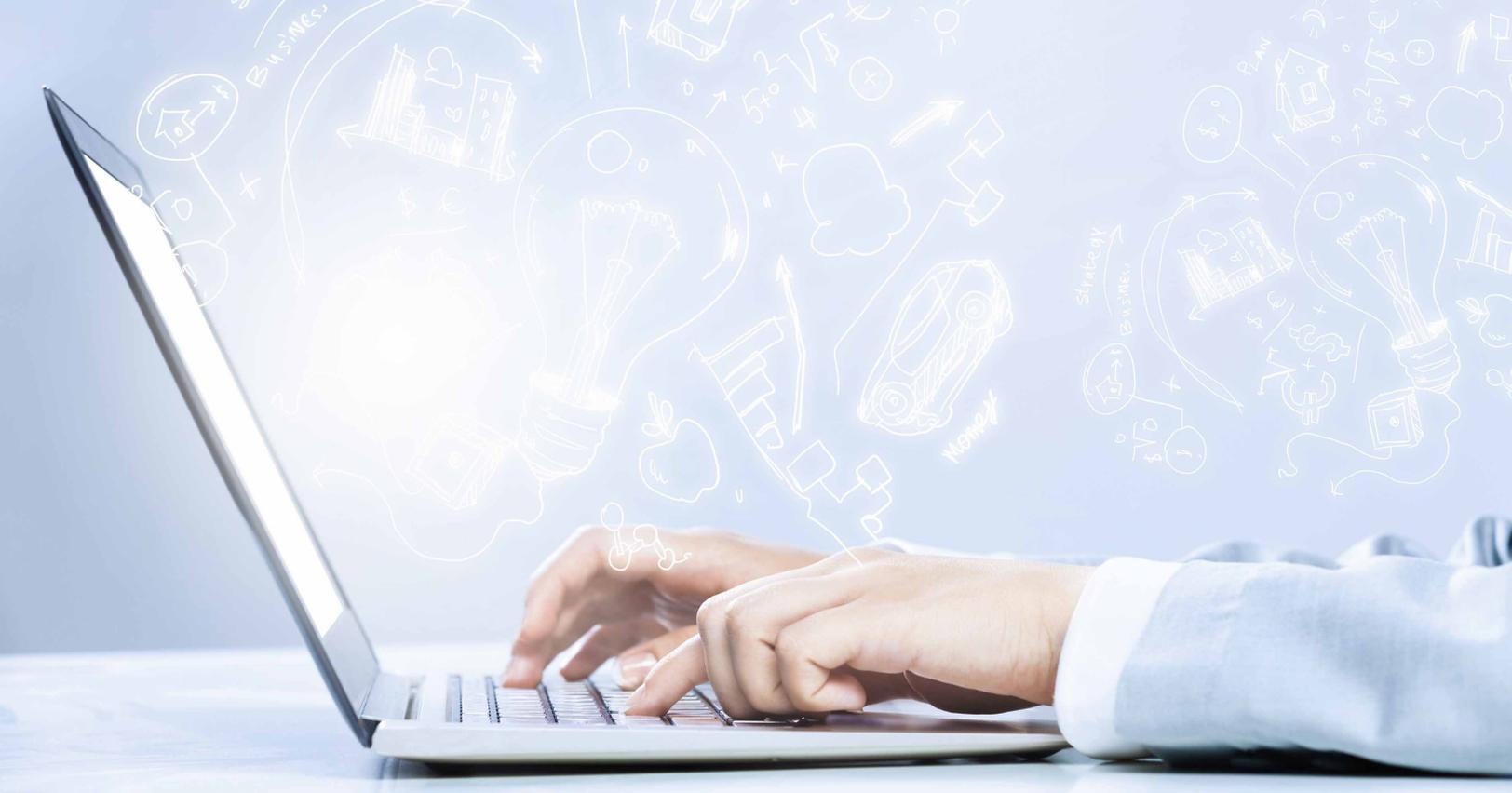
States and the education community at large recognize the need to improve the education standards upon which the tests that measure student achievement are based. In 2009-2010, the National Governors Association (NGA) and the Council of Chief State School Officers (CCSSO) undertook an ambitious project to develop new education standards linked to those skills required for success in the 21st century. The new standards, referred to as the Common Core State Standards (CCSS), were designed to measure deeper learning and were benchmarked against those of top performing nations on international assessments of student achievement. To date, 45 states have adopted the standards. However, as those states consider new approaches to student assessment, the most pressing concern continues to focus on the relative cost of the alternatives.

In 2010, two state student assessment consortia, the Partnership for the Assessment of Readiness for College and Careers (PARCC) and the Smarter Balanced Assessment Consortium (SBAC), representing 44 states¹ formed to develop the next generation of common assessments based on the CCSS. These new assessments, to be administered for the first time in 2014-2015, promise to result in a significant improvement in the ability to measure those skills students will most need to be successful in the 21st century. Also, by pooling the resources of multiple states, the consortia are attempting to lower the assessment cost per student.

In the past, and in most current assessments, the most effective measure of student critical thinking skills comes through questions that generally require student written responses, which are scored by human readers. This is a time intensive and expensive endeavor, often costing \$1 to \$2 per student and per question scored, taking several weeks to process all of the student responses. In 2010, the Assessment Solutions Group (ASG) estimated that the cost of a higher quality assessment, in which multiple choice questions would be reduced by half and replaced with fewer but higher quality essay questions, would cost a single state \$56 per student to implement if those responses were graded by hand (Topol, Olson, & Roeber (2010), *The Cost of New Higher Quality Assessments: A Comprehensive Analysis of the Potential Costs for Future State Assessments*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.). This compares to the roughly \$20 to \$25 per student that states currently spend (as of 2010) on assessments, which rely heavily on multiple choice to assess student achievement in math and English Language Arts (ELA) (Ibid.) A follow-up survey (unpublished) of 42 states that ASG recently completed shows states spending \$25 per student for their NCLB math, reading and writing assessments.

This leads to a fundamental research question: Can a better test be delivered, around the current cost of \$25 per student, that measures deeper learning, is as efficient to deliver as a “bubble test,” and can pass the scrutiny of such a wide range of state interests? The most prevalent cost reduction techniques include pooling resources through the state assessment consortia, using other technologies to deliver the assessments, using teachers to score some of the essay type questions, and using machine scoring engines to grade the remaining student essay questions. Both PARCC and SBAC are considering whether to employ these techniques in developing their new assessments and over time should be able to deliver the new assessments at close to the \$25 per student range.

[1] Some PARCC states and Pennsylvania (PARCC and SBAC) have switched membership status from governing state to member state and will not take the assessments in 2014-2015.



II PURPOSE OF THE STUDY

As noted above, one critical assumption in the affordability of the next generation of assessments will be the ability of machine scoring engines to accurately grade student essay responses and to do so at a reasonable cost. To determine the feasibility of both of these issues, the William and Flora Hewlett Foundation funded the Automated Student Assessment Prize (ASAP) in 2012-2013 and this pricing study.

In ASAP Phase One, the ability of machine scoring engines to accurately score long form essay responses (>150 words) as well as humans was examined. The responses came from six participating state departments of education and encompassed writing assessment items from three grade levels: 7, 8 and 10. The items were evenly divided between source-based prompts (that is, essay prompts developed on the basis of provided source material) or those drawn from traditional writing genre (narrative, descriptive, persuasive). Nine different engines were evaluated and the results of the study were that existing machine scoring engines are generally capable of producing scores similar, in the aggregate, to the scores of human raters for long form ELA essays (Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 313-346). New York, NY: Routledge.).

In ASAP Phase Two, the ability of machine scoring engines to score short form essays (<150 words) was studied. The essays came from three participating state departments of education and encompassed assessment items from two grade levels: 8 and 10. On average, each answer was approximately 50 words in length. Some responses were more dependent upon source materials than others, and the answers covered a broad range of disciplines (from ELA to science). 256 engines competed in the study, and the results indicated that the machine scoring engines could be used in some applications to score short form essays but failed on most measures to compete with hand scoring. In fact, while the findings of ASAP Phase Two showed some promise across some testing item types, by and large the study called for a deeper investigation before machine scoring of short form responses should be applied in a high stakes environment (Shermis, M. D. (2013). *Contrasting state-of-the-art in the machine scoring of short-form constructed responses*. Paper presented at the National Council on Measurement in Education, San Francisco, CA.).

Having demonstrated the current capabilities to grade certain types of student written responses using machine scoring technology, the next step in the process was to measure the cost of scoring the most effective application (i.e., student long form essay responses) of machine scoring and to compare the results to the cost of scoring similar questions with human scorers. Should the cost of scoring essay items with machine scoring be significantly lower than the cost of scoring similar items with humans, higher quality tests could then be implemented at a lower, more affordable price. The goal of this phase of the project was to determine and understand the important pricing elements inherent in machine scoring, understand the factors that cause variations in machine scoring costs, predict machine scoring costs for the item types being developed by PARCC and SBAC, and develop recommendations to states and consortia of states looking to procure and implement machine scoring for their assessment programs. Through this process, the options presented to states could include a range of applications or cost savings, from utilizing machine scoring as the primary grading system to relying on machine scoring for “read-behind” or cost-effective validation of hand scoring. In any case, the purpose was to provide states with alternatives that would include the cost implications to adopt them.



METHODOLOGY / PROCESS



In late 2012, a team of researchers working on this project decided that an open call to the vendor community would be issued in the form of a Request for Information (RFI) to compile the factors involved in assessing the cost of machine scoring. The educational assessment machine scoring community is rather small with eight vendors accounting for the dominant majority of the market (Shermis, Hamner 2013). The vendors are (in alphabetical order) American Institutes for Research (AIR), CTB/McGraw Hill, Educational Testing Service (ETS), Measurement Incorporated, MetaMetrics, Pacific Metrics, Pearson Education and Vantage Learning.

While machine scoring has been in existence for many years, its use in scoring student responses in high stakes assessments in the United States has been quite limited. Only a handful of states currently use or plan to use machine scoring to score student responses for their high stakes NCLB tests prior to the consortia tests becoming available in 2014-2015. Consequently, most states are not well educated in procuring machine scoring services, and a common complaint of vendors in the community is the lack of clarity and specific detail included in the few RFPs for machine scoring services that are issued. Therefore, it was decided that the vendor community should have significant input as to the content of the RFI. It was also hoped that including vendor input in drafting the RFI and promising vendors anonymity in their responses, particularly those related to costs, would lead to a higher response rate and better quality responses from the vendors.

Prior to writing the RFI, a survey was developed (see Appendix A) to identify the key information that vendors would need to understand in order to estimate pricing for machine scoring services. The questionnaire was sent to all eight vendors and interviews were held with seven of the eight companies to record their responses to the survey questions and gather other information that the vendors identified during the interviews as important factors toward being able to provide pricing estimates for machine scoring services. Vendors were free and open with their input and the RFI was largely based on the content gathered from them. Additionally, some of the information presented in this paper has been taken from vendor responses to the RFI.

The RFI was comprised of several sections:

- Introduction and overview of ASAP and the importance of the RFI
- General instructions on responding to the RFI
- General company information, experience and capabilities in machine scoring
- General assumptions and item type assumption information
- Description of PARCC assessment item types for machine scoring
- Description of SBAC assessment item types for machine scoring
- RFI response request detail

Important assumptions were provided in the RFI so that vendors could work from a common set of parameters regarding the nature of the items to be scored, nature and security requirements of the assessment itself, technology used to deliver the assessment, machine scoring system interface needs, scoring turnaround time expectations, student and item counts for both the PARCC and SBAC assessment consortia, agreement rate with human scorers required, item field test assumptions/number of readily available scored papers, etc.

Both the PARCC and SBAC assessment consortia were supportive of the effort to develop the RFI and each provided letters of support that were included in the final document released to the vendor community (see Appendix C). Each assessment consortium assigned a representative to work with the research team to provide important information about the items they were developing, so that vendors could better prepare their cost estimates. Descriptions and rubrics were provided for each different item type that was expected to appear on the common assessment in which a pricing estimate was being solicited. The item descriptions were reviewed and approved by each consortium. Note that the RFI itself was not reviewed by either assessment consortium. (To view the final RFI go to: http://www.assessmentgroup.org/uploads/Machine_Scoring_RFI-3.12.13.pdf)

Four of the seven vendors that the team spoke with responded to the RFI. These vendors submitted significant qualitative and quantitative information in their responses, which the research team has reflected in this paper. Vendor pricing submissions covered a wide range of costs and some differing assumptions, so the research team decided to augment the vendor pricing data submitted in response to the RFI with an excellent real world example.

The state of Michigan issued an RFP in late 2012 to implement the SBAC assessment system in the state for the 2014-15 administration year. (Note that the SBAC plan is to have the consortium manage the content development, psychometrics and other centralized activities while the member states are responsible for the actual implementation and delivery of the assessment in their states). The research team gathered the vendor responses to the RFP and used that pricing information, as well as the information vendors submitted in response to the RFI in the analysis.

The following sections of this report include, based on the information submitted by the vendors and additional data gathered from the Michigan RFP responses, the factors to consider in implementing machine scoring, the areas impacting costs, general cost data, and recommendations for implementation of machine scoring.



IV

KEY COST ELEMENTS IN IMPLEMENTING MACHINE SCORING

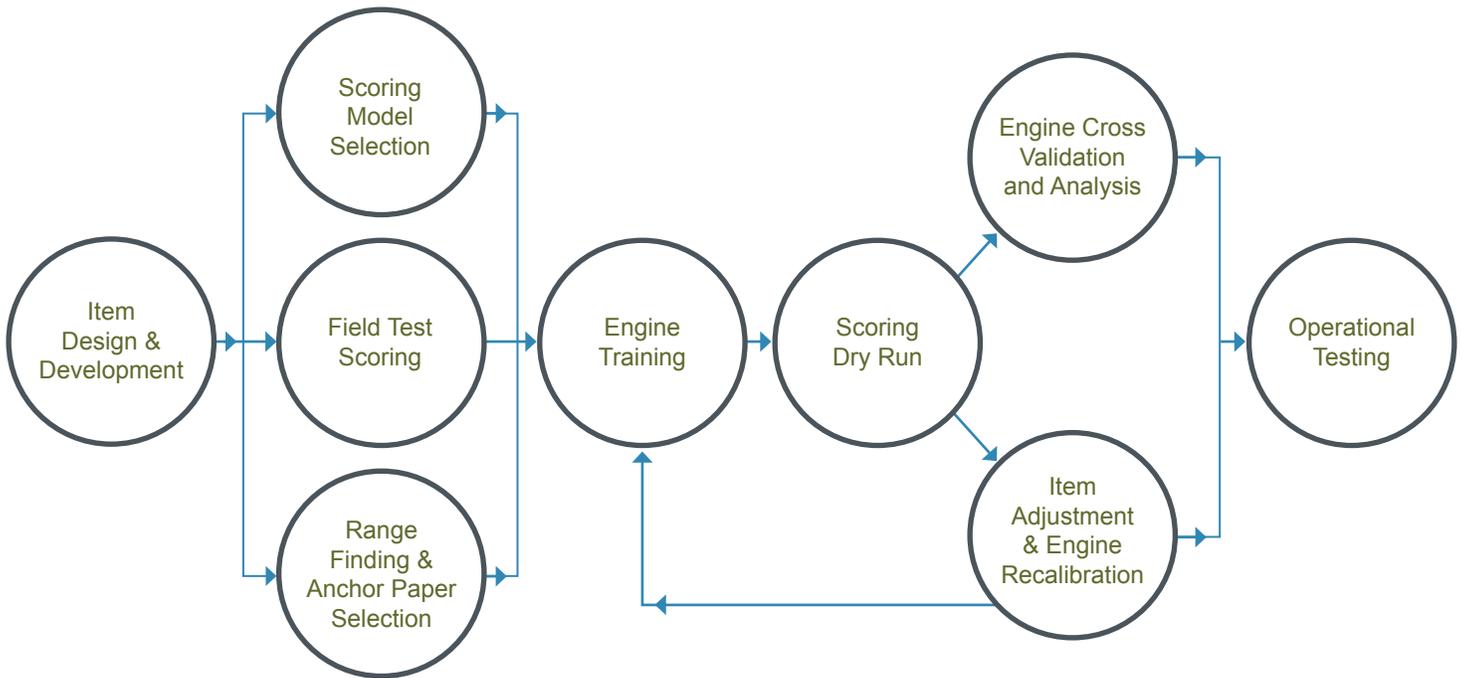
This section of the paper presents the key cost elements in implementing machine scoring, a definition of the cost elements, and factors influencing cost variability related to the particular cost elements.

There can be many different types of functions, and therefore costs, involved in the implementation of a machine scoring system. Not all of the categories of cost information presented below will be incurred for any particular implementation, and quite often some of the costs for the services described below are contained within other, more primary, cost elements. For the study, the cost elements were divided into two categories:

- Per response and per item fees, which are 100 percent variable and tied to either the item being evaluated or the student response to the item itself; and,
- General system implementation cost elements, which are related not to the scoring of any particular item or response but rather to the fidelity of the scoring implementation as a whole. General system implementation costs can be quoted separately or included as part of the per response and per item fees.

Prior to engaging in a discussion of the key cost elements related to machine scoring, it is useful to provide a brief overview of the process of machine scoring. This will provide valuable context for the discussion to follow. The diagram below outlines the primary elements in a typical machine scoring system:

EXHIBIT 1: MACHINE SCORING PROCESS OVERVIEW



MACHINE SCORING PROCESS OVERVIEW

- A. **Item design and development** –The first step in the scoring process for any type of open-ended or constructed-response item, whether it is to be scored by a machine or humans, is to develop the item itself. The vendors recommend that item developers work with their machine scoring staffs in order to create the types of items that are able both to elicit student critical thinking skills AND be readily scored with existing machine scoring systems. This is an important step in the process and one that is often overlooked.
- B. **Scoring Model Selection** – Once the item(s) have been developed, the machine scoring vendor will determine which of its existing scoring models is best suited to score the student responses. Many vendors have several different models they use to score item responses, depending on the nature of the item and type of response that is called for. Vendors may also adjust or modify their engines to enable them to deliver more accurate scores for a given item.

- C. **Field test/scoring** – The item(s) are then field tested. The field testing is used not only to determine if the items are functioning properly for assessment purposes but to also generate enough student responses, at each score point and trait, to “train” the scoring engine in how to render the appropriate scores. Generally speaking, it takes 100 to 200 papers at each score point to train the system. Generating the appropriate number of papers to train the system to score a five or six point high stakes assessment item can require 1,500 to 2,500 total responses in order to get enough papers at both the high and low ends of the scoring rubric.
- D. **Range finding and anchor paper selection and engine training** – Range finding and anchor paper selection is done concurrently with machine scoring engine training, and, therefore, enough papers must be selected and evaluated to train and validate the machine scoring engine. These papers are also used to train human scorers to score the items if humans are going to be used either to audit machine scored papers or provide primary scores of student responses (where machine scoring serves as the “read-behind” feature to validate hand scoring).
- E. **Model Formation** – Prior to the operational assessment, the scoring model is formed to determine if the machine scoring engine is scoring papers appropriately.
- F. **Engine cross validation and analysis** – Once these new papers have been scored, psychometricians and other specialists will determine if the machine scoring algorithm is working properly. Usually the “test” data is split into two parts—a test file and a cross-validation file. The prediction formulas (usually based on multiple regression) constructed from the training set are first applied to the test file and reliability coefficients are calculated. Multiple regression prediction formulas based on one set of data are not as accurate as those based on another set of data drawn from the same population because of the least squares nature of the formulas, so the prediction equations are applied to the cross-validation set and reliability coefficients are re-calculated.
- G. **Item adjustment and engine recalibration** – Once the engine cross validation and analysis work has been completed, the machine scoring analysts may recalibrate the scoring algorithm to produce better results.
- H. **Operational testing** – Once the engine has been trained, a sample of responses properly scored and the scoring algorithm properly calibrated, the user is ready to use the system in a live assessment environment.

With the process review description completed, the various cost categories and cost elements related to machine scoring are now presented. Procurers of machine scoring services should ask vendors to provide quotes for these services.

A. **Per Item/Response Fees** – Per item/response fees represent the major costs associated with machine scoring and are generally included in all implementations. The major per item/response fees are presented below.

- a) *Engine training fee* – As described above, the engine training fee is the cost associated with calibrating the machine scoring engine to score student responses accurately at all score points for all traits. This may include engine recalibration after some initial testing is completed if the engine is not meeting benchmarks. It is important to note that the engine training fee does not include the cost of generating the hand scoring of sample papers.

Engine training fees in our survey and response to the Michigan RFP ranged from \$1,100 an item (excluding one low end response) to \$6,000.

The major factors influencing the amount of the engine training fees will be the complexity of the item and rubric used to score the item, and the amount of work the vendor needs to do to adjust its existing engine to score the item. Rubric complexity will also influence the sample size required for engine training. The number of responses to be scored for the item also can have an impact on the amount of the engine training fee, with a greater number of responses potentially resulting in a higher training fee. Additional important factors include the number of scoring models being built simultaneously and the validation requirements for the models, with a higher number of scoring models and greater validation requirements resulting in higher engine training fees. Other factors such as the time of the year the items are to be scored, the length of the scoring window and other vendor resource commitments can also play a role in determining the engine training fee. Most testing occurs during the spring, and vendor resources can get tied up during this period. A fall testing program will not put the same type of strain on vendor personnel resources and could result in lower fees for services.

- b) *Per response scoring fee* – The per response scoring fee is the cost for scoring each individual student response to the prompt during the operational administration. Once the engine has been trained and calibrated, the marginal cost to score an additional student response is quite low. Additionally, student responses can be machine scored quite rapidly, resulting in a far faster scoring turnaround time than is the case in hand scoring of open items.

Per item response scoring fees in our RFI survey ranged from \$0.006 to \$2.00. Responses to the Michigan RFP indicated per response scoring fees ranging from \$0.10 to \$0.36. The nature of the items, responses required and the numbers of responses to be scored are the major factors influencing the cost of scoring each student response. Student responses requiring a lot of inference or the use of multiple scoring models will generally cost more to score. On the other hand, a large quantity of responses should result in a lower per response charge.

- c) *Item setup fees* – Item setup fees are charged to recover vendor costs for work to acquire the responses and score data from the repository where they are stored, pre-process the data to address any quality or consistency issues and format the data so that they meet the input specifications of vendor scoring and reporting engines. Item setup fees are not charged consistently by all vendors. When they are charged, these fees are typically \$200 to \$300 an item.

The difficulty of retrieving the data and the nature of the programming to retrieve it are the major factors influencing cost variability for item setup fees. If the item data are not clean on import, programming and labor costs will increase significantly in order to retrieve and clean up the data. Additionally, for groups of similar items, significant cost savings can be realized in multiple models that are built at once rather than spread out over time.

- d) *Human second scoring of a sample of scores* – This cost element pertains to the expense of having human readers score a certain number of student responses to ensure that the machine scoring engine is generating accurate scores. This function is no different from a human read-behind function of hand scoring of open-ended items typically seen in high stakes assessments today.

The cost for human scoring of a sample of scores will run from \$0.60 to \$1.75 per response depending on the length and complexity of the student responses, the rubric against which the responses are scored and the total number of responses to be scored. For long form ELA essays such as the kind discussed in this paper, the cost for human scoring of these item responses is at the higher end of the range, \$1.25 to \$1.75. This assumes adequate time to prepare training materials and gather enough readers to score the responses in a reasonable time period. Additional considerations include training facility costs, if the scoring is centralized (versus distributed scoring) and costs for imaging hardware and software that readers use to both train and score. These costs can become significant, on a per item basis, if a small sample of items is to be scored. These additional considerations are not significant in high volume scoring situations but can be important in low volume scoring.

It should be noted that some vendors recommended a double blind human read-behind of responses with expert adjudication of discrepant scores as the best way to compare the results of human and machine scoring. Obviously, there are cost considerations in implementing this type of validity model as double human scoring would increase the per item fees noted above by about 67 percent.

B. General System Cost Elements – The second set of cost elements associated with a machine scoring implementation are general system costs. These costs are highly dependent on client needs, and the range of costs associated with each implementation will vary based on the magnitude of the effort the client determines appropriate for its assessment. Some of these costs may be included in the per item fees listed above depending on the magnitude of the effort involved in each particular area.

a) *Prompt design and consulting and item revision* – Prompt design and consulting, as well as item revision fees, are associated with using the machine scoring vendor’s staff to assist in the item development process. As mentioned earlier, the machine scoring vendor’s staff have experience in the type of items that both elicit student responses that are reflective of higher order critical thinking skills and are readily scored by the vendor’s machine scoring engine. Having the customer’s item development staff work in conjunction with the machine scoring vendor’s staff can yield higher quality open-ended items for the operational assessment. In a similar vein, item revision fees can be incurred if an item is not performing properly as is and can be revised to better fit the machine scoring engine parameters.

Charges for item design consulting are generally based on labor hours and will vary based on the extent of the work to be performed. The key factor is the experience of the item writers in writing items that are designed to be machine scored. If item writers have little experience in this area, as much as a week of training may be required. Additionally, machine scoring consultants may be needed to review and revise items during the initial few weeks of item development. Fees will generally range in the \$1,000 to \$1,500 per person, per day range.

b) *Infrastructure quality control* – Infrastructure quality control charges relate to the “care and feeding” of the technical architecture related to machine scoring of student essays. With each release of machine scoring software, and with each unique testing situation, load testing should be performed to ensure that the scoring software and hosting array perform as expected over a range of traffic volumes. Generally

speaking, these fees are included within the per item or per response fees, but can be separately quoted in certain cases, such as a major change mid-contract in item types to be scored, scoring methodology/ scoring model(s) or technology architecture.

- c) *Storage of student responses* – Costs for storage of student responses vary depending on whether responses are batch-scored by the machine scoring engine at the end of the test or streamed and scored in real time, as in computer adaptive testing (CAT). CAT requires careful planning to anticipate times of peak volume and ensure adequate server capacity. Capacity requirements are compounded if a large number of students are directing responses to hundreds (or thousands) of different constructed response items, each of which uses a different machine scoring model. Storage is relatively inexpensive, so in either case, the authors would not expect storage fees to be a significant part of the total cost of machine scoring.
- d) *Rubric validation and revision* – The scoring rubric for an item outlines the key scoring points and traits the scoring system should be searching for in a student's response. Response characteristics are provided at each score point level for each trait.

The scoring rubrics are validated and revised based on an early sample of student scores. For this reason, rubric design, validation and revision should occur as early as possible in the process, ideally in concert with item and prompt development. This is work the vendor will typically do as part of the engine training fee. In some cases, additional work may be priced separately based on labor hours of the vendor employees doing the work. Fees are generally quoted on a daily rate which can run from \$1,000 to \$1,500 per day.

- e) *Range finding* – The range finding process is an essential phase in the cycle of performance assessment scoring (human or machine scored). The process sets the standards for performance at each score point level for each trait, establishes the “standard” scores for a representative sample of responses for each item or prompt, and identifies exemplar responses that are then used to develop training materials for readers who will score each machine scoring training set of papers. When range finding is used to develop machine scoring models, special care needs to be taken to ensure that the sample drawn to train the scoring engine accurately represents the skill range and response variety that will be found in the general population. Response selection is a time consuming process and project schedules must allow sufficient time for skilled scoring personnel to evaluate the training responses for each item that will be machine scored.

Range finding is generally done in one meeting for all items. Significant travel and per diem costs are incurred for both vendor and customer staff, as well as for local educators who may attend the sessions. These costs can be allocated across the items reviewed to estimate a range finding cost per item.

- f) *Score file creation and integrity with data repository* – The machine scoring system will create a file of student scores that will need to be uploaded to the larger scoring system containing the student responses to other test items. It is essential that the files created contain the correct student scores and that these scores can then be matched to the remaining item scores for each student. The system may also be used to tell which scoring model should be used for each response.

There are two key considerations that have an impact on costs: 1) volume of simultaneous users over a definite period of time, and 2) response time in the case of instantaneous reporting on a computer adaptive test.

- g) *Data review and validity studies* – A client may wish to engage the machine scoring vendor to perform a number of studies to ensure that the data being generated by the machine scoring system is correct and valid and that the items are measuring the intended constructs. Data review will include the typical review and analysis of item statistics (Item Response Theory (IRT), item parameter characteristics, item difficulty, discrimination, reliability, Differential Item Functioning, etc.).

The cost involved in these studies depends on the additional analyses beyond classical or IRT item statistics requested by the client, such as those for various types of validity studies. The number of prompts undergoing review, as well as the number of grades/students to review, will impact the cost of data review and analysis. Also, the more data reviewed, the greater the length of the review meetings. Travel costs may also apply.

Data review and validity studies can cost from \$10,000 to \$50,000, depending on the nature and depth of the study.

- h) *System monitoring and audit fees* – System monitoring continually checks system availability, security, and performance and provides real time alerting of detected abnormalities. To accomplish this, several independent services continually ping the system and report stability, bandwidth and response time, and other metrics. Continual monitoring of scoring engine performance is also required to ensure that scoring quality and performance are maintained because infrastructure can fail and there can be “drift” in the scoring of student responses. Each service can cost hundreds of dollars per month for each monitored site.

Auditing is a more intensive and costly process, particularly when it is conducted by an outside agency. An audit is performed periodically to ensure correct system functionality, adequate security, and appropriate usage through active means, as well as through a review of system audit logs. Comprehensive system auditing can cost up to \$20,000 for an initial study, plus \$1,200 per site annually on an ongoing basis.

The exhibits below and on the following page summarize the key cost elements in machine scoring of student essays.

EXHIBIT 2: KEY COST ELEMENTS IN MACHINE SCORING - PER ITEM / RESPONSE FEES

Key Cost Elements In Machine Scoring

Per Item / Response Fees		Price Range		Factors Influencing Costs
Cost Element	Description	Low	High	
Engine Training	Calibration of the scoring engine to score papers accurately at all score points for all traits	\$1,000	\$6,000	<ul style="list-style-type: none"> Item and rubric complexity Amount of work required to adjust vendor scoring engine Number of responses to be scored Availability of vendor resources
Per Response Scoring	Scoring of individual student responses to a prompt	\$0.10	\$2.00	<ul style="list-style-type: none"> Volume of responses for a given prompt Total number of prompts to be scored Difficulty of prompts to be scored
Item Setup Fees	Work to acquire the response and score data from the repository where they are stored, pre-process and format the data	\$ --	\$300	<ul style="list-style-type: none"> Can be included in per response scoring Difficulty of retrieving data Difficulty of cleaning up data

EXHIBIT 3: KEY COST ELEMENTS IN MACHINE SCORING - GENERAL SYSTEM COSTS

Key Cost Elements In Machine Scoring

General System Cost Elements		Price Range		Factors Influencing Costs
Cost Element	Description	Low	High	
Prompt Design and Consulting / Item Revision	Assistance provided by vendor machine scoring staff to customer item writers to assist in writing items that can be readily machine scored	\$1,000 / day	\$1,500 / day	<ul style="list-style-type: none"> Experience of item writing staff in developing items that can be readily machine scored. Amount of assistance required.
Infrastructure Quality Control	The “care and feeding” of the technical architecture related to machine scoring of student essays	\$ --	\$20,000	<ul style="list-style-type: none"> General monitoring of the system is included. A full blown audit of the system is expensive.
Storage of Responses	Storage of student responses on easily retrievable media	\$ --	\$10,000 / month	<ul style="list-style-type: none"> If responses needed for CAT item presentation. Volume of responses. Length of time responses need to be stored.
Rubric Validation and Revision	Testing the accuracy of the scoring rubric by machine scoring a small sample of items and measuring the results. Revising the rubric if required.	\$ --	\$1,500 / day	<ul style="list-style-type: none"> Can be included as part of the scoring package. Difficult items can require assistance from vendor staff.
Range Finding	The process of setting the standards for performance at each score point level for each trait of the rubric, and identifying exemplar responses that are then used to develop training materials for human readers and the machine scoring engine for actual response scoring	Significant Travel Expenses	Significant Travel Expenses	<ul style="list-style-type: none"> Range finding is done in large meetings with many educators, vendor and assessment staff in attendance. The process is essentially the same for both human and machine scoring of student responses.
Score File Creation / Data Integrity	Creation of the file containing the student response scores and integration of the file with the data storage and reporting system	\$ --	TBD	<ul style="list-style-type: none"> Generally included in the cost of machine scoring but can be separately priced if the customer has certain unique requirements. Use of the system in a CAT environment could result in additional costs.
Data Review and Validity Studies	Ensuring that the data being generated by the machine scoring system is correct and valid and that the items are measuring the intended constructs. Data review will include the typical review and analysis of item statistics.	\$10,000	\$50,000	<ul style="list-style-type: none"> This work is done as a part of the general data review of the assessment and also needs to be done for items that are machine scored. Costs will depend on the number and depth of the various studies performed.
System Monitoring and Audit	Continual checking of system availability, security, and performance and real time alerting of detected abnormalities	\$ --	\$20,000	<ul style="list-style-type: none"> Normal system monitoring is part of infrastructure quality control. Full blown audit of the system is extra.

SUMMARY OF KEY COST ELEMENTS IN IMPLEMENTING MACHINE SCORING

This section provided an overview of the many cost elements involved in machine scoring. The most important cost elements are per item and per response fees. In some cases, the general systems costs are amortized within the per item and per response fees. Customers considering implementing machine scoring of student essays should understand the nature and components of the per item engine training fees and per response processing fees, as these items make up the majority of the costs of machine scoring.

Additionally, potential customers of machine scoring services need to consider that some human scoring still may be involved in the process. Typically, a small percentage of student responses will not be able to be scored by the engine and will require human scoring. Furthermore, states can consider whether to use machine scoring as the primary grading system or as the read-behind process for validating hand scoring. Finally, customers should include a factor for the costs of other general system cost elements based on their own needs and discussions with the vendor. Later in this paper the authors present a model for machine scoring implementation costs.



V
**KEY
 ASSUMPTIONS
 AND INPUTS
 FOR
 INCLUSION
 IN RFPs**
 (vendor provided
 information)

Machine scoring vendors were asked during interviews to describe the information that should be included in a state or assessment consortium RFP for services that would be sufficient to allow them to provide a quality response to the proposal, including an accurate estimate of the costs of machine scoring. This section presents the key assumptions and inputs for inclusion in machine scoring RFPs, a definition of the assumption/input, and factors influencing cost variability of the assumption/input.

A. **Item descriptions** – The most important information to include in any RFP for machine scoring services is a thorough description of the nature and types of items that will be used to elicit the student responses, the nature of the student responses expected and how the responses will be scored.

Key item and response information includes the item scoring rubric (including number of traits and point scales for each), whether the student response will be source dependent (students referencing the source document or prompt in their response) or not, if the response is to be scored for writing construction and the length of the stimulus material and expected student responses. All of these things should be addressed in the RFP.

B. **Item rubrics** – As described above and also in section IV, the item rubrics outline the key scoring points and traits the scoring system should be searching for in a student’s response. Response characteristics are provided at each score point level for each trait. Entities issuing RFPs for machine scoring services should include a copy of the rubrics, or examples of the types of rubrics that will be developed, for the open-ended and/or essay item types in the proposal.

Similar to the discussion above, item and response complexity will directly impact the nature of the scoring model and have an impact on how expensive the model development and scoring will be.

- C. **Item counts** – The number of different items to be scored using machine scoring will have an impact on the price of the service. More items to be scored equates to more work for the vendor, so the potential for volume pricing and discounts increases. On the other hand, too many items requested to be scored using machine scoring may put a crimp on available vendor resources and actually result in higher prices for engine training, item consulting and revision, etc.
- D. **Item release rates** – Item release rates, or the number/percentage of items that will no longer be used on a future assessment because they were retired or released to the public, will impact the nature of the pricing model. What is important in this instance is actually the inverse of the item release rate or the item reuse rate. Once the machine scoring engine has been trained and validated for a particular item, it should not have to be retrained, and therefore there is very little marginal cost in scoring responses for the same item in a subsequent assessment year. This should result in a lower total price to score the item responses.

It is important to remember that many performance items/tasks or innovative/unique constructed response items tend to be much more “memorable” than a selected response item, and consequently the item release rate for these should be quite a bit higher than for selected response items.

- E. **Field test scoring assumptions** – The field testing assumptions related to the items revolve primarily around the number and quality of the sample papers that will be available for the vendor to use to train the scoring engine. As noted above, anywhere from 1,500 to 2,500 responses may be required to get enough papers at each score point (for each trait) required to train the engine. Since student responses typically do not have an even distribution across all the score points, of particular concern is having enough papers at the highest and lowest points, particularly the former. Having the responses accurately scored, with an available rubric, and ready to be input to the scoring system will save time and money. Should the vendor be required to assist in any of the field testing, rubric creating, human scoring, etc., the price will rise accordingly.
- F. **Student counts and volumes** – Student counts and volumes provide information on the number of responses that will require scoring for each item. The greater the number of student responses for a particular item, the lower the per response item fees will be (and vice versa).

- G. **Exam locations** – If the vendor is not using a web-based machine scoring system, the number of different exam locations and scoring installations will be important factors in determining vendor costs. It is expected that most vendors will provide web-based services. In this case, overall hosting/server capacity will determine if any surcharges for multiple locations/servers will be required.

- H. **Scoring turnaround time requirements** – The time period between when the item is presented to a student and when the scores are required to be delivered to the primary scoring data repository could impact vendor pricing. The machine scoring engine will score student responses very quickly and efficiently. However, it generally does take some time to make sure the scoring process worked as planned, review/audit the scoring file and make any required adjustments. If vendor resources are constrained in the effort to turn around scores, pricing could be impacted.

It is unclear how the use of machine scoring in a CAT environment will work if the student open-ended responses are to be scored immediately and used to determine future items to be presented to the student. Procurers of machine scoring services should require bidders to explain in detail how this will be done and question vendors responding to their solicitations as to how the machine scoring implementation will work with computer adaptive testing, and what the associated costs will be.

- I. **Scoring time of year** – The time of the year the assessment and scoring are to take place is another important factor in determining some of the pricing elements in machine scoring. Most assessments and most scoring occur in the spring when vendor resources are at their busiest. Testing programs that put a constraint on vendor available resources (engine training, prompt consulting, prompt revision, etc.) can be expected to be priced a bit higher than fall testing programs.

- J. **Human read-behind** – Potential customers issuing RFPs for machine scoring services should include any requirements or expectations regarding the percentage of responses that will also be scored by humans in order to check the accuracy of the machine scored responses. Many machine scoring vendors also provide human scoring of open-ended items; a bigger pool of services that a vendor can propose could positively impact the total price of the program. Additionally, the degree of human audit and the required scorer agreement rates (inter-rater reliability) are important items for the machine scoring vendor to take into account when creating the scoring model and calibrating the accuracy of the model scores.

Factors influencing the cost of human scoring include the time of the year the scoring is required (see above discussion on scoring time of year), where the scoring will be done and the location of the vendor scoring centers (and therefore the cost of the labor pool from which the vendor hires its readers) and/or if scoring is to be distributed via systems and not be site-based.

- K. **Agreement rates required** – Agreement rates (inter-rater reliability) measure the percentage of time a human scorer and machine scoring engine will give the same score or a score within one score point of each other. The higher the required agreement rate, the more precise the scoring model required. An acceptable agreement rate for most high stakes testing is usually about 80 percent.
- L. **Exam purpose** – The purpose of the exam (such as grade promotion, graduation requirement and teacher accountability) will have some impact on the cost of machine scoring services. The purpose of the assessment is directly tied to the security requirements of the data and can have an impact on the vendor’s technical architecture that is set up for the particular client and assessment. The more secure the data need to be, the greater the cost of the technology infrastructure. Additionally, the more high stakes the exam, the more exact the scoring model needs to be. More precise scoring models will generally cost more than less precise scoring models.
- M. **Security requirements** – See L, Exam purpose. In addition, the size of the testing window, the physical settings used for testing, the possibility that students will share items with others, the need to do data forensics on the student responses as well as other factors may also have an impact on the exams. Vendors will need to work closely with their customers to monitor the security requirements and possibly propose enhancements to the process.
- N. **Student response input method** – Student essay responses can be typed into a computerized system and delivered to the machine scoring engine or handwritten and scanned into the scoring system. It is highly recommended that all high stakes testing implementations of machine scoring use only direct input responses (not handwritten) for machine scoring purposes. Optical character recognition (OCR) software may not be able to accurately translate 100 percent of a given student’s response, and often issues such as capitalization, carriage returns and other challenges specific to OCR protocols present other issues that threaten valid representations of student work. This can make it difficult for the machine scoring engine to deliver an accurate score. For purposes of this paper, the authors assumed that all responses to be machine scored are typed directly into the system.

- O. **System interface issues (item presentation and scoring file)** – System interface issues have to do with how the items are presented to students, how the responses are captured and loaded into the scoring system and how the scores are stored and fed into the customer’s or other vendor’s scoring system. In many instances, the machine scoring system will operate in the background of the main test delivery system. The main test delivery system will deliver the items to the students, and the students will type their responses into the primary testing system. The primary testing system will then deliver the responses to the machine scoring system for scoring. The machine scoring system will score the responses and deliver the scores back to the primary testing system scoring repository.

The testing and machine scoring configurations should be noted in the RFP and any unique item or data presentation requirements outlined to the potential vendors.

Exhibit 4 on the following page, highlights some of the items recommended by vendors that should be included in RFPs for machine scoring, and their relative importance.

SUMMARY OF KEY ASSUMPTIONS AND INPUTS FOR INCLUSION IN RFPs

This section of the paper reviewed the items that potential customers of machine scoring services should consider including in any RFPs for services they develop. Vendors interviewed stated that the RFPs they have reviewed generally do not provide enough information for them to accurately project the costs of a machine scoring implementation. Including these items in RFPs for machine scoring services should result in more accurate vendor pricing proposals.

EXHIBIT 4: VENDOR RECOMMENDED ITEMS FOR INCLUSION IN RFPs FOR MACHINE SCORING

Key Information to Include In RFPs for Machine Scoring Services

Element	Description	Importance	Comments
Item Descriptions	Nature/description of the type of items to be scored using machine scoring	High	Whether the item response will be source dependent, require inference or be scored on writing structure is key; nature of the scoring rubric and expected length of the response are also important.
Item Rubrics	A description of the item scoring rubric including number of score points and number of traits	High	Sample rubrics should be included in the RFP.
Item Counts	The actual number of items to be scored using machine scoring	Moderate-High	Larger volumes of work can generate greater economies of scale.
Item Release Rate(s)	The number or percentage of items that will not be used on a future assessment because they are released to the public or retired	Moderate-High	Items that will be re-used on the upcoming assessment do not have to go through engine training again.
Field Test Scoring Assumptions	Relates to the number and quality of sample papers and scoring rubrics that are available for engine training	Moderate-High	Generally, 1,500 to 2,500 scored papers are needed to generate enough papers at each score point and trait.
Student Counts / Volumes	The number of students that will be responding to a given prompt	High	The larger the volume of responses per item, the lower the per item response fee.
Exam Locations	Number of sites where scoring will take place	Low-Moderate	Scoring sites will have an impact on the number of vendor servers needed for scoring.
Scoring Turnaround Time	The time period between when the item is presented to a student and when the scores are required to be delivered to the primary scoring data repository	Low-Moderate	This can be important if the item response score is used to determine future items to be presented to a student in CAT.
Scoring Time of Year	Calendar week(s) when testing occurs	Low-Moderate	This can impact vendor resources if several programs require services at the same time.
Human Read-Behind	Percentage of items that also will be scored by humans as a check on the machine scoring engine	Low	Many machine scoring vendors also provide human scoring services.
Exam Purpose and Security Requirements	Nature and purpose of the exam, i.e. summative high stakes or low stakes diagnostic	High	High stakes tests will have higher accuracy and security requirements, which can impact the model building and validation work, human scoring read-behind and infrastructure required to run the program.
Student Response Input Method	How student responses are delivered to the machine scoring engine (hand-written or typed into a computer)	High	Only student typed responses should be used for machine scoring.
System Interface Issues	Covers various issues such as how the items are presented to students, how the responses are captured and loaded into the scoring system, and how the scores are stored and fed into the customer's or other vendor's primary scoring system	Low-Moderate	If the machine scoring system operates in the background of a primary test delivery engine, APIs will be required to integrate the two systems for scoring and reporting purposes.



VI

ADDITIONAL INFORMATION FOR INCLUSION IN RFPs (non-vendor provided information)

In addition to the machine scoring vendors, others in the assessment industry who are very knowledgeable in the area of machine scoring were interviewed. In this section, additional information is presented that the authors recommend be included in any RFP for machine scoring services.

- A. **Vendor client list** – A listing of the vendor’s more recent clients, a description of the service(s) provided (such as types and numbers of items, number of responses scored, etc.) will give the potential customer a good feel for the vendor’s experience in the area. The client list can also serve as a set of references for the potential customer.
- B. **Average and peak volumes scored by day, week, month** – This data, which can be requested along with the vendor client list, will provide the potential client with a sense of whether the vendor has had prior engagements with similar capacity and load requirements to those that the potential customer is planning.
- C. **Company capabilities and capacities related to machine scoring of student essays, and infrastructure currently in place to handle contract volume requirements** – A general discussion of vendor capacity, load balancing system availability, back-up and recovery capabilities and other security capabilities are all important system aspects for the potential vendor to describe.

- D. **Description of machine scoring engines and the types of items and length of student responses the vendor is best suited to score** – As noted previously in this paper, different machine scoring vendors will use different scoring engines depending on the nature of the item prompt and responses. Some engines are better than others at scoring short essay answers. Some are better at scoring responses that are source based and some are better at scoring responses for writing structure. It is important to understand the type(s) of items the customer will be using and the strength of the potential vendors' scoring engine(s) at scoring the responses generated by those items.

In addition, any studies the vendor has that compare the accuracy of its proposed scoring system to human scoring and agreement rates between its scoring engine and human scoring should be requested. For the true aficionado of machine scoring, the nature and description of the scoring algorithm used to score the responses can be requested.

- E. **Sample service agreement** – A sample service agreement should be requested as part of the RFP submission. A review of the vendor service agreement can point out various customer requirements the vendor may have, as well as other operational issues to be considered in vendor selection.
- F. **Listing of any operational or legal use restrictions on vendor technology** – A potential purchaser of machine scoring services should determine if there are any restrictions on the use of the vendor technology being proposed either from an operational or legal standpoint.
- G. **Quality control and audit processes in place** – If not included in any of the prior vendor descriptions of its capabilities and services, the quality control and audit features the vendor has in place to ensure accurate and valid scoring should be addressed in the RFP.
- H. **Format and design used to return scores to a third party test delivery system** – It is helpful for both customers and potential vendors of machine scoring services to be aware of the format and design in which test scores must be delivered to a third party test delivery system. Any unique requirements can be determined and properly assessed by each side.
- I. **Other technical issues** – The vendor should describe how its machine scoring of student essays meets technical standards for
- a) Validity
 - b) Reliability
 - c) Fairness and freedom from bias

- J. **Number of responses required to train the system** – Different vendors of machine scoring services may require different numbers of responses in order to train their machine scoring engines. It is important to understand the expected number of responses required so the customer can plan appropriately for field testing.
- K. **Item security protocol** – Item security protocol will establish the particular procedures and processes the vendor uses to maintain security over the items and responses.
- L. **Vendor documented application program interfaces (APIs) that are provided to the prime test delivery interface provider** – A look at the machine scoring vendor APIs for interfacing with other systems can be requested if the customer has the IT staff available to examine the APIs and determine the amount of additional IT work to customer systems that will be required to properly interface with the machine scoring applications.
- M. **System auto save and back-up function description** – It is desirable for vendor machine scoring software to include automatic save and back-up features so if there is an outage during testing student work will not be lost.

SUMMARY OF ADDITIONAL INFORMATION FOR INCLUSION IN RFPS

This section of the paper provided details on additional information to be included in customer machine scoring solicitations (RFPS). This information was not mentioned by the vendors during the interviews but was added based on the authors' discussions with others in and around the assessment and machine scoring communities. Potential customers of machine scoring services can pick and choose the items they ultimately include in a solicitation for services based on the level of detail they choose to review and in-house technical expertise they may have at their disposal. Costs for the items that are selected by the customer should be spelled out by the vendors in their responses to the RFP, preferably in detail.



VII PRICING DATA

Section VII reviews the pricing data gathered and analyzed in this paper. The authors collected data from industry vendors in response to the RFI and augmented that data with information obtained from the vendor responses to the Michigan RFP for implementation of the SBAC assessment system in the state. Based on the data analyzed, the authors put forward some estimates of what procurers of machine scoring services can expect to pay for the services.

A. Information from vendors in response to the RFI

Per Item Pricing

Table 1 below shows a summary of the pricing data vendors supplied in response to the RFI. The table shows the high, medium and low prices submitted by vendors at different student counts for both the field test and operational years of the assessment. One vendor provided a high and low range of pricing data, so in some instances one or both figures may be shown in the table. Generally speaking, the vendors did not supply significantly different pricing information between the type of items PARCC and SBAC will be using on their assessments, so the authors have chosen to present a summary of the per item pricing data in one table.

TABLE 1: Per Item Pricing Information

VENDOR RESPONSES TO RFI PRICING REQUEST
Per Item Pricing information

	Field Test*	Volume Small <500K Students			Volume 1M-2M Students			Volume >5M Students		
		Low \$	Med \$	High \$	Low \$	Med \$	High \$	Low \$	Med \$	High \$
Item Setup Fees	\$190	\$51	\$79	\$200	\$51	\$79	\$200	\$51	\$79	\$200
Engine Training Fees	\$200	\$93	\$1,400	\$3,000	\$93	\$1,400	\$3,000	\$65	\$1,100	\$3,000
Item Per Response Scoring	\$4.50-\$2.20	\$0.04	\$0.55	\$2.20	\$0.04	\$0.48	\$2.20	\$0.04	\$0.35	\$2.20
Human Double Scoring**	\$25-\$45	\$1.75-\$2.40	\$1.75-\$2.40	\$2.75-\$3.75	\$1.75-\$2.40	\$1.75-\$2.40	\$1.75-\$2.40	\$1.65-\$2.35	\$1.65-\$2.35	\$1.65-\$2.35
Other**	\$227-\$1,333									

* Average or Range of Responses
** One vendor submitted range

Note: Where a vendor differentiated between a short and long response to score, the reduction in price from a long essay to a short essay was \$500 in engine training fees and \$0.05 in the per response scoring fee.

Overall, vendors reported a wider spread of pricing data than expected. For the two most significant per item pricing line items, engine training and per response processing fees, the range of pricing differences is significant. However, the authors believe the high and low end of the reported range for engine training fees and item per response scoring fees represent, somewhat, extreme cases. The price quote at the low end of the range assumed no modification to vendor existing engines or scoring models. Similarly, the price quote at the high end of the range assumed major modifications and/or development of new scoring engines. The authors hope that the large majority of machine scoring implementations will involve scoring engines with typical programming and modifications. If this is the case, the midpoint of the pricing range represents a more realistic, yet conservative, view of pricing expectations.

Item Setup Fees

Item setup fees also varied somewhat, but regardless of where they ultimately end up within the quoted range, these costs will not have a significant impact on the total price paid to vendors for machine scoring services.

Engine Training Fees

Engine training fees went from a low response of \$65 per item to a high of \$3,000 per item in the highest volume case (>5 million students). Most pricing in the market today (based on the authors' experience and other anecdotal evidence) for engine training fees is about \$3,000 to \$5,000 per item, so the vendor responses were all at or below current market pricing. The authors did not see significant reduction in per item engine training fees as student volume increased. This is most likely due to two reasons: a) Vendors were conservative in responding to the RFI, and b) Per item engine training fees are relatively constant and do not vary based on the number of responses that will ultimately be scored.

In conducting the cost study, the team expected vendors to be relatively conservative in their pricing responses, as all vendors expressed some level of concern in submitting pricing data related to the RFI, despite the fact that the data would be anonymous. The authors expect vendors to be more aggressive in their pricing for actual revenue-generating opportunities. Indeed, this has been validated somewhat in responses that were seen in the Michigan RFP for implementing the SBAC assessment system in the state. It is the authors' understanding, based on data from a limited sample and anecdotal information, that most machine scoring implementations currently in place have engine training fees in the \$5,000 per prompt range and per response fees in the \$0.75 to \$2.00 neighborhood.

Per Response Processing Fees

Per response processing fees went from a low of \$0.04 per response to a high of \$2.20. It should be noted that both extremes were stated as being such by the vendors submitting each pricing estimate. The \$0.04 per item at the low end of the range (this vendor also quoted \$0.006 for one constructed response item type) was based on the vendor's current engine being able to score the responses without any additional adaptations to the engine. In a similar vein, the \$2.20 per response processing fee was at the high end of range of pricing of a vendor submission and based on extensive rework and calibration of its scoring engine. In the authors' opinion, the mid-range of pricing of \$0.35 to \$0.55 a response is probably slightly high but in the reasonable range of what might be expected for a single state or small consortium assessment.

Human Double Scoring of Items

An expected normal cost to human score these items would be about \$1 to \$2 (depending on the nature of the item), provided volumes are high enough (per ASG's Cost Model). Costs will be higher should scoring volumes be low. It should also be noted that one percent to five percent of the student responses for a given item typically cannot be scored by the machine scoring engine and are set aside for human scoring.

TABLE 2: Other Services Pricing

VENDOR RESPONSES TO RFI PRICING REQUEST
General System Pricing information

	Field Test*	Operational**	
		Low Range	Med Range
a. Prompt Design Consulting	\$185	N/Q	N/Q
b. Infrastructure Quality Control - Audit	\$20,000	\$20,000	\$20,000
c. Storage of Students Responses	\$1,000	\$1,000	\$1,000
d. Rubric Validation and Revision	\$270	N/Q	N/Q
e. Range Finding	\$3,750	N/Q	N/Q
h. Data Review and Validity Studies	\$3,250	\$12,500	\$52,300
i. Documentation (peer review, tech manuals)	\$20,500	\$15,310	\$25,735
j. System Monitoring	\$850	\$420	\$1,270

Specific Notes:

- a. and d. are per item/rubric, with range dependent upon revision complexity. This assumes no new items in Yr1 and Yr2.
- b. is per audit event. Yr1 and Yr2 may not require audit.
- e. is per item, paper selection and 5 person committee establishing exemplars. Range is dependent upon rubric/score scale complexity. This assumes no new items in Yr1 and Yr2.
- h. and i. are per study/manual, with range dependent upon manual/study complexity.
- c. and j. are per site per month, with range dependent upon simultaneous usage.

More Notes:

- Years 3, 4, and 5 (if applicable) will be the same as Year 2.
- Years 1-5 do not include COLA adjustments, which may apply.
- System monitoring will be directly related to the length of the administration window.

* Average or Range of Responses
** One vendor submitted range

Other Systems Costs

The RFI requested vendors to respond with pricing information for a number of items defined as “other systems costs.” These items tend to be incurred based on individual customer needs. One vendor responded with a range of costs for these items and the data are presented in the table below.

States and assessment consortia should take careful note of the items in Table 2 and their respective costs and make sure they include cost estimates for items of interest in their pricing estimates. As mentioned previously, some of the items in Table 2 (infrastructure quality control, storage and system monitoring) are often included in the per response fees. A minimal amount of prompt design consulting and rubric validation and revision might also be added into per response fees. Anything more than minimal amounts of prompt design and rubric validation, along with a full system audit, validation studies and documentation for peer review, will most likely result in separate charges. Later in this paper, a planning level for spending on these system components is recommended.

B. Pricing information from vendor responses to the Michigan RFP

As noted previously, in late 2012 the state of Michigan issued an RFP to implement the SBAC assessment system statewide to test roughly 840,000 students. Machine scoring of items was included in the RFP, and three vendors responded. Two of the three vendors subcontracted with the same machine scoring vendor in responding to the RFP. The following data summarize the per item fees submitted by each vendor.

Michigan RFP Responses

Cost Item	Vendor 1	Vendor 2	Vendor 3
Engine Training Fee	\$2,000	\$4,850-\$6,075	\$1,100
Per Response Processing Fee	\$0.10	\$0.35	\$0.35
Other	N/A	N/A	N/A

The engine training fees quoted by Vendors 1 and 3 are at the lower end of the range of quotes received in response to the RFI, and the per response processing fees of all three vendors are also at the lower end of the range of price estimates submitted.

SUMMARY OF PRICING DATA

As noted above, pricing estimates submitted by vendors in response to the RFI covered a significant range from high to low. Additionally, pricing figures obtained from the vendor responses to the Michigan RFP tended to be at the lower end of the midrange of vendor responses to the RFI (which the authors believe is reasonably accurate).

Machine scoring vendors have made significant investments in their machine scoring departments, and the use of machine scoring in high stakes assessment has been somewhat limited to date. Pricing of existing implementations of machine scoring services may include a higher amount of investment cost recovery than will be required in future implementations. Indeed, with both PARCC and SBAC seriously considering the use of machine scoring in their assessment systems, volume implementations of machine scoring services should be increasing in the near future. Thus, the authors expect to see lower prices for machine scoring services in the near future.



VIII

PRICING EXPECTATIONS

What can an implementer of machine scoring services expect to pay?

A great deal of data has been presented in this paper and in this section the cost study team will consolidate and integrate all the information to estimate a range of costs a state or assessment consortium might expect to pay in implementing machine scoring of long form (>500 words) student essays. The data are presented for small, medium and large student counts and are the authors' best estimate based on the industry data presented in this paper and in discussions with machine scoring industry personnel.

Machine scoring costs are highly variable and depend on the nature of the items being scored, the item rubrics, volume of responses to score and whether the item lends itself to being accurately scored by the vendor's machine scoring engine. For purposes of this exercise, the authors assume that items can be scored by current vendor machine scoring engines with some reasonable modifications/programming to current engine types. Note that since the assumption is for long form essays, the authors did not consider the low quote received from one vendor of \$0.006 per CR response (although potential procurers of machine scoring services should keep this figure in mind).

It is also important to remember that even with accurately calibrated engines, one percent to five percent of student responses will not be able to be scored by the engine and will require human scoring. Additionally, at least in the initial year(s) of a machine scoring implementation, a reasonable percentage of papers (10 percent to 20 percent) should be checked by human readers. Once the customer is satisfied with the performance of the machine scoring engine, the read-behind rate can be reduced.

What about fees for other machine scoring system costs?

Finally, customers should “bake” into their cost estimates fees for some of the other system costs described in section VII. The table below provides pricing estimates for states and consortia of states to consider when implementing machine scoring of student essays.

TABLE 3: Price Ranges for Long Form ELA Machine Scoring Services at Different Volumes

AUTHOR ESTIMATE OF MACHINE SCORING COST RANGE, BY COST ELEMENT, FOR LONG FORM ESSAY ELA ITEM

Per Item / Response Cost Element	Low Volume <500K Students		Medium Volume 750K--1.5M Students		Multi-State Volume 1.5M--3M Students		Consortium Volume >5M Students	
	Low Range	High Range	Low Range	High Range	Low Range	High Range	Low Range	High Range
Item Set-up Fees	\$200	\$200	\$200	\$200	\$200	\$200	\$200	\$200
Engine Training Fees	\$2,000	\$5,000	\$1,000	\$3,000	\$750	\$2,750	\$500	\$2,500
Item Per Response Fees	\$0.25	\$0.65	\$0.15	\$0.45	\$0.075	\$0.35	\$0.05	\$0.25
Per Response Human Scoring	\$1.28	\$1.75	\$1.27	\$1.74	\$1.26	\$1.73	\$1.26	\$1.73
Other Systems Costs - % of Per Item Costs	12%	15%	8%	12%	5%	10%	5%	10%

Table 4 on the following page, details what a customer of machine scoring services might expect to pay for scoring of an ELA long form essay item.

In looking at the tables above one can see that, in general, machine scoring of student essays, including a 20 percent human read-behind of machine generated scores and five percent fall out rate of student responses requiring human scoring, is significantly less expensive than full human scoring (which also includes a 20 percent read-behind). For the two lowest volume scenarios, the cost of machine scoring at the high end of the range is very close to human scoring costs at the low end of the range. However, once higher volume scenarios are achieved, machine scoring is significantly less expensive than human scoring regardless of where the item and responses end up in the cost range.

Furthermore, distributed scoring is generally 20 percent to 30 percent less expensive than site-based human scoring. For high volume scoring situations, a mix of site-based and distributed scoring is generally necessary. A distributed scoring implementation would narrow the gap somewhat between machine and human scoring in the above scenarios, but not by a significant amount. The authors of this cost study believe the findings above hold for both site and distributed human scoring.

TABLE 4: Cost Comparison of Machine vs. Human Scoring of Student Essays

AUTHOR ESTIMATE OF MACHINE VS. HUMAN SCORING COSTS FOR LONG FORM ESSAY ELA ITEM

Per Item / Response Cost Element	Low Volume <500K Students		Medium Volume 750K--1.5M Students		Multi-State Volume 1.5M--3M Students		Consortium Volume >5M Students	
	Low Range	High Range	Low Range	High Range	Low Range	High Range	Low Range	High Range
Assumed Student Count	500,000	500,000	1,000,000	1,000,000	2,250,000	2,250,000	5,000,000	5,000,000
Number of Items	1	1	1	1	1	1	1	1
Human Read-Behind Rage	20%	20%	20%	20%	20%	20%	20%	20%
Item Set-up Fees	\$200	\$200	\$200	\$200	\$200	\$200	\$200	\$200
Engine Training Fees	\$2,000	\$5,000	\$1,000	\$3,000	\$750	\$2,750	\$500	\$2,500
Item Per Response Fees	\$125,000	\$325,000	\$150,000	\$450,000	\$168,750	\$787,500	\$250,000	\$1,250,00
Per Response Human Scoring	\$160,000	\$218,750	\$317,500	\$435,000	\$708,750	\$973,125	\$1,575,000	\$2,162,500
Other Systems Costs - % of Per Item Costs	\$34,464	\$82,343	\$37,496	\$106,584	\$43,923	\$176,358	\$91,285	\$341,520
Machine Scoring Cost*	\$321,664	\$631,293	\$506,196	\$994,784	\$922,373	\$1,939,933	\$1,916,985	\$3,756,720
Cost Per Student Per Question	\$0.64	\$1.26	\$0.51	\$0.99	\$0.41	\$0.86	\$0.38	\$0.75
Human Scoring Cost*	\$768,000	\$1,050,000	\$1,524,000	\$2,088,000	\$3,402,000	\$4,671,000	\$7,560,000	\$10,380,000
Cost Per Student Per Question	\$1.54	\$2.10	\$1.52	\$2.09	\$1.51	\$2.08	\$1.51	\$2.08

* Excludes costs of range finding, anchor pulling and site visits which occur for both human and machine scoring



IX FINAL CONCLUSIONS AND RECOMMENDATIONS

The use of machine scoring of student essays is expected to increase significantly over the next few years. Both PARCC and SBAC are considering using machine scoring services to enable their next generation assessments to better measure the skills students will need in the 21st century at an affordable price. As of June 2013, two SBAC member states already have issued RFPs for assessment services that include machine scoring components, with several more expected. Additionally, SBAC issued RFP 16-17 which includes machine scoring of open-ended items that will be field tested in 2014.

The expected increased demand for machine scoring services is attracting new entrants to the field, and existing service providers are looking to add new talent and further upgrade their capabilities. All these factors should result in better scoring engines and lower pricing in the not-too-distant future. Offsetting this decreasing price trend somewhat may be the additional investment required for vendors to develop the improved engines required to score new types of items being developed by the two assessment consortia.

ASAP has thus far tested the performance of machine scoring engines against humans for long form (>150 words) and short form (<150 words) student responses. Vendor scoring engines were able to produce scores on long form essays that tended to match those of human raters, in the aggregate. While performance in machine scoring short form constructed responses was not quite as good as that of humans, there still may be some low stakes applications where machine scoring services can be used, perhaps as a read-behind of human scoring. Moreover, the longer the student response, the more costly it is to score with humans, so it is significant that many long form essays currently can be scored successfully by machine scoring engines. Therefore, machine scoring engines offer the promise of being able to score performance type items at a reasonable cost with the additional benefit of dramatically reducing time

to process scores and report results. At this time, however, it is still too early to tell if current engines will be able to score any or all of the types of items being developed by the two common assessment consortia.

Due to the lack of use of machine scoring in high stakes assessment to date, states and consortia of states are not well-versed in the practice and have not issued quality RFPs for these services. One of the goals of this paper was to provide the customer community with a thorough review of the machine scoring process, cost elements related to machine scoring, and factors to consider in implementing machine scoring in their state or assessment consortium.

Implementing machine scoring should be a joint process between vendor and customer. Customer development staff should work with the machine scoring vendor to gain an appreciation as to how to develop items that can measure student critical thinking skills and be accurately scored by a machine scoring engine. This process is often overlooked to the detriment of the customer and the students being assessed. Additionally, the customer and machine scoring vendor should work together to define all of the machine scoring service and system elements required in a high stakes testing environment.

The major cost elements in a machine scoring implementation are the per item and per response fees. However, other systems-related costs also need to be considered (whether they are amortized in the cost of the per response fees or separately quoted) in any implementation of machine scoring services.

The authors have generated a range of expected costs for purchasers of machine scoring services. As shown in Table 4, Section VIII, machine scoring services can be significantly less expensive than human scoring, and there is ample reason to believe that the next generation of higher quality assessments that make use of these services will be affordable for states and consortia of states in the future.

REFERENCES

Topol, B., Olson, J., & Roeber, E. (2010). *The Cost of New Higher Quality Assessments: A Comprehensive Analysis of the Potential Costs for Future State Assessments*. Stanford, CA: Stanford University, Stanford Center for Opportunity Policy in Education.

Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 313-346). New York, NY: Routledge.

Shermis, M. D. (2013). *Contrasting state-of-the-art in the machine scoring of short-form constructed responses*. Paper presented at the National Council on Measurement in Education, San Francisco, CA.).

ACKNOWLEDGEMENTS

This study was made possible with support from Jaison Morgan and Lynn Van Deventer of The Common Pool, LLC, and Dr. Mark Shermis of The University of Akron.

Design by Kelley Tanner.

APPENDIX A

Initial survey to identify the key information that vendors would need to understand in order to estimate pricing for machine scoring services.

November 14, 2012

To: Distribution list

Subject: ASAP contest – RFI

Automated Scoring Professionals:

The Assessment Solutions Group has been retained by the Hewlett Foundation, the PARCC and SBAC assessment consortia and the Board of Advisors for the ASAP project to write an RFI for pricing information related to automated scoring of student essays.

In order to write an RFI that is comprehensive and that vendors feel comfortable responding to the information requested, I would like to have a brief conversation with each of you. The conversation will provide an opportunity to discuss the type of information you think important to include in the RFI, the pricing information your company feels comfortable providing, the confidentiality we plan on providing related to the pricing information gathered and other relevant information. Accordingly, it may make sense for you to review some of this information with the appropriate people in your organization prior to the call.

Topics we will cover include:

1. Pricing components – item set up fees, scoring engine training fees, item per response fees, other fixed and variable fees
2. Item types – long response (>150 words), short response (<150 words) and different scoring engines for each
3. Essay types – source dependent (looking for a specific answer), open ended (tell me what you did on your summer vacation), items requiring inference (describe three facets of the Treaty of Versailles and how they led to World War II), other
4. Math items – discuss the types of math items that can and can't be scored with AI technology both today and in the near future (2014/2015). Discuss different engines used to score different math item types.
5. Volumes – impact of volume on scoring costs, locations, number of different items, N counts (volume discounts)
6. Confidentiality – discuss how your company's response may vary depending on the level of confidentiality in which we hold the information provided
 - a) Data only seen by ASG and the ASAP team – any reports will be blind as to vendor identification
 - b) Data only seen by ASG, the ASAP team and the consortia – any reports will be blind as to vendor identification
 - c) Data publicly available if requested
 - d) Other relevant information

I suspect that the calls will last somewhere between a half hour to an hour and would like to schedule them for one hour. I am fairly open as to timing. Can you review this information and let me know when we can schedule some time to talk? My contact information is below.

Thank you for your time and participation.

Sincerely,

Barry Topol
Managing Partner
CC: Dr. John Olson

APPENDIX B

Letters of support from assessment consortia



Partnership for Assessment of
Readiness for College and Careers

On behalf of the Partnership for Assessment of Readiness for College and Careers (PARCC), thank you for taking the time to complete this Machine Scoring Pricing Study: Request for Information (RFI).

PARCC is a consortium of 23 states plus the U.S. Virgin Islands working together to develop a common set of K-12 assessments in English and mathematics. These new K-12 assessments will build a pathway to college and career readiness by the end of high school, mark students' progress toward this goal from 3rd grade up, and provide teachers with timely information to inform instruction and provide student support.

The PARCC Vision

Our states have committed to delivering a K-12 assessment system that:

- Builds a pathway to college and career readiness for all students;
- Creates high-quality assessments that measure the full range of the Common Core State Standards;
- Supports educators in the classroom;
- Makes better use of technology in assessments; and,
- Advances accountability at all levels.

PARCC States

PARCC States educate about 25 million students and include 10 of the 12 Race to the Top winners. PARCC states include: Alabama, Arizona, Arkansas, Colorado, District of Columbia, Florida, Georgia, Illinois, Indiana, Kentucky, Louisiana, Maryland, Massachusetts, Mississippi, New Jersey, New Mexico, New York, North Dakota, Ohio, Oklahoma, Pennsylvania, Rhode Island, and Tennessee.

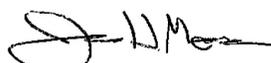
About This RFI

Your response to this RFI is critical in understanding how we can leverage the current capabilities of machine scoring to administer testing as cost effectively as possible. We intend to analyze your response to set cost assumptions and pricing expectations, as we publish various contracting announcements to you and others. Your cooperation is greatly appreciated. Thank you for taking the time and energy to respond.

Sincerely,

The PARCC Management Team:


Mary Ann Snider


James Mason


Laura Slover



The Smarter Balanced Assessment Consortium (Smarter Balanced) thanks you for taking the time to complete this Machine Scoring Pricing Study: Request for Information (RFI).

Smarter Balanced is a state-led collective effort working to develop next generation assessments that accurately measure student progress toward college and career readiness.

The Smarter Balanced Vision

The work of Smarter Balanced is guided by the belief that a high-quality assessment system can provide information and tools for teachers and schools to improve instruction and help students succeed – regardless of disability, language or subgroup. Smarter Balanced involves experienced educators, researchers, state and local policymakers and community groups working together in a transparent and consensus-driven process.

Smarter Balanced States

Our state partners educate more than 19 million of the nation's public K to 12 students. They include: Alabama, California, Connecticut, Delaware, Hawaii, Idaho, Iowa, Kansas, Maine, Michigan, Missouri, Montana, Nevada, New Hampshire, North Carolina, North Dakota, Oregon, Pennsylvania, South Carolina, South Dakota, Vermont, Washington, West Virginia, Wisconsin and Wyoming.

About This RFI

While Smarter Balanced has published other contracting opportunities which include requests for similar information, as part of the PARCC and Smarter Balanced consortia we support this specific request for information, to inform any deeper attention to the machine scoring of constructed response items and any implied pricing assumptions or cost expectations. We appreciate your response, which will be compiled into a summary analysis of the currently available requirements to administer machine scoring on state administered assessments.

Sincerely,

A handwritten signature in black ink that reads "Anthony Alpert".

Anthony Alpert, Chief Operating Officer
Smarter Balanced Assessment Consortium