THE EFFECTS OF CALCULATOR USE DURING NCCCO TESTING

International Assessment Institute

October 2011

 Ver. 1211F

EXECUTIVE SUMMARY

The National Commission for the Certification of Crane Operators (NCCCO) has asked for research on the topic of calculator use during its CCO certification examinations. This report was created using existing data and with inspection of previously published research and is organized by an introduction, literature review, methods, results and conclusion sections.

Summary of the literature review:

- Previous research has found negligible evidence that using calculators during examinations has an effect on the construct being tested.

- Research suggests that if there is an effect of calculator use, it is related to the type of math problem (computational or problem solving).
  - Computational problems tend to become easier with calculator use while problem solving items are not affected by calculator use.

Summary of the Methods and Results:

- Currently there are no NCCCO exams that are taken with the use of calculators and as such a direct effect of calculator use cannot be determined.

- This research attempted to inspect the effect of math questions on the examinations and from that infer if calculator use was appropriate.

- T-tests were conducted to analyze the differences between:
  - P-Values of the math items as compared to the rest of the items on the Core (Form 3002) exam.
    - The math items on the core are easier than the remainder of the exam with statistical significance.
  - P-Values of the load chart calculating items in the specialty exams as compared to the rest of the items on the specialty exams.
    - The load chart calculating items in each specialty exam are more difficult than the remainder of the exam with statistical significance in most instances.

- Correlations between math items on the Core and individual specialties revealed the relationship between scores.
  - There was a moderate positive relationship between the Core exam and all of the specialties, further indicating they are different types of items.

Summary of Conclusions:

This report reaches two main conclusions:

1. The use of calculators for the Core (Form 3002) exam **would not be** appropriate because the items are computational in nature and calculators would make them too easy.

2. The use of calculators for the specialty exams **would be** appropriate as (i) the math items are more difficult than the rest of the exam and (ii) a calculator would not help a candidate deduce the correct answer on load chart questions since there are multiple problem solving steps required for a correct answer.

It is recommended that current laws, regulations and standards be included in the decision-making process. If calculators are approved, consideration needs to be given to training of candidates on their use.

REPORT ON EFFECTS OF CALCULATOR USE DURING TESTING

INTRODUCTION

The question of whether the National Commission for the Certification of Crane Operators (NCCCO) should allow the use of calculators by candidates taking its certification examinations has been a topic of much discussion. In general, calculator use during testing has been a topic of interest in the assessment field for many reasons. The main reason is to ensure that calculator use does not confer an unfair advantage, thus decreasing the validity of the testing instrument. In essence, the question is: Does using a calculator change the meaning of the score on the test? If there were no effect from using calculators, NCCCO could review its current policies and consider allowing calculator use during examinations.

NCCCO strives to ensure the utmost safety in the operation of cranes. A very important safety aspect of operating a crane is the operator's ability to compute math in order to calculate the maximum load a crane can lift. If this calculation is done incorrectly, a serious injury, or even loss of life could occur. As part of promoting safety, NCCCO has included several items in its certification examinations that address a test taker's ability to perform load chart calculations.

NCCCO attempts to comply with best practices, industry standards and mandates when considering the use of calculators during assessment. Specifically, the ASME B30.5 Standard requires crane operators to be able to "…exhibit arithmetic and load chart/capacity usage..." – Section 5-3.1.2(b)(3). In the past the B30 Standards Committee has interpreted this requirement as not allowing calculators. More recently, the Committee has stated: "5-3.1.2(b)(3) does not specify what tools are to be used to demonstrate the ability to read, write comprehend and use arithmetic and a load capacity chart." OSHA in Subpart CC of the revised 29 CFR 1926 regulations, provides for the "use of and ability to calculate (manually or with a calculator) load/capacity information on a variety of configurations" – OSHA 1926.1427 (j) (1) (i) (B). This debate has increased interest in the possible effect of calculator use were they to be allowed.

Currently, no NCCCO crane operator assessments are being administered with the aid of a calculator, so therefore, a true calculator effect cannot be determined. However, math questions were added for the first time in the form of the Core exam released in August 2011 and the effects of adding these items can be determined. In theory, if the math items on the test were more difficult for the examinees than the remainder of the exam, it might be appropriate to permit the use of calculators. On the other hand, if the examinees appear to have an easier time getting the math items correct than the remainder of the test, then the use of a calculator during testing would not be appropriate as this would make the items even easier.

The questions this investigation addressed were:

1) What difference, if any, would there be in the Core (form 3002) exam if the math questions were included in the final score?
2) Is there a difference in the Core (form 3002) between math items and non-math items?
3) Is there a difference in each of the specialty exams between load chart calculation items and the rest of the exam?
4) How did the same candidates score on the Core math items versus specialty calculation items?

Current Status On Calculator Use

To date, the use of calculators during NCCCO certification exams has not been permitted even though some crane operator certification bodies do allow their use. NCCCO is not alone in its policy as several other licensing and certification organizations do not permit the use of calculators during assessments. Unfortunately, there appears to be no consistency in the assessment field regarding calculator use. Educational Testing Service (ETS), a company that administers many educational and industry examinations, allows the use of calculators on the Graduate Record Examination, but not on the PRAXIS exam for teacher certification.
The use of calculators in the Commonwealth of Pennsylvania for the professional engineers licensing exam is not allowed, but the state of Alaska permits them. To obtain an electrical license from the Department of Labor and Industry in Minnesota, calculator use is permitted, as it is during the exam to become a Certified Financial Planner or a Certified Safety Professional. The reasons behind the permission to use calculators during testing are not transparent and appear not to be related to the type of industry the exam is related to. Knowing that there is no consensus across the country allows NCCCO considerable latitude in any decision it makes in regard to its own calculator policies.

Factors that could affect the reliability and the validity of the current instruments should be understood prior to any policy changes.

Validity

Understanding what occurs in the field (in real life) in regard to calculator use needs to be considered in the decision making process. If calculators are currently used and readily available on job sites, calculator use during examination would increase the content validity of the test by overlapping what happens on the job and what is done in the test setting (Bing et al., 2009). This would then signify that the skills needed and used in the field would be tested more accurately.

One must also consider that, if calculators are not currently used in the field, or are not easily accessible to most operators, the validity of the test could be diminished. In this situation the examinees would not be testing in situations similar to those they would encounter when they operate cranes. Though this alone should not be a deciding factor, it deserves attention.

Another validity aspect to consider is the reason behind the math items on the test. If the purpose of the math questions is to examine the candidate's reasoning ability, allowing a calculator would not interfere with this cognitive domain. On the other hand, if the purpose is to examine the candidate's calculation skills, and if calculators are not used or permitted in real life situations, permitting a calculator during the examination might artificially increase test scores.

Reliability

A decision to change the policy on the use of calculators during testing might change the current reliability of the instruments. Though reliability is sample dependent and slight changes are expected when calculated with different samples, it is crucial that, if calculators were introduced, there be no statistically significant change in reliability.

In order to have good reliability, variability in the scores is essential. If all of the scores on the test are the same, then reliability would be very low. Allowing calculator use during testing might increase all the scores, thus potentially decreasing reliability.

Literature Review

Considerable research exists on the use of calculators for educational testing. However, there is less research on the effects of calculator use for industry related exams.

Many studies have attempted to discover the effects of calculator use on assessments, specifically for the service sector. Bing et. al., attempted to identify the effects of calculator use on tests that would be used for employment (2009). After administering two tests to the same population with and without the ability to use a calculator (investigating within person effects), they found no statistically significant decrease in the validity or the reliability of the tests. This study indicated that the use of calculators did not increase the test scores to a level of homogeneity that would decrease reliability. Further, the researchers established that the construct that was being tested did not change because of the use of calculators.

In this same study, the authors did find that questions that looked at the computational domain as opposed to the reasoning domain were answered correctly more often. There was an increase of correct answers for items that required calculating a solution as opposed to those items that required reasoning and a series of steps to get to the correct solution. This finding is an indication that there may be some effect of the use of calculators on assessments.

In fact, these results, indicating that calculator use increases scores on computational items without having much effect on problem solving items, have been replicated many times in research studies (Bridgeman, Harvey, & Braswell, 1995; Dye, 1981; Ellington, 2003; Lawrece & Dorans, 1994; Lloyd, 1991; Long, et al., 1989; Scheuneman & Camera, 2002; Schwarz, et al., 2002). This evidence indicates further that when a decision needs to be made in regard to calculator use during assessment, the type of item in question plays an important role.

Concluding thoughts on the review of literature are that calculator use does not appear to change the construct being tested (mathematical knowledge) and as such, validity of the instrument remains the same. Also, it appears that though there might be an increase in total scores because of the use of calculators, this increase does not statistically affect the reliability of the instrument. Any changes made to an instrument or how it is administered require individual inspection for verification of these statements.

Methodology

Data was received and inspected for normality as well as inaccuracies. Duplicates were identified and removed from the datasets. Scores on the original data were joined in one cell and were properly coded into one item per cell. Lastly, the pilot questions had no score and had to be recoded into binary responses.

Adequate power for the calculations was established by sample size. It was estimated that the difference between math items and the rest of the exams would have a medium effect size, one that was visible to the untrained eye. Cohen's Power Primer calculations inform that to have adequate power for a T-test a sample size of at least 51 people would be necessary. For all calculations conducted in this study this guideline was followed.

The Core exam was recently modified and released in August 2011 (N=2,491). This version of the core exam contains a total number of 90 questions of which six math items are embedded throughout the exam that were not in the previous version of the examination. The newly added math questions were not counted as a part of the final score and, as such, are considered pilot items. To answer the first research question *(What difference, if any, would there be in the Core exam if the math questions were included in the final score?)* these items were scored and added to the total of the test to see what difference having math items made on the overall test.

To answer the second research question *(Is there a difference in the Core exam between math items and non-math items?)* the item difficulties were inspected for comparison of the math questions in the Core (form 3002) exam against the remainder of the exam. An independent means t-test was used to calculate the difference between the math items, and the non-calculation problems were statistically significant at the 95% confidence level. It was important to understand if there was a difference between the two types of items on the test since, depending on which types of items were considered more difficult, a decision could be made regarding the potential use of calculators during examination.

To answer the third research question *(Is there a difference in each of the specialty exams between load chart calculation items and the rest of the exam?)* specialty certification exams were used for analysis. Only exams that had been taken by a sufficient sample size (more than 51 examinees) were used in this study. If not enough people had taken the exams, the item

difficulties would not be accurate or representative of what they might really be. The specialty exams that had appropriate sample size for analysis included:

- Manitowoc - Lattice Boom Crawler exam (*N*=646),
- Grove - Telescopic Boom Crane -swing- cab exam (*N*= 1,774),
- Link-Belt - Telescopic Boom Crane -swing- cab exam (*N*=508),
- Broderson Telescopic Boom Crane - fixed cab exam (*N*=573), and
- Manitex - Telescopic Boom Crane - fixed cab exam (*N*=1,475).

The analysis for the specialty exams was the same as that for the Core exam. All specialty exams follow the same format of 26 total items, with nine load chart questions at the end of the exams (question 18-26), six of which require calculations. The goal was to determine if load chart math calculation questions were significantly different from the other items on the test. Whether the calculation questions were considered easier or more difficult to answer correctly by examinees can provide an indication of the appropriateness of calculator use once the type of math item is inspected.

Analysis of the final research question consisted of computing new final scores for only the math items in both the Core and the specialty exams. These scores were then correlated to see the relationship between how a candidate scored on the math items of the Core (form 3002) exam math questions compared with the load chart math calculation questions in the specialty exams.

Results

Results from the Core exam demonstrate that 393 (15.8%) examinees failed the Core (form 3002) exam (*N*=2,491). When adding the math items into the test and recalculating the passing score, nineteen (19) more people would have passed, reducing the number of failing candidates to 374 examinees (15%). The passing rate of candidates would have increased from 84% to 85%. Adding the math questions into the total score makes the test easier to pass.

Table 1 - Results of t-tests

Exam Category	# of Exams	Mean P-value Math Items	Mean P-value Non-math	Statistically Significant?
Core (form 3002) - Pilot six math questions	2,491	91.83	78.60	Y

As stated earlier, NCCCO does not permit calculator use during examination. Because of this it was impossible to see the effect that using a calculator would have on individual items. Since a review of the literature indicates that the type of item makes a difference in the decision to use a calculator during examination, inspecting the differences between math items and non-math items in regard to how difficult the examinees found the questions was appropriate.

For the second research question *(Is there a difference in the Core exam between math items and non-math items?)*, inspection of the Core (form 3002) exam was done between the difficulties (p-

values) of items that contained calculations and items that did not. Cognitive inspection of these items revealed that they were more computational in nature and simply required adding subtracting or dividing. An independent means t-test was calculated inspecting for the assumptions. Data was within normal limits and there was no evidence that the assumption of homogeneity of variances was violated. The mean for the P-value of the math items in the Core (form 3002) exam was 91.83, while the mean for the P-value for the remainder of the items on the test was 78.60 as seen in Table 1. This difference was statistically significant indicating that the examinees found the math items easier than the non-math items with a 95% confidence interval.

Knowing that the math items on the Core (form 3002) exam were computational in nature and that the examinees found these items easier, it would not be recommended that calculators be allowed during examination of the Core (form 3002) exam. Allowing calculators to be used could, in fact, cause the reliability of the test to decrease by not having sufficient variance within the scores. Further, not allowing calculators on this exam would not appear to breach any standards and allows NCCCO the freedom to consider the possibilities of allowing calculator use during other certification examination.

Table 2 - Results of t-tests

Exam Category	# of Exams	Mean P-value Math Items	Mean P-value Non-math	Statistically Significant?
Manitowoc - Lattice Boom Crawler	646	67.67	79.35	Y
Grove - Telescopic Boom Crane swing-cab	1,774	52.78	79.18	Y
Link-Belt - Telescopic Boom Crane swing cab	508	60.00	83.06	Y
Broderson - Telescopic Boom Crane fixed cab	573	57.56	83.59	Y
Manitex - Telescopic Boom Crane fixed cab	1,475	76.67	80.53	N

The third question addressed whether there was a difference between the P-values of the math items on the load charts calculation items and the remainder of the specialty examinations. There were a total of five specialty exams that were inspected. Results of the t-test can be seen in table 2. All data was within normal parameters and there was no evidence that the homogeneity of variance assumption was violated. The P-value means of the math items were always lower than the means for the remainder of the assessments, even when the difference was not statistically significant.

For all of the specialty assessments, the load chart math calculation items were consistently considered more difficult by the examinees and as such answered correctly less frequently. This

is opposite from what was observed in the Core (form 3002) exam where the math items in the test were considered to be easier by the examinees and answered correctly more often.

Cognitive inspections of the math calculation items for the load chart questions revealed that these are not simply computational items, but that in fact they are higher order thinking questions. These questions do not simply involve a math calculation, but require the examinee to understand what the question is asking for, look in a table for the appropriate information that is needed, understand what items they will need from the aforementioned table, and then compute a math calculation. Incorrect answers from these questions may arise from the test takers not understanding the true nature of the problem, not knowing how to read the load chart tables, not pulling all of the deductions necessary from the table or not computing the math calculation appropriately.

Because of all of the steps in answering the question, and the complexity of the items themselves, the use of a calculator would not change the steps necessary to answer the item and might be considered appropriate during testing. Further, since the means of the p-values of these items are consistently lower than the means of remainder of the instruments, calculator use would only increase the p-values of these items if the reason for an inaccurate response was solely the calculation portion. Since these items require several steps to compute, the use of a calculator would further serve to discriminate between examinees that understand how to solve the problem and those who do not. Test takers who understand the problem and made a calculation error at the end would have a better probability of getting the item correct. Examinees who did not understand how to solve the problem would have no added benefit with the use of a calculator.

The fourth research question (*How did the same candidates score on the Core math items versus specialty calculation items?)* examined the relationship between correct answers on the Core (form 3002) exam math items and correct responses on the specialty exams math items. A correlation informs of the relationship of the scores. A person who gets a high math score on the Core and a high math score on a specialty would have a high positive correlation. An examinee that obtained a low score on the Core and a low score on the specialty would also have a high positive correlation. Negative correlations would indicate examinees had high scores on one test and low scores on the other.

Results from the correlation can be seen in Table 3. All exams have a small to moderate positive correlation. This implies that examinees that scored the math items on the core correctly also scored the math items on the specialty exams correctly some of the time. The inverse is also true.

Table 3 - Table of Correlations

	Grove	Broderson	Manitex	Manitowoc	Link Belt
Core	.368	.329	.384	.296	.301

Since there are no strong correlations between these scores, this is yet another indication that the types of items math in the Core exam and the specialty exams are different in nature and should be treated differently when making decisions.

Conclusions

There are two types of math items used in the exams. The first kind is the calculation problem that requires only computation of basic math skills such as adding, subtracting, dividing, and multiplying. The second type of math question on the specialty examinations involves higher order thinking and requires several steps to solve.

Taking into account what other researchers have found in their studies, and consideration of the results of the differences in difficulties between math items on a test and the remainder of the test, a decision can be made regarding calculator use during examinations.

In general, it is not recommended that calculators be allowed during the Core (form 3002) examination. The math items in this examination are considered to be easier than the remainder of the test and they are computational type items. The use of a calculator would make these items even less of a challenge.

In contrast, a calculator might be deemed appropriate for use during the specialty examinations. The math load chart calculation questions require many more steps than a simple calculation of addition or subtraction. The use of a calculator will not assist an individual who does not know what steps should be followed to answer the question.

However, even though applicants for this examination may be competent in calculator use, if a decision were to be made to allow calculators during assessment, a brief training or orientation session prior to the test should be considered (Hembree & Dessart, 1986; Roberts, 1980).

If the decision to allow calculators in testing is made, it is suggested that a pilot study be administered that observes the effects of examinees taking the assessment with and without the aid of a calculator. The results could then be compared and inspected for statistical difference. If none were found this would provide further evidence that allowing calculator use during testing was an appropriate and prudent decision.

References

Bing, M. N., Stewart, S. M., & Davison, H. K. (2009)
An investigation of calculator use on employment tests of mathematical ability: Effects on reliability, validity, test scores, and speed of completion. *Educational and Psychological Measurement, 69*, 322-350.

Bridgeman, B., Cline, F., Levin, J. (2008)
Effects of Calculator availability on GRE quantitative questions. (Report No. RR-08-31). Princeton, NJ: Educational Testing Service.

Bridgeman, B., Harvey, A., & Braswell, J. (1995)
Effects of calculator use on a test of mathematical reasoning. *Journal of Educational Measurement, 32*, 323-340.

Brooks, T., Case, B. J., Cerrillo, T., Severance, N., Wall, N., & Young, M. J. (2003)
Calculator use on Stanford Series Mathematics Tests. (Assessment Report). San Antonio, TX: Pearson.

Dye, D. L. (1981)
The use and non-use of calculators on assessment testing. St. Paul, MN: Minnesota State Department of Education.

Ellington, A. J. (2003)
A meta-analysis of the effects of calculators on students' achievement and attitude levels in precollege mathematics classes. *Journal of Research in Mathematics Education*, *34*, 433-463.

Hembree, R., & Dessart, D. J. (1986)
Effects of hand-held calculators in precollege mathematics education: A meta-analysis. *Journal of Research in Mathematics Education*, 17, 83-99.

Lawrece, I., & Dorans, N. J. (1994)
Optimal use of calculators on a mathematics test: Effect on item difficulty and score equating. (Report No. RR-94-40). Princeton, NJ: Educational Testing Service.

Lloyd, B. H. (1991)
Mathematics test performance: The effects of item type and calculator use. *Applied Measurement in Education*, 4, 11-22.

Long, V.M., Reys, B., & Osterlind, S. J. (1989)
Using calculators on achievement tests. *Mathematics Teacher*, 82, 318-325.

Roberts, D. M. (1980)
The impact of electronic calculators on educational performance. *Review of Educational Research, 50*, 71-98.

Scheuneman, J., & Camera, W. J. (2002)
Calculator use and the SAT I Math. (Report No. RN-16). New York, NY: College Board.

Schwarz, R., Rich, C., Arenson, E., Podrabsky, T., & Cook, C. (2002)
An analysis of functioning based on calculator type. Paper presented at the National Council on Measurement in Education, New Orleans, LA.