Validity and diagnostic attributes of a Mathematics Olympiad for Junior High School Contestants

Johann Engelbrecht* and Jeanine Mwambakana

University of Pretoria, South Africa

Corresponding author Johann Engelbrecht, email johann.engelbrecht@up.ac.za

Abstract

The purpose of mathematics competitions, and in our case the South African Mathematics Olympiad (SAMO), is to promote problem solving skills and strategies, to generate interest and enthusiasm for mathematics and to identify the most talented mathematical minds. SAMO is organised in two divisions – a junior and a senior division - over three rounds. We analysed the results of the junior second round over seven years 2006-2012. Based on the literature a mathematical content framework was developed, dividing the mathematical content into seven broad content areas. In this paper we investigate the face validity, diagnostic attributes and predictive criterion validity of mathematics olympiad question papers over the period by focussing on the frequency of content area occurrence in the different items. We also look at performance of contestants in the different content areas as a broad diagnosis. Lastly we investigate the item performance, comparing the expected performance by the problem committee of experts setting the question papers and the actual performance of contestants. Topics such as numbers, algebra, patterns and functions, measurement, applications, modelling and logic were used abundantly whereas (quite surprisingly) there were few items on graphs, decimal fractions, spatial logic and vertices and edges of polygons, indicating that the face validity can be improved. Contestants performed best in items on algebra and weakest in items on statistics. The ability of the problem committee to anticipate student item performance varied considerably and was significantly worse in 2012 than in 2006, indicating better predictive criterion validity in 2006.

Keywords: mathematics competitions, mathematics olympiads, face validity, predictive criterion validity, diagnostic attributes

Background

School mathematics in several countries has moved closer to mechanical calculation or numeracy (Taylor, 2008). However, creative problem solving skills are crucial to be competitive in the global market of mathematics-based careers (Kenderov, 2006). Across the world a shortage exists of young people taking up mathematics related careers, often caused by a negative attitude to the subject. According to the literature mathematics competitions may address both issues as they do not consist of a collection of routine tasks to be executed, but purposely emphasise that mathematics is about creative thinking and the development of problem solving methods (Kahane, 2009; Kenderov, 2006). There is some evidence that participation in mathematics competitions improves performance in school mathematics (Bicknell & Riley, 2012) and attitudes towards mathematics as a subject (Gyöngyösi, 2002; Bicknell & Riley, 2012). In order for mathematics competitions to also identify ability, the competition instruments need to be valid, i.e. measure ability in creative thinking and problem solving.

Little research has been done on the impact and efficiency of mathematics olympiads and various questions arise about mathematics competitions (Somers & Callan, 1999; Gleason, 2008). Issues that justify investigation include gender issues (boys generally perform better in olympiads than girls), the impact that exposure to competition mathematics has on university preparedness, the diagnostic strengths of olympiad papers and the validity and reliability of the question papers in mathematics competitions.

In this study we investigate the validity and diagnostic attributes of mathematics olympiad question papers - do they measure what they are supposed to measure. The investigation uses the South African Mathematics Olympiad as a case study. It focuses on how the spectrum of relevant content areas in the South Africa mathematics curriculum is covered in the olympiad question papers; how performance in the different mathematics content areas compares; and how well the experts setting the question papers predict the level of difficulty of the different items, i.e. at an appropriate level and with appropriate progression of difficulty.

Gleason (2008) evaluated mathematics competitions using item response theory. His analysis showed that the multiple-choice format in the mathematics competitions provided sufficient information to discriminate ability levels of contestants, but that the most valid information is provided for discriminating between participants whose ability levels are near the mean. This would imply that this format can best be employed as an initial round of the competition to reduce the number of contestants.

The South African Mathematics Olympiad (SAMO)

The South African Mathematics Olympiad (SAMO), organised by the South African Mathematics Foundation (SAMF) (http://www.samf.ac.za/Default2.aspx) is the biggest mathematics olympiad in the country. The olympiad has been running since 1966. Participation has grown from 5 234 contestants in the first round of the first event to more than 86 000 contestants who participated in the 2016 olympiad. The olympiad involves high school students and consists of a junior division (grades 8-9) and a senior division (grades 10-12).

A new structure with three rounds came into operation in 1992. A separate second round paper for juniors was introduced in 1994. A third round for juniors was introduced in 2004. The first round is written in March every year and the first round for the junior division consists of separate papers for grades 8 and 9. In this round schools are provided with the solutions - the teachers mark the papers and send the marks to the SAMF office. Consequently no detailed records of the first round answers are available.

Contestants who attain 50% or higher in the first round qualify for the second round. This time the grade 8 and 9 contestants write the same paper. Contestants have two hours to complete twenty multiple-choice questions. Second round answer sheets are sent to the SAMF office where they are marked electronically. In this round comprehensive data about contestants' answers are available. For this reason our study focuses on an analysis of the second round answers specifically for the junior division. The best 100 senior and junior contestants from the second round qualify for the third round. The junior third round paper consists of 15 open ended problems which have to be completed in four hours.

The objective of SAMO is not only to find winners. More important is mass participation. The aim of the SAMO is to promote mathematics as a subject and all learners are

encouraged to take part. The organisers believe that learners who take part in the olympiad benefit from the exposure to mathematics going beyond the curriculum which assists them to think out-of-the-box.

Mathematics competitions worldwide

The main goal of mathematics competitions is to enrich the study of mathematics. Although inspiring for the better students, tasks can be developed at different levels, also allowing average students to be exposed to the various benefits of competitions (Gyöngyösi, 2002). Bright students need challenges to keep their minds actively focussed on mathematics and prevent them from moving to endeavours outside mathematics they may find more appealing. Students of every level, background, ability or motivation should be challenged, not only bright students (Gyöngyösi, 2002). For students with less motivation, challenging mathematical tasks can serve to attract them to mathematics learning rather than to the mastery of algorithms or routine methods. However, it has been found that even the learning of routine material is improved when taking place in a challenging environment (Barbeau & Taylor, 2009). Rather than focusing on a small group of winners, broad participation in competitions is more important since by preparing for the competition and trying to solve the problems during the competition itself, all participants increase their knowledge significantly (Kenderov, 2006).

Competitions have been considered as elitist even though there has been a growth in participation in competitions in recent years, e.g. the European Kangaroo (Taylor, 2008). Gender issues have also been raised by critics - there is evidence that boys are more successful than girls (Niederle & Vesterlund, 2010). A closer examination reveals that a gender gap in mean scores is so small so as to be of little practical importance, even though the gender gap in the upper tail can be quite large (Niederle & Vesterlund, 2010). However, even for the high achievers Desjarlais (2009) established that after controlling for ability no statistically significant gender differences in competition performance are evident. There are also arguments that competitions provide unnecessary pressure, stress and feelings of failure from excessive competitiveness (Davis, Rimm & Siegle, 2014).

In spite of these concerns, there is a significant body of committed competition supporters worldwide who highlight many advantages (Gyöngyösi, 2002; Bicknell & Riley, 2012). Competitions stimulate interest in mathematics. Questions are often set in real world situations to which the students can relate, rather than pure mathematical situations. The impact competitions such as the European Kangaroo with more than 3 million participants, have, is difficult to overestimate (Kenderov, 2006). Probably the most important advantage of competitions is that they fill a gap in the curriculum, providing an opportunity for students to be exposed to real problem solving and to appreciate some of the aesthetically pleasing parts of mathematics. Taylor (2008) appreciates the variety of mathematical approaches in competitions in comparison with normal classroom assessment tasks in school that are becoming more and more predictable. Kahane (2009) argues strongly for mathematics competitions since they particularly lend themselves to free investigation, imagination and creative activities. Bicknell and Riley (2012) plead for competitions to be acknowledged in school policy as part of the official mathematics programme and that equitable opportunities should be provided for students to participate in mathematics competitions.

Kenderov (2006) sees competitions as providing a tool to identify and develop students with higher abilities and talent who do not experience any challenge in the standard curriculum

and their mathematical abilities and talent then remain undiscovered and undeveloped.

Performance in mathematics competitions does not always correlate with classroom performance (Ridge & Renzulli, 1981). However, experiences in competitions and related activities improve the preparation of the student for university study. Specifically, Taylor (2008) mentions the fact that many former olympiad participants have become research mathematicians.

The social impact of competitions is also mentioned by Bicknell and Riley (2012) and Kenderov (2006). Mathematics enrichment activities can be viewed as events generating discussions among the students, since competition problems can often be solved in more than one way thus provoking discussions. These informal social interactions might be as important as participation in the competition itself for acquiring new mathematical knowledge. These social interactions can happen in preparation for the competition, working on problems from previous competitions, or sharing after a competition.

One of the important benefits of participating in mathematics competitions is the exposure learners get to problem solving. Problems used in problem solving create a challenge for the student, which occurs when there seems to be no standard method of solution. You have to reflect and analyse the situation, possibly bringing together some diverse factors (Barbeau & Taylor, 2009). Although the ultimate objective is to meet the challenge, i.e. to solve the problem, the process of grappling with its difficulties often results in better understanding, new insights and a sense of personal power:

> The joy of confronting a novel situation and trying to make sense of it - the joy of banging your head against a mathematical wall, and then discovering that there may be ways of either going around or over that wall (Olkin & Schoenfeld, 1994, p. 43).

Students who can handle unexpected situations and solve new problems are in great demand. Problem solving as part of mathematics has been reported on frequently in literature (e.g. Cai, 2003). Lester and Cai (2010) define problem solving as "mathematical tasks that have the potential to provide intellectual challenges for enhancing students' mathematical understanding and development." (p. 1).

With overfull current mathematics curricula in South Africa, teachers tend to focus on "technical" mathematics and recipe–driven manipulations get preference. The syllabus is contracted to manipulation skills only, leaving little time for using these skills in various ways to solve problems in everyday life (Taylor, 2008). Performance in examinations has become almost too important, causing teachers to spend any available time on examination coaching. This approach again does not promote problem-solving activities. Formulas tend to hide the real content of concepts and to create stereotypes (Gyöngösi, 2002). A stronger focus on problem solving not only contributes to the development of students' higher-order thinking skills but also improves positive attitudes towards mathematics. (Lester & Cai, 2010).

Setting olympiad question papers

SAMO's question papers for mathematics competitions – as with most academic competitions – are set by problem committees of 5-7 people comprising some senior university academics (mathematics professors), some former olympiad contestants and some

school teachers. The latter are able to judge whether the cognitive level of the questions is appropriate for the contestants. We will refer to all problem committee members as *experts.*

Research on the way competition papers are set, and the validity and reliability of olympiad papers is almost non-existing - this points to a serious research need. In relating his experience as member of problem solving committee, Miguel (2012) highlights the dilemma of developing problems that are both beautiful and have the right difficulty. To address the face and criterion validity requirements, rather than developing individual problems, a set of problems is developed that are balanced with regard to content areas and difficulty.

The working procedure for problem committees is fairly standard (for instance, see Kenderov, 2006). Committee members would come to committee meetings well prepared. Each member has to contribute a few problems that can serve as possible items for the eventual question paper. New problems are designed by committee members consulting a variety of sources, including question papers of other international mathematics competitions. They would often get ideas from existing problems to develop a "new" problem.

At the SAMO problem committee meetings (twice a year) members work through the available problems and specifically focus on the following validity issues:

- Face validity: A representative coverage of the mathematics content.
- Predictive criterion validity: Coverage across different levels of difficulty and with progression of difficulty from the beginning to the end of the paper.

Items in the SAMO test are ranked from 1 to 20 according to the level of difficulty as perceived by the problem committee who sets the papers. Items 1-5 are considered to be easy or accessible to most contestants (attracting 4 marks each); items 6-15 are considered to be moderately difficult (for five marks each) and items 16-20 are perceived to be difficult (for six marks each). In our analysis we weighted items according to their levels of difficulty, using the number of marks contributed to the total. This means that in finding the average performance for a particular content area, the level of difficulty was accounted for, i.e. more difficult questions contributed more to the average than easier items.

Validity and reliability form an over-arching backdrop during the entire process. After the committee has reached agreement on the final paper, an external moderator would work through the paper to ensure that the question paper complies with the validity issues mentioned.

Research questions

The main purpose of this study is to investigate the validity and the diagnostic attributes of mathematics olympiad question papers. The study focuses on three issues:

- How well do the question papers cover the spectrum of mathematics content areas relevant to the contestants' level of mathematical development? (Face validity)
- How does performance in the different mathematics content areas compare, and what is the variation in performance in the different content areas over time? (Diagnostic attributes)

- How well does the problem committee of experts predict the level of difficulty of the different items? Are the question papers set at an appropriate level and with appropriate progression of difficulty? (Predictive criterion validity)

Mathematical content framework

In the USA the NAEP (National Assessment of Education Progress) report is a national data source for achievement by learners (4^{th}, 8^{th} and 12^{th} grades) in mathematics and other subjects (Neidorf, Binkley, Gattis & Nohara, 2006). Their assessment framework is based on the collaborative input of a wide range of experts from government, education and business. Similarly, the development of the assessment framework for TIMSS (Trends in International Mathematics and Science Study) involves mathematics experts and education professionals from many countries (Mullis, Martin, Ruddock, O'Sullivan & Preuschoff, 2009).

Both the NAEP and TIMSS mathematics frameworks have five content areas in the content dimension: numbers; measurement; geometry; data; and algebra. Looking at cognitive dimensions, NAEP has three categories (conceptual understanding; procedural knowledge; and problem solving), whereas TIMSS has four categories (knowing facts and procedures; using concepts; solving routine problems; and reasoning).

In South Africa the recently introduced CAPS (Curriculum Assessment Policy Statement) mathematics learning areas for grades 8-10 include numbers, operations and relationships; patterns, functions and algebra; space and shape (geometry); measurement; and data handling (statistics). The cognitive distinction (mathematical ability) is between knowledge; routine procedures; complex procedures; and problem solving (DBE, 2011).

The content of SAMO over the years has been close to the South African national curriculum but with a somewhat different focus on subtopics. For this study an assessment framework with seven content areas was developed by the authors using the CAPS framework as well as the TIMSS and NAEP frameworks as a basis. We refer to this framework as the *MANGSLO* classification:

M. Measurement, applications, modelling
A. Patterns, functions and algebra
N. Numbers, operations and relations
G. Geometry, space and shape
S. Statistics, data handling
L. Logic
O. Others

Essentially the MANGSLO framework is the same as the CAPS, NAEP or TIMMS frameworks, but because of the nature of competition mathematics, we added *logic* as a separate content area. Each of these main content areas was divided into subtopics in the following classification. We illustrate some of the subtopics with an example from one of the question papers. The actual test papers consisted of multiple-choice questions (MCQ) but for the purpose of this description the MCQ format is dropped.

M: Measurement, applications and modelling

M1: Rate (speed), distance and time

Example: *John cycles* 10 *km/h faster than Dave, and takes one third of the time that Dave takes. They both cover the same distance. Calculate Dave's speed in km/h.*

M2: Date and clock arithmetic
M3: Length and perimeter
M4: Mass, area and volume

A: Algebra, patterns and functions

A1: Patterns and sequences
Example: *What are the last two digits of* 7^{2009}?
A2: Substitution and manipulations
Example: *If x, y and z are real numbers such that* $(x-3)^2 + (y-4)^2 + (z-5)^2 = 0$, *determine x + y + z.*
A3: Algebraic expressions
A4: Linear equations
A5: Simultaneous equations
A6: Graphs

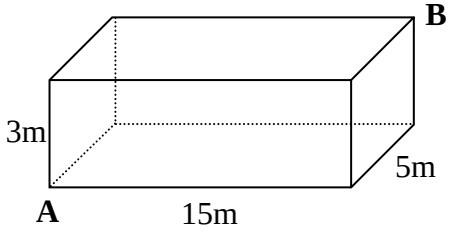
N: Numbers, operations and relations

N1: Calculation with numbers
Example: *Calculate the value of* $1 - 4 + 9 - 16 + 25 + \ldots + 625$.
N2: Multiples and factors
N3: Order of operations
N4: Properties of numbers
Example: *What is the smallest positive integer which must be added to 2009 in order to get a perfect square?*
N5: Powers and exponents
N6: Averages
N7: Approximations
Example: *Give an estimate of* $\frac{2008 \times 1710}{3421}$
N8: Common fractions
Example: *Determine the value of* $1\frac{1}{2} \times 1\frac{1}{3} \times 1\frac{1}{4} \times \ldots \times 1\frac{1}{19}$
N9: Decimal fractions
N10: Percentages
N11: Ratio and proportion
Example: *If the ratio x : y is* 3 : 4 *and the ratio y : z is* 3 : 5, *then what is the ratio x : z?*
N12: Magnitudes of numbers
N13: Working with digits

G: Geometry, space and shape

G1: 2D shapes
G2: 3D shapes
G3: Vertices and edges
G4: Angles
G5: Theorem of Pythagoras

Example: *The diagram shows a long room. An ant wants to walk from A to B. It can walk along the walls and ceiling of the room. What is the shortest distance it could walk*?



S: Statistics and data handling

S1: Counting

Example: *How many four-digit multiples of* 9 *are there in which all the digits are odd and distinct?*

S2: Statistics and probability

L: Logic

L1: General logic

Example: *I recently returned from a trip. Today is Friday. I returned four days before the day after tomorrow. On which day did I return*?

L2: Spatial logic

O: Others

O1: Miscellaneous

Example: *The digits of* 20098 *can be arranged in any order. For each arrangement, the 'score' is the sum of the positive differences between successive digits. What is the maximum score that can be achieved?*

Research design

This study investigates the second round of the junior SAMO test, compiled by a problem committee of experts as described earlier. Contestant responses for seven years (2006-2012) were considered. The number of contestants in each year is given in Table 1. These numbers depend on how many contestants made it through round 1 of the competition. Contestants need a mark of at least 50% in round 1 to qualify for round 2.

Table 1: Numbers of contestants in different years in the junior SAMO second round

Year	Number of contestants
2006	3817
2007	2851
2008	2281
2009	3508
2010	2052
2011	5813
2012	4142

All 140 items in these seven question papers were allocated to one or more of the mathematical topics in the MANGSLO classification scheme mentioned in the previous section. The classification was done independently by the researchers, assisted by members of the problem committee, and topic allocations were then discussed. There was a remarkable agreement between different raters, indicating a high inter-rater reliability.

Some items fall into more than one content area and such items were considered in all content areas in which they were classified as illustrated by the following item.

> *What is the smallest value of n such that the product n! = 1×2×3×⋯× n ends in at least 10 zeroes?*

This item was considered to belong to three topics, N4 (Properties of numbers), L1 (General logic) and N5 (Powers and exponents).

To compare expected with empirical performance, the question papers of two years, 2006 and 2012, were considered as case studies. The anticipated level of difficulty as envisaged by the experts who set the question papers was compared with the actual performance of contestants in the papers.

Results

Frequency of topics (face validity)

Since some items were classified into more than one topic, rather than 140 (7 papers of 20 items each) there were 245 item classifications. The frequency distribution of item classifications in the main- and sub-content areas (over the seven years) is given in Table 2.

Table 2: Frequency of item occurrence in different topics (n = 245)

Content area		**Number of occurrences**	**Percentage of Occurrences**
M: Measurement, applications, modelling		**40**	**16.3%**
M1	Rate (speed), distance and time	8	3.3
M2	Date and clock arithmetic	1	0.4
M3	Length and perimeter	12	4.9
M4	Mass, area and volume	19	7.8
A: Algebra, patterns and functions		**52**	**21.2%**
A1	Patterns and sequences	16	6.5
A2	Substitution and manipulations	5	2.0
A3	Algebraic expressions	11	4.5
A4	Linear equations	12	4.9
A5	Simultaneous equations	8	3.3
A6	Graphs	0	0.0
N: Numbers, operations and relations		**89**	**36.3%**
N1	Calculation with numbers	9	3.7
N2	Multiples and factors	7	2.9
N3	Order of operations	4	1.6
N4	Properties of numbers	17	6.9
N5	Powers and exponents	12	4.9
N6	Averages	1	0.4
N7	Approximations	4	1.6
N8	Common fractions	3	1.2
N9	Decimal fractions	0	0.0
N10	Percentages	6	2.4
N11	Ratio and proportion	15	6.1
N12	Magnitudes of numbers	2	0.8
N13	Working with digits	9	3.7
G: Geometry, space and shape		**18**	**7.3%**
G1	2D shapes	5	2.0

G2	3D shapes	1	0.4
G3	Vertices and edges	1	0.4
G4	Angles	5	2.0
G5	Theorem of Pythagoras	6	2.4
S: Statistics and data handling		**16**	**6.5%**
S1	Counting	10	4.1
S2	Statistics and probability	6	2.4
L: Logic		**25**	**10.2%**
L1	General logic	24	9.8
L2	Spatial logic	1	0.4
O: Others		**5**	**2.0%**

The most popular sub-topics were L1 (General logic) with 24 item occurrences, M4 (Mass, area, volume) with 19 occurrences and N4 (Properties of numbers) with 17 occurrences. It is surprising that there were no items in two sub-topics, i.e. A6 (Graphs) and N9 (Decimal fractions). The sub-topics M2 (Date and clock arithmetic), L2 (Spatial logic) and G3 (Vertices and edges) were also not popular with only one item occurrence each over the seven years.

Considering face validity, these results can be compared to the CAPS weighting of content areas as in Table 3 - the CAPS weighting is the average between the prescribed weightings for grades 8 and 9 in DBE (2011).

Table 3: Comparison of weighting of content areas in SAMO and CAPS

Content area	SAMO	CAPS
M: Measurement, applications, modelling	16.3%	10.0%
A: Algebra, patterns and functions	21.2%	32.5%
N: Numbers, operations and relations	36.3%	20.0%
G: Geometry, space and shape	7.3%	27.5%
S: Statistics and data handling	6.5%	10.0%
L: Logic	10.2%	
O: Others	2.0%	

As is clear from Table 3, amongst the SAMO items content areas A (Algebra, patterns, functions) and G (Geometry, space and shape) are under-represented as compared to the school curriculum. In contrast, content areas such as M (Measurement, applications, modelling) and N (Numbers, operations and relations) are over-represented. The face validity of the SAMO question papers can at best be described as *moderate.*

Performance in content areas (diagnostic attributes)

To address the question relating to the diagnostic value of the question papers over the seven years, we compare performance in different content areas over the seven years. As explained earlier, the average performance in the different items was weighted according to the expected level of difficulty. Table 4 shows the weighted performance in the different content areas over the entire period of seven years together with the overall weighted performance over the entire period.

Table 4: Weighted performance in the various content areas for 2006 -2012

Content area	Weighted performance (%)							
	2006	2007	2008	2009	2010	2011	2012	Overall
M: Measurement, applications, modelling	40.3	43.6	26.4	43.6	40.7	35.1	15.4	34.9
A: Algebra, patterns, functions	35.6	35.9	36.7	44.3	34.0	29.4	58.0	39.2
N: Numbers, operations, relations	34.3	28.7	29.0	36.5	20.3	36.7	32.2	31.1
G: Geometry, space and shape	60.0	28.4	16.4	29.0	26.2	47.6	31.4	34.1
S: Statistics and data handling	25.2	3.8	21.9	13.8	18.4	24.4	15.6	17.6
L: Logic	47.0	14.7	29.2	21.0	28.1	43.2	38.7	31.6

The last column in Table 4 shows that contestants performed best in A (algebra, patterns, functions), and performed relatively poorly in S (Statistics and data handling).

Table 4 also indicates that student performance in M (Measurement, application, modelling) does not show any real trend over the years, but there was relative poor performance in 2008 and 2012. In contrast, performance in A (Algebra, patterns and functions) was fairly consistent over the seven years, with 2012 showing the highest peak (58%) and 2011 the relatively low weighted average of 29%. In terms of student performance in N (Numbers, operations, relations) Table 4 shows no real trend over the seven years, but there were bad years (2007, 2008 and 2010) where the weighted performance of students was below 30%. Equally, student performance in G (Geometry, space and shape) does not show any real trend over the seven years, with four bad consecutive years (2007 – 2010) when the weighted performance of students was below 30%. As already identified from the overall performance, student performance in S (Statistics and data handling) was consistently poor, with 2006 showing the highest peak (25.2%) and 2007 the lowest weighted average of only 3.8%. Lastly, student performance in L (Logic) does not show any definite trend over the seven years, but in 2007 and 2009 in particular, the weighted performance of students was low.

Comparison between expected and actual performance (predictive criterion validity)

The problem committee of experts sets the papers, grades the items with the level of difficulty increasing from Item 1 to Item 20. As a case study the question papers for 2006 and 2012 were analysed. The actual performance in the items of 2006 and 2012 (no weighting) is presented in Figures 1 and 2 respectively. The X-axis shows the number of the item in the question paper – indicating the expected level of difficulty. The numbers on the Y-axis represent the actual percentage of contestants who got the item right. So if the correspondence between expected and empirical performance is good, one would expect this graph to decrease from Item 1 to Item 20.

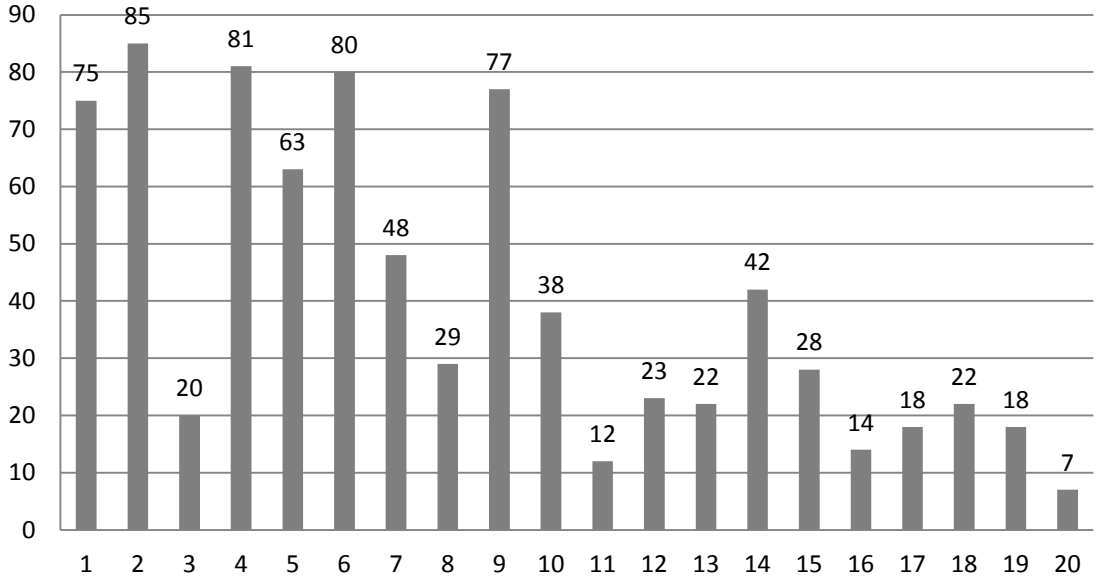


Figure 1: 2006 comparison between expected and actual performance

Table 5 compares the expected performance with the empirical performance by the students for the two years 2006 and 2012. The experts considered items 1-5 to be easy or accessible to most contestants, items 6-15 are classified as moderately difficult and items 16-20 as difficult.

According to Table 5, the actual student performance for 2006 shows that the easy items are (in this order) items 2, 4, 6, 9 and 1. In this category of questions, items 6 and 9 – planned as moderately difficult items - were found easy by students whereas item 5 and 3 were experienced as moderately difficult by the students with item 3 occupying 15th position. Items 5, 7, 14, 10, 8, 15, 12, 13, 18 and 3 were experienced as moderately difficult by the students. According to the problem committee, beside items 6 and 9 that did not make it in this category, item 11 also is missing, it was perceived difficult by the students and item 18 in this category should be in the last category. Finally, items 17, 19, 16, 11 and 20 were experienced as difficult by the students. Except for Item 11 these agree with how the problem committee planned the paper. So for 2006 the overlap between expected and empirical difficulty is fairly high in all three the groupings of items.

Table 5: Comparison of expected difficulty and empirical difficulty

	Design of the test items ranked from less difficult to greater difficulty	Empirical ranking of items ranked from less difficult to greater difficulty		Overlap between expected and empirical difficulty	
		2006	2012	2006	2012
Easy	1→2→3→4→5	2→4→6→9→1	16→6→4→13→1	60% (3/5)	40% (2/5)
Moderate difficulty	6→7→8→9→10	5→7→14→10→8	11→8→15→7→9	80% (8/10)	60% (6/10)
	11→12→13→14→15	15→12→13→18→3	17→2→14→19→3		
Difficult	16→17→18→19→20	17→19→16→11→20	5→18→20→12→10	80% (4/5)	40% (2/5)

The Spearman rank correlation coefficient was calculated between the rankings as made by the experts and the actual performance. For 2006 this correlation coefficient is 0.73, which on the sample of 20 items, is significant on a 0.01 level.

For 2012 the picture looks slightly worse. From the students' performance, it can be seen (Figure 2, Table 5) that items 3, 2, 14, 19, 17, 9, 8, 15, 7 and 11 were experienced as moderately difficult. Items 3 and 2 are supposed to be in the "easy" category and item 19 and 17 were planned to be in the "difficult" category according to the problem committee. Finally, items 5, 18, 20, 12 and 10 were perceived as difficult by the students. Item 5 is the biggest surprise as it is expected to be accessible to most students (easy items category) and items 12 and 10 were intended to be moderately difficult by the problem committee that set the paper. So for 2012 the overlap between expected and empirical difficulty is poor for all three item groupings.

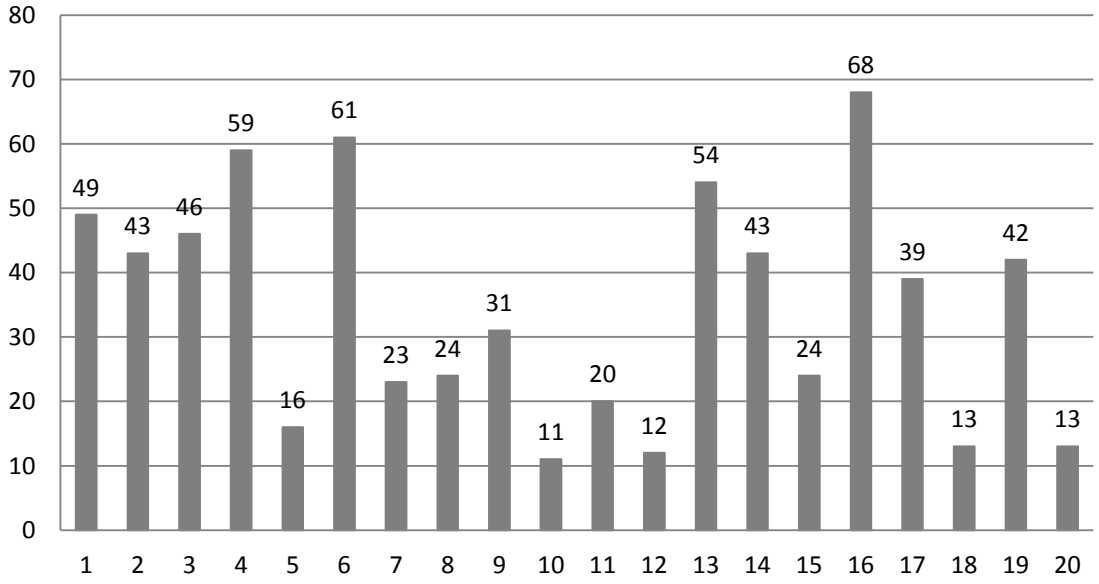


Figure 2: 2012 comparison between expected and actual performance

The Spearman correlation coefficient between expected and actual level of difficulty in 2012 is 0.29. On the sample of 20 items, this correlation is poor and not significant. From Table 4 the low value of the 2012 coefficient is clear. The expected and actual performance of students vary greatly: easy items according to contestants' performance were: 16, 6, 4, 13 and 1. In this group, item 16 belongs to difficult category and items 6 and 13 to the moderately difficult category as per the problem committee's classification.

Discussion and conclusions

The objective of this study was to investigate the validity and the diagnostic attributes of mathematics olympiad question papers. We focussed on face validity - how the mathematics spectrum is covered regarding the different relevant content areas in mathematics; on diagnostic attributes - how performance in the different mathematics content areas compare; and on predictive criterion validity - how well the experts setting the question papers predict the level of difficulty of the different items, i. e. at an appropriate level and with appropriate progression of difficulty.

If we compare the frequency of items in the different content areas with what is suggested in the CAPS curriculum, there is a clear difference in representation, indicating a face validity

that is moderate at best. However, the objective with an olympiad is somewhat different from the aims in the standard curriculum. As mentioned earlier, CAPS provides for four different cognitive levels, knowledge, routine procedures, complex procedures and problem solving. In an olympiad the focus is on problem solving. This may result in expecting different weightings for the various content areas compared to the standard curriculum. Content areas such as N (Numbers, operations and relations), A (Algebra, patterns and functions), M (Measurement, applications, modelling) and L (Logic) were well represented in the items over the seven years of the study. On the other hand, it is quite surprising that there were no items on graphs and on decimal fractions at all and that there were almost none on spatial logic and vertices and edges of polygons.

By considering the analysis of the frequency and performance of the different content areas in the second round question papers over the seven years, the question immediately arises whether these data can be interpreted as a spontaneous indication of the relative importance of these content areas or whether this indicates that the committees setting the problems should try to be more balanced with regard to addressing different topics in mathematics. We recommend that the problem setting committee consider deciding on a framework for proportional allocation of content areas in question papers, securing a balance between the different content areas to improve the face validity of the question papers.

Our second research question addresses the diagnostic power of the question papers, comparing performance in the different content areas over the seven years in question. Contestants throughout performed best in A (algebra, patterns and functions) with a weighted average of 39%. This could probably have been expected. School curricula in South Africa are technically driven and procedures get a lot of attention at the cost of creative thinking or real problem solving. Participants in the olympiad are therefore more familiar with algebraic manipulations and working with patterns and functions than in most other topics in mathematics. The fact that contestants performed worst in S (Statistics and data handling) can be explained using the same argument. Statistics is a content area which has been introduced into the curricula in schools fairly recently and many teachers tend to neglect this topic. Our results do not indicate a significant change in performance in specific content areas over the period in question. Using the diagnostic results of the study, educators could consider putting a stronger focus on teaching statistics.

Regarding the predictive criterion validity of the olympiad papers, the correlation between the actual performance of contestants and the anticipated ranking as set by the problem setting committee was significantly worse in 2012 compared to 2006. This fact indicates that the committee of experts sometimes gets it right but in other instances it is out of touch with what could be expected from contestants. This ability to anticipate contestants' performance could be improved by involving more school teachers in these committees who are working with the learners on ground level.

References

Barbeau**,** E. J., & Taylor**,** P.J. (Eds.) (2009). *Challenging mathematics in and beyond the classroom. The 16th ICMI Study. New ICMI Study Series 12*. New York: Springer.

Bicknell, B., & Riley, T. (2012). The role of competitions in a mathematics programme. *The New Zealand Journal of Gifted Educatio*n, 17(1), 1-9.

Cai, J. (2003). What research tells us about teaching mathematics through problem solving. In F. K. Lester, Jr. (Ed.), *Research and issues in teaching mathematics through*

problem solving (pp. 241−254). Reston, VA: National Council of Teachers of Mathematics.

Davis, G.A., Rimm, S.B., & Siegle, D. (2014). *Education of the gifted and talented. 6th edition.* Harlow, UK: Pearson.

DBE (Department of Basic Education) (2011). *Curriculum and Assessment Policy Statement (CAPS), Grades 7-9, Mathematics*. Pretoria: DBE.

Desjarlais, M. A. (2009).*Gender differences on the American Mathematics Competition AMC8 contest*. Unpublished PhD, University of Nebraska-Lincoln.

Gleason, J. (2008) An evaluation of mathematics competitions using Item Response Theory. *Notices of the AMS*, 55(1), 8-15.

Gyöngyösi, (2002). Mathematics competitions and their role in education. *Acta Academiae Paedagogicae Agriensis, Sectio Mathematicae*, 29, 115-124.

Kahane, J-P. (2009). Cooperation and competition as a challenge in and beyond the classroom. In *Challenging mathematics in and beyond the classroom*. In E.J. Barbeau & P.J. Taylor (Eds.) ICMI Study 16. New ICMI Study Series, Vol. 12. Retrieved 11 May 2016 from http://www.amt.edu.au/pdf/icmis16pkahane.pdf

Kenderov, P. S. (2006). *Competitions and mathematics education.* Proceedings of the International Congress of Mathematicians, Madrid, Spain, (pp. 1583-1598).

Lester, F. & Cai, J. (2010). *Why is teaching with problem solving important to student learning*. NCTM report. Retrieved on 6 May 2014 from http://www.nctm.org/uploadedFiles/Research_News_and_Advocacy/Research/Clips_and_Briefs/Research_brief_14_-_Problem_Solving.pdf

Miguel, J. (2012). Quora blog. Retrieved 22 February 2016 from https://www.quora.com/How-do-problem-setters-set-questions-for-maths-computer-Olympiads-and-competitions-IOI-IMO-ACM-ICPC-etc

Mullis, I.V.S.,Martin, M.O., Ruddock, G.J., O'Sullivan, C.Y., & Preuschoff, C. (2009). *TIMSS 2011 Assessment Frameworks*. Boston, USA: TIMSS & PIRLS International Study Center Lynch School of Education, Boston College.

Neidorf, T.S., Binkley, M., Gattis, K. and Nohara, D. (2006). Comparing mathematics content in the National Assessment of Educational Progress (NEAP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 assessments. Technical report NCES 2006-029. Darby, PA: DIANE Publishing.

Niederle, M, & Vesterlund, L. (2010). Explaining the gender gap in math test scores: The role of competition. *Journal of Economic Perspectives*, 24(2), 129-144.

Ridge, H. L., & Renzulli, J. S. (1981). Teaching mathematics to the talented and gifted. In V. Glennon (Ed.) *The mathematical education of exceptional children and youth* (pp. 191-266). Reston, VA: National Council of Teachers of Mathematics.

Olkin, I., & Schoenfeld, A. H. (1994). A discussion of Bruce Reznick's chapter (Some thoughts on writing for the Putnam). In A.H. Schoenfeld (Ed.) *Mathematical Thinking and Problem Solving* (pp. 39-51). Hillside, NJ: Lawrence Erlbaum.

Somers, L. & Callan, S. (1999). *An examination of science and mathematics competitions*. Technical report , Westat , Rockville, MD, Retrieved 20 December 2015 from http://www.wmich.edu/evalctr/competitions.pdf

Taylor, P. (2008). *ICMI Study 16: Current Perspectives.* Plenary Lecture presented at the 5th International Conference on Creativity in Mathematics and the Education of Gifted Students, Haifa. Retrieved on 15 June 2014 from http://www.amt.edu.au/icmis16haifapaper.pdf