Psychology Science Quarterly, Volume 50, 2008 (2), pp. 147-172

The *optimal achievement model* and underachievement in Hong Kong: an application of the Rasch model

SHANE N. PHILLIPSON[1]

Abstract

Termed the *optimal achievement model*, a new method for the estimation of underachievement relies on measuring student potential (*P*) and achievement (*A*) using Rasch models and calculating an achievement index, I_A, for each individual. This study extends a previous report (Phillipson & Tse, 2007) that estimated the proportion of Hong Kong students in Primary 5 to now include a sample of students from Primary 3 (n = 1406), Secondary 1 (n = 756) and Secondary 3 (n = 578), across six districts of Hong Kong. The students were administered a standardized test of mathematical achievement and the Ravens Progressive Matrices Test. Using the *optimal achievement model*, estimates of underachievement at six percentile bands showed that the proportion of students who were underachieving ranged from 10 % at the 50-59th percentile band up to 30 % at the >95th percentile band for Primary 3, Primary 5 and Secondary 1 students, and 50 % of Secondary 3 students. The estimation of I_A at the level of the individual allows the researcher the possibility to directly study the interaction of the environment on student potential.

Key words: Mathematics, optimal achievement model, underachievement

[1] Correspondence regarding this paper should be addressed to Shane N. Phillipson, PhD, The Department of Educational Psychology Counselling and Learning Needs, The Hong Kong Institute of Education, 10 Lo Ping Rd, Tai Po, NT, Hong Kong; Phone: (852) 2948 8548, email: shane@ied.edu.hk

Introduction

The study of student underachievement continues to diversify, including the characterization of gifted underachievers (Dixon, Craven, & Martin, 2006), the EEG differences between students of average and above average intelligence (Staudt & Neubauer, 2006), motivational orientations amongst high- and under-achievers (Ziegler & Stoeger, 2004), self-regulation in gifted mathematics underachievers (Stoeger & Ziegler, 2005), students with emotional and behaviour disorders (Lane, Gresham, & O'Shaughnessy, 2002), students with learning disabilities (Fletcher, Denton, & Francis, 2005), and the relationship between parenting styles and achievement (Jacobs & Harvey, 2005). Despite general agreement that underachievement is the discrepancy between what can be expected and what is actually achieved, there is a growing realization that the current methods used to identify underachievement are themselves problematic (Fletcher et al., 2005; Lau & Chan, 2001; Phillipson & Tse, 2007; Stoeger & Ziegler, 2005; Ziegler & Stoeger, 2003) with some studies beginning to propose alternate methods (Fletcher et al., 2005; Lau & Chan, 2001; Phillipson & Tse, 2007).

At a fundamental level, this article argues that our understanding of the conceptual basis of underachievement is not well defined, and that there is a need to rethink the nexus between potential and achievement. Furthermore, it is significant that over-achievement has not received the same degree of attention as underachievement, except that they are seen as opposite ends of a continuum, with the achievement of intellectual potential being arbitrarily defined as the mid-point between the two. However, any conceptual basis for the detection of underachievement must be a unified model, including explanations for both the phenomenon of under- and over-achievement.

Furthermore, there appears to be little adherence to the requirements of fundamental measurement when using instruments that address issues of underachievement, despite progress in related domains such as intelligence (Styles, 1998; van der Ven & Ellis, 2000; Vigneau & Bors, 2005), science education (Liu & Boone, 2006) and school achievement (Lissitz, 2005). These measurement requirements include the need for unidimensionality of the measurement instrument, invariance in the order of the items used to construct the instrument, and the units of measurement should correspond to an interval level scale (Bond & Fox, 2001, 2007; Styles, 1999). In the development of measurement models that meet these requirements, Georg Rasch (1980) and others (see Bond & Fox, 2007; Smith & Smith, 2004) have developed the methodology to assess the specific objectivity of the test instruments that produce dichotomous (True/False) data and polytomous data, for example, and to transform raw counts of participant responses from these instruments into interval level measures.

In Hong Kong, Phillipson and Tse (2007) showed that across all ability levels, the percentage of Chinese Hong Kong students in Primary 5 who are underachieving in mathematics is approximately 9.4 % (90 out of 957). Rather than using the more common methods of estimating underachievement, such as the absolute split method, simple difference method or regression method, Phillipson and Tse (2007) based their calculation on Rasch measures and the explicit assumption that both tests of mathematical achievement and intellectual ability reflected the same underlying psychological construct or latent trait.

This article has two specific purposes. The first is to describe the psychometric basis of a new model for the detection of underachievement, termed the optimal achievement model (OAM). The second is to complement the findings reported in Phillipson and Tse (2007) by estimating the proportion of Hong Kong Primary 3, Secondary 1 and Secondary 3 students who, according to this proposed model, are underachieving in mathematics. The first section begins with a brief review of the three more commonly used methods before describing the theoretical basis of the OAM and the research design.

Limitations in the current models of underachievement

The three most commonly used methods for the estimation of underachievement are the *absolute split method*, the *simple difference method* and the *regression method* (Chan, 1999; Lau & Chan, 2001; Phillipson & Tse, 2007). In the *absolute split method*, percentiles are created for both standardized scores in both tests of intellectual ability (for example the Ravens Progressive Matrices test) and achievement. If a given student's score is in the top 25 % of intellectual ability and below the bottom 25 %, for example, then that student is underachieving.

For the *simple difference method*, student scores in both tests of intellectual ability and achievement are concerted into standardized scores (z-scores). When the difference between z-score achievement and z-score ability is either < -1 or > +1, then the student is regarded as under or overachieving respectively. Similarly, the *regression method* relies on standardized scores of both tests of intellectual ability and achievement. When the achievement scores are regressed against the scores of intellectual ability, students whose deviations from the regression line are less than -1 are defined as underachieving.

The estimation of underachievement depends on a direct comparison of scores from the tests of intellectual potential and achievement. This is usually done by converting deviation IQ scores (or percentiles) and achievement scores into z-scores, thereby equating the scores on the two tests. Implicit in these models is the assumption that achievement (A) equals potential (P) plus the sum of all environmental factors (E) that contribute to (or inhibit) achievement. This "additive" model can be shown mathematically as:

$$A = P + \sum_{i=1}^{N} E_i$$

where A = z-score achievement, P = z-score potential, and $\sum E$ = *motivation* + *self-efficacy* + *education* + *chance* ... When the net sum of the environmental components is zero, then the individual is achieving his or her potential, but if $\sum E$ is < -1 standard deviation or > +1 standard deviation, then the individual is defined as under- or over-achieving respectively. In this model, however, the contributions of the various components of E on P are not well defined.

Leaving aside arguments against the arbitrary use of cut-scores such as top 25 % and standard deviations of 1, all three methods are highly dependent on sample parameters such as means and standard deviations. As pointed out in Lau and Chan (2001), when samples change so will these values, thereby reducing the generalisability of these methods. At a more fundamental level, however, there is no justification for the use of mean scores (and standard deviations), no matter how accurately they reflect the population means (or standard deviations), to determine the absolute potential or achievement scores of that student, much less to assess whether or not a student is under (or over-) achieving. Statistical parameters such as means and standard deviations describe the characteristics of groups, not individuals. Using the analogy of student height, it makes little sense to argue that an assessment of a student potential (or actual) height, for example, is determined by the mean potential (or actual) height of his or her classmates. The requirements for our current purpose are objective assessments of both the student's intellectual potential and his or her achievement rather than relative comparisons.

Fletcher, Denton and Francis (2005) have also highlighted a number of other problems in using these methods. Assessments of underachievement must take into account problems of test reliability, particularly when the tests are used at a single point in time, and measurement errors arising out of issues such as assumed normality. These measurement errors are compounded when two or more tests are used concurrently in assessing underachievement. Furthermore, if a student is underachieving in a given subject because of poor attitude and motivation, it is possible that the same student will also underachieve on a test of intellectual ability.

Fletcher et al. (2005) also highlighted ongoing issues of validity, arguing that the considerable research evidence to date does not support the classification of students as underachievers on the basis of discrepancy scores. In contrast, classifications based on IQ scores or learning disability has empirical evidence to support this distinction. In the absence of a sound conceptual basis to the discrepancy models, Fletcher et al. recommend that low achievement models, together with response to instruction models best serve the needs of identification of students with learning disabilities.

As Fletcher et al. (2005) argued, underachievement is a latent variable and, hence, imperfectly measured by test instruments such as tests of intellectual ability and achievement. A new conceptual basis of achievement must take into account the objective limitations of both the instrument (its reliability and validity) and the possibility that the student may themselves influence the parameters of the instrument. In the description of the optimal achievement model for identifying achievement, the use of Rasch models might overcome many of the limitations in the test instruments and take into account the possibility of interactions between the test and testee.

Developmental models of achievement

Two recent models of giftedness describe the development of expertise and are of relevance to this article. Gagné's (2005) DMGT described the development of natural abilities (or gifts) in an individual into talents. The developmental process is facilitated by both intrapersonal catalysts (IC) and environmental catalysts (EC). A third catalyst, chance, acts on gifts, IC and EC. As in the physical sciences, Gagné's catalysts are not affected by the developmental process but return to their natural state. According to Gagné, components within each of the catalysts interact with each other in complex ways, sometimes inhibiting as well as promoting talent development.

In contrast, the Actiotope Model of Giftedness (AMG) (Ziegler, 2005) focused on an individual's action repertoire (including both cognitive processes and behaviours) in the development of expertise, rather than gifts as defined by Gagné. As the repertoire develops and expands over time in response to the environment, they also directly affect the environment.

In the development of mathematical expertise, an infant's Actiotope is thought to be built upon an innate set of mathematical skills that are common to all other infant (Phillipson & Callingham, *in press*). Differences in the infant's subsequent environment will lead to differences in each individual's Actiotope. Although the DMGT and AMG differ in respect to the importance they give to gifts, both emphasize the role played by chance factors, and complex relationship between the individual and the environment. In the DMGT, chance plays a role in determining the level and combination of an individual's natural abilities as well influencing both IC and EC. Ziegler (2005) maintained that the AMG is a probabilistic model and that any given Actiotope can never guarantee expert status. However, Gagné's DMGT is more restrictive because it does not acknowledge the possibility that both IC and EC may change over time.

Apart from chance, Gagné (2005) ranked natural abilities, particularly cognitive abilities, as the primary causal agent (p. 136) of academic achievement, particularly in grade and high school, followed by IC and EC factors. Gagné cited the influential work of Jensen (1998) and others in demonstrating the explanatory power of IQ scores and the *g* model of intelligence as a predictor of academic achievement. Ziegler (2005) didn't distinguish between internal and external factors, noting, as did Gagné, that each can influence the other. Irrespective of the model, it is clear that measuring the achievement of any one individual at one particular point in time during the developmental stage represents the culmination of a large number of very specific and complex environmental conditions that are unique to that individual.

If cognitive abilities, as reflected by IQ (Gagné, 2005, p. 137), play a causal role in achievement, then for any given IQ score a range of achievements can be expected. Because development doesn't occur within an optimally conducive environment, it is likely that very few individuals will be achieving close to their potential, with most underachieving. Although there is no reason to expect that achievement levels will be normally distributed, sample parameters such as mean (and standard deviation) achievement level can be estimated for any given potential, with patterns of achievement emerging if a large number of individuals with a common context, such as school history or SES background, are examined.

g and achievement

Of fundamental importance in the estimation of underachievement is the distinction between tests of intellectual potential and tests of achievement. Recent tests of intellectual potential for mathematics have included the Ravens Progressive Matrices (RPM) (Lau & Chan, 2001; Phillipson & Tse, 2007; Stoeger & Ziegler, 2005; Ziegler & Stoeger, 2004), the WISC-R (Dixon, Craven & Martin, 2006) and the "Intelligenz-Struktur-Test 2000 R" - a German psychometric test (Staudt & Neubauer, 2006).

The RPM is a routinely used test of mental ability[2] (Jensen, 1998) with RPM scores most often reported as deviation IQ scores or percentiles. In this way, the raw score of an individ-

[2] Jensen (1998) preferred the term "mental ability" rather than intelligence, saying that the latter term "should be discarded altogether in scientific psychology" (p. 48). In broad terms, Jensen defined mental ability as a voluntary performance on any distinct behavioural and observable act that cannot be attributed to reactions of the autonomic nervous system, nor dependent on sensory or physical dexterity.

ual is compared to scores obtained by other individuals of the same age, and hence, are dependent on the availability of accurate norm tables. In other words, raw scores are not the basis for understanding the relationship between mental ability and its biological and cognitive correlates.

The RPM is one of many IQ tests that are highly loaded with *g* (or general) factor of mental ability (Jensen, 1998). The RPM, in particular, has been specifically designed to "maximize the relation eduction and to minimize group factors" (p. 90) associated with different types of content. The predictive value of IQ scores derived from the RPM for school achievement (rather then grades) is very high, ranging up to .7 for all ability levels. In other words, school learning is highly dependent on *g*, especially in mathematics (Jensen, 1998, p. 278, 279). According to Jensen

a person's level of g *acts only as a threshold variable that specifies the essential minimum level required for different kinds of achievement. Other, non-g special abilities and talents, along with certain personality factors, such as zeal, conscientiousness, and persistence of effort, are also critical determinants of educational and vocational success. (p. 542)*

On the other hand, tests of mathematical ability are also *g* loaded, but they also reflect non-*g* factors such as personality, motivation and environment. When *g* is held constant, education becomes the major predictor of performance (p. 557), including tests of mathematical achievement. Furthermore, the effect of education on *g* is "multiplicative" rather than additive (Jensen, 1998, p. 557).

The relationship between *g* and achievement is more complex than implied thus far. Using a fiscal analogy, Jensen (1998) described the relationship of *g* with achievement in terms of Spearman's "law of diminishing returns", where higher levels of *g* become less important in explaining achievement. In other words, achievement performance in lower-IQ groups is much more dependent on *g* than achievement performance in higher-IQ groups (Deary, Egan, Gibson, et al., 1996). By extension, it is clear that achievement might mean different things at different IQ levels and, hence, a different criterion of underachievement might apply across different IQ groups.

Rasch models

To date, all of the studies that estimate underachievement do not seem to address issues relating to the psychometric properties of the two tests, but assume that each test objectively measures intellectual potential or achievement. Furthermore, when raw scores from either the RPM or achievement tests are used it is required that the items in each test are able to contribute meaningfully to the estimate of the individual's potential or achievement. These assumptions are rarely, if ever challenged in the literature on achievement.

A family of mathematical transformation techniques has been developed since the pioneering work of Georg Rasch (1980). Known as Rasch models, these models include transformation of dichotomous data, responses to rating scales and partial credit data into interval level measures (see Bond & Fox, 2001; 2007; Kubinger, 2005; Phillipson & Tse, 2007). Furthermore, the transformation takes into account the individual variability in the use of the test instruments, allowing for estimates of the instruments reliability as well as individual item difficulty. Items in tests such as RPM and tests of achievement are ranked in order of difficulty and expressed on an interval level scale of measurement termed the logit (log-odds units) scale and allow for the identification of redundant as well as other ill-fitting items.

This research takes the view that the Rasch model is the preferred model for developing interval level measurement scales (e.g., Andrich, 1988; Bond & Fox, 2007; Wilson, 2004), although it is often regarded as being just one of the many item response theory or latent trait models (e.g., Kubinger, 2005; van der Linden & Hambleton, 1987). Although a detailed discussion of these two perspectives is beyond the scope of this article, in this research the fit of the data to the Rasch model is based on residual based statistics (infit and outfit mean squares and their standardized forms (Wright & Stone, 1979) rather than likelihood ratio tests adopted in the European tradition (Andersen, 1973, Kubinger, 2005). This distinctions are reflected in the design of Rasch modeling software such as Quest (Adams & Khoo, 1992) and Winsteps (Linacre & Wright, 2004), the former being the software used for these analyses in this research.

Person estimates are also produced by Rasch models, allowing for the ranking of individuals according to their scores on the test instrument. Thus, persons are ranked from most capable to least capable after allowing for acceptable discrepancies in patterns of their responses to the test items. Again, person estimates are reported in terms of logits and, hence, are measured on an interval level scale. Rasch models include a collection of fit statistics to allow for estimates of the fit of the data to the model as well as estimates of the unidimensionality of the test. Thus, both estimates of an individual's potential and achievement are produced on the same interval level scale (Bond & Fox, 2007).

Styles (1999) argued that the use of fundamental measurement in the study of intelligence is not strong. Some notable exceptions have used Rasch models of student responses to Raven's Matrices to illustrate developmental changes in intellectual growth (Andrich & Styles, 1994; Styles, 1999) that were otherwise not possible to see. In accordance with the principles of fundamental measurement, Rasch models of responses to the RPM have shown that the test is unidimensional (Styles, 1999), although other researchers using more stringent fit statistics have questioned this (Kubinger, Formann, & Farkas, 1991; Lynn, Allik & Irwing, 2004; Mackintosh & Bennett, 2005; van der Ven & Ellis, 1999).

In Rasch models, it is possible to test the hypothesis that two tests measure the same underlying construct (Bond & Fox, 2001; 2007). This hypothesis is tested by creating scatterplots of participant responses to the two tests and by examining the extent by which the points lie between 95 % confidence bands. According to Bond and Fox (2001), 95 % confidence bands are useful when inferring whether two tests essentially measure the same construct.

Using this technique, it is possible to compare directly the participant responses to the RPM and tests of mathematical achievement to determine if they measure the same psychological construct. As noted earlier, both tests are highly loaded in *g*, although we assume that the RPM is more loaded in *g* than the achievement test. In particular, the latter test reflects aspects of self and the wider educational environment and, hence, we are interested in the degree of discrepancy between the two tests rather than agreement. Following this argument, points outside the 95 % confidence bands would indicate a high degree of discrepancy between the two tests because of the relatively high positive or negative influence of the environment.

The optimal achievement model

In the optimal achievement model (OAM), an individual's responses to the RPM and tests of mathematical achievement are converted into logit scores, where the RPM logit score represents their ranking in terms of *g*, and the achievement logit score represents *g* plus environmental components, some of which enhance achievement while others inhibit achievement. As outlined in Gagné (2005) and Ziegler (2005), the relationship between the various components of the environment on potential is interactive, noting that the net effect of the environment (E) on *g* for any individual is multiplicative rather than additive. Accordingly, the relationship between the two categories of environmental components, *enhancing* and *inhibiting*, and P in predicting A for person n is shown as:

$$A_n \propto P_n e^E \text{ or}$$
$$A_n = k.P_n e^E$$

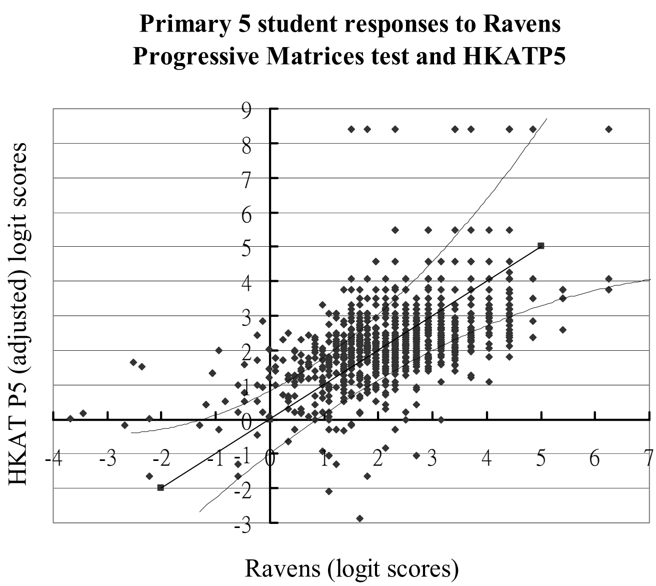
where $E = \left(\sum_{i=1}^{N} e_i - \sum_{j=1}^{N} i_j \right)$, the difference between the sum of the enhancing (e) and inhibiting (i) components. When A and P are measured using the same interval level scale, then k = 1. Rearranging this equation yields an achievement index, I_A, for person n:

$$A_n - P_n e^E = I_{A,n}$$

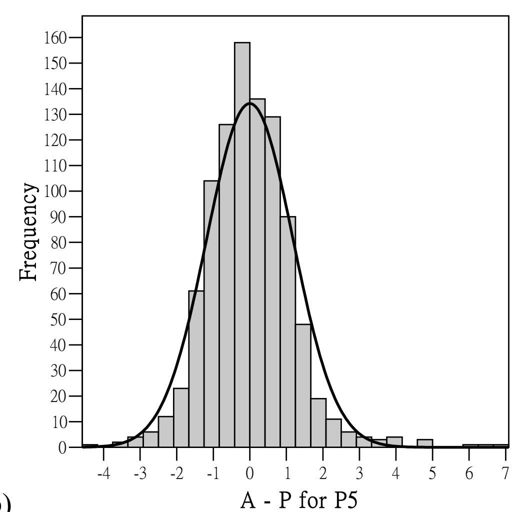
According to this model, when the net effects of the *enhancing* and *inhibiting* components on P is zero (i.e. $\sum_{i=1}^{N} e_i = \sum_{j=1}^{N} i_j$), then I_A is zero because $A_n = P_n$. Similarly, if the sum of *enhancing* components is greater than the sum of the *inhibiting* components (i.e. $\sum_{i=1}^{N} e_i > \sum_{j=1}^{N} i_j$), then I_A is positive in value. In the converse situation, I_A is negative.

In this model, achievement is optimal when the relative contributions of the *enhancing* components in the environment far exceed that of the *inhibitory* components. When an individual's achievement is below this optimal level, he or she is "underachieving" although the level of underachievement will differ in severity. Also note that this model is at the level of the individual and does not depend on knowing population parameters such as mean and standard deviation, although these can be conveniently estimated if required. Although the model can make estimates of both the upper and lower levels of achievement for any given potential, it makes no assumptions of the distribution of scores, suggesting that for any given context, the distribution may, in fact, be non-normal.

Using the data from Phillipson and Tse (2007), it is possible to illustrate several key features of the OAM. This study reported the responses by a random sample of 953 Primary 5 students in Hong Kong to the RPM and a standardized test of mathematical achievement. The modal age of the students was approximately 11 years, ranging from 8.5 to 14.5 years. Hence, the range of RPM logit scores reflected both age dependent and ability dependent variability, and the range of logit scores from the test of achievement reflected their study of the same mathematical syllabus. After transforming the participant responses to both tests into logit scores it is possible to rank order the participants from most capable to least capable. Scatterplots of these responses (Figure 1A) showed the degree of discrepancy between participant responses to the RPM and achievement test and 95 % confidence bands. The correlation between the HKATP5 and RPM logit scores was estimated as .53. The correlation between age and HKATP5 or RPM logits, on the other hand, was negligible with correlations of -0.03 and .01 respectively.



a)



b)

Figure 1:
a) Scatterplot of RPM and HKATP5. Person logits from the RPM and HKATP5 (n = 953 students) were plotted after adjusting the HKATP5 logit scores in order to align the two scales (Bond & Fox, 2007, pp. 84-90). The centre line represents the ideal modeled relationship between the RPM logit and HKAT logit scores, and the curved lines represent the upper and lower 95 % confidence bands. Note that each point may represent more than one student.
b) Frequency distribution of I_A (A – P) for Primary 5 students (n = 953 students).

Also shown in Figure 1B is the frequency distribution of I_A – the difference between the logit scores for the HKATP5 (or ability) and RPM (or potential). The distribution appears non-normal (Kolmogorov-Smirnov $Z = 1.461$, $p = .028$, $n = 953$).

Consider a RPM logit score of 4 in Figure 1. At this relatively high level of ability, the level of achievement ranges from 1.07 to 5.55 logits ($\bar{x} = 2.95$, $s = .65$, $n = 34$). The maximum level of achievement attained by students a little lower and higher in ability, however, is greater than 8 logits so this upper value of 5.55 logits may be an artifact of the sample. As described by the OAM, the upper level of achievement reflects the maximum contribution of the *enhancing* environmental factors and the least contribution of the *inhibiting* factors. Likewise, the lower level of achievement reflects the converse. The equal but opposing contribution of the *enhancing* and *inhibitory* components is shown by the ideal modeled relationship (dotted line).

Similarly, a HKAT score of 2 logits, for example, was achieved by students with a wide range of abilities as reflected in the RPM logit scores, ranging from -.94 to 4.4 ($\bar{x} = 2.01$, $s = .89$, $n = 65$) logits. This range reflects students with very different abilities and/or ages, some of whom are achieving close to their maximum level and some of whom are not. Each end of this ability spectrum highlights students who show evidence of unexpected over-achievement and underachievement respectively. An examination of the frequency distribution of I_A for a given RPM logit and HKAT logit may yield further information regarding the specific contexts of achievement.

Figure 2A shows the frequency distribution of I_A for the RPM logit score of 4. The distribution appears normal (Kolmogorov-Smirnov $Z = .78$, $p = .58$, $N = 34$) with a mean I_A of -1.1 ($s = .65$) and, surprisingly, with over 95 % of scores below the I_A of 0. In other words, over 95 % of Primary 5 students are experiencing an environment where the *inhibiting* components exceed the *enhancing* components.

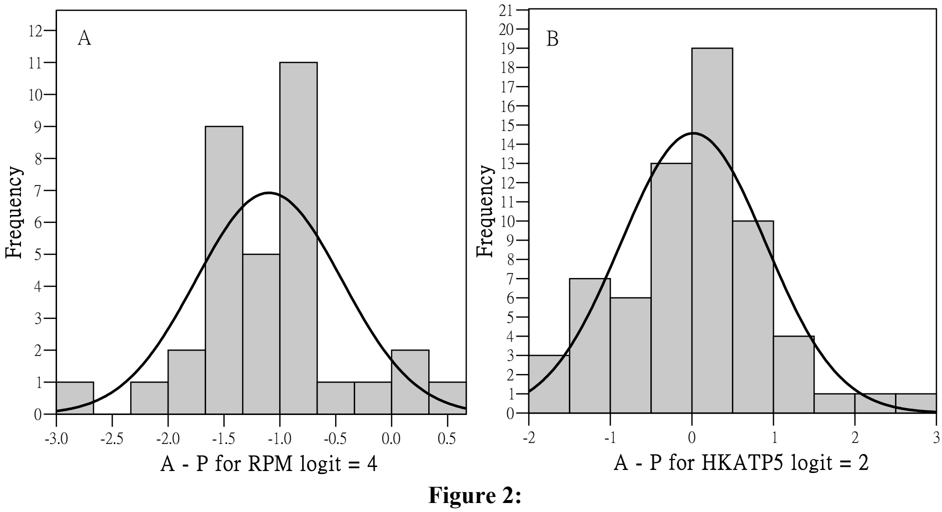


Figure 2:

a) Frequency distribution of I_A (A – P) for a RPM logit score of 4 ($\bar{x} = -1.1$, $s = .65$, $n = 34$ students); **b)** Frequency distribution of I_A (A – P) for a HKAT logit of 2 ($\bar{x} = -.02$, $s = .89$, $n = 65$ students). Data from Phillipson & Tse, 2007.

Figure 2B shows the frequency distribution of I_A for a HKAT logit of 2. Again, the distribution appears normal (Kolmogorov-Smirnov $Z = .72$, $p = .689$, $n = 65$) with mean I_A of .02 ($s = .89$). In this sample of Primary 5 students, a HKAT logit score of 2 is just below the sample mean of 2.2 ($s = 1.22$), and hence, these students would be considered to be "achieving" in comparison with the sample mean. This initial conclusion is misleading because only half of the students have an I_A greater than 0, with the balance "underachieving".

In deciding whether the level of underachievement is of practical significance, the 95 % confidence bands can provide a useful criterion. The 95 % confidence bands describe critical values of $I_{A,\ RPM}$ whereby values of I_A outside the lower 95 % confidence band would determine if a student is underachieving. For a logit score of 4, the corresponding value of HKAT on the 95 % confidence band is 2.7, hence $I_{A,\ 4} = 2.7 – 4 = -1.3$. Using this criteria, 13 (38 %) have a I_A value less than -1.3 and, hence, could be deemed to be underachieving. The values of I_A less than -1.3, however, are not constant and although there appears to be a lower limit to the value of I_A, it is not possible to objectively quantify these values.

Outline of study and hypotheses

This study extends a previous report in Phillipson and Tse (2007) by including students from Primary 3, Secondary 1 and Secondary 3 in estimating the proportion of students who are underachieving in mathematics. It tests the hypothesis that there are substantial proportions of students who are underachieving in mathematics (Lau & Chan, 2001; Phillipson & Tse, 2007), although the proportions are not distributed equally across levels of ability or grade level.

As in the previous study, the aim of the current study was to obtain a representative sample of students across Hong Kong and to administer the Ravens Progressive Matrices (RPM) test and a standardized (Hong Kong) test of mathematical achievement (HKAT) suitable for the grade level. The responses to the RPM and HKAT were subjected to Rasch models for dichotomous data and polytomous data respectively and examined initially for fit to determine if the responses fitted the Rasch model. Person estimates of potential (P) and achievement (A) in terms of logit scores were then analyzed using scatterplots as well as other parametric and non-parametric tests as appropriate. As well as analyzing the responses from students in Primary 3, Secondary 1 and Secondary 3, this report also reinterprets the data obtained from Phillipson and Tse (2007).

Method

Schools

The Education Bureau[3] (EDB) of the Hong Kong SAR has divided the territory into 16 administrative regions. These 16 districts were collapsed into six according to the basis of their common SES in order to facilitate the sampling procedure. Random lists of government

[3] In 2007, the Education and ManPower Bureau (EMB) of the Hong Kong SAR changed its name to Education Bureau (EDB).

schools and their classes at each of the three grade levels, Primary 3, Secondary 1 and Secondary 3, were created and oversampled until a sufficient number of students were obtained. In practice, however, there were difficulties in procuring students from particular districts within a reasonable timeframe because of the reluctance of personnel in certain schools.

Participants

A total of 3,790 students across the four grade levels participated in the study as consenting volunteers (*n* = 1,433 Primary 3 students, *n* = 957 Secondary 1 students, *n* = 595 Secondary 3 students), including *n* = 957 students from Primary 5. All students were advised that the results to the tests were confidential and to be used solely for research purposes. Descriptive statistics of the students across the three grade levels, together with data from Primary 5 (Phillipson & Tse, 2007) are shown in Table 1. A log likelihood ratio test (Sokal & Rohlf, 1981) showed that the sampling did not produce a representative cross section of students.

Table 1:
Number of students from the four grade levels according to district

	Grade level							
	Primary 3[1]		Primary 5[2]		Secondary 1[3]		Secondary 3[4]	
Frequency District	Observed	Expected	Observed	Expected	Observed	Expected	Observed	Expected
1	74	51.6	106	32.9	30	24.7	0	18.8
2	213	198.2	211	127.4	256	128.9	76	94.0
3	223	188.4	120	132.1	119	90.0	157	67.2
4	258	341.2	53	219.5	98	192.4	164	135.7
5	269	247.2	142	168.6	25	155.3	18	118.2
6	396	406.3	325	276.5	277	213.6	180	161.2
TOTAL	1433	1432.9	957	957	805	804.9	595	595.1

Total number of students is 3,790.
In order to determine if the observed number of students at each grade level were a proportional representation of students from each district, expected frequencies were calculated using data from the (then) Education and ManPower Bureau (EMB) of Hong Kong for the 2005-06 school year. The agreement between the expected and observed proportions is estimated by the log likelihood ratio test (Sokal & Rohlf, 1981). This test calculates a G index (for df = n - 1) and is tested against a χ^2 distribution. The results show that the observed proportions do not match the expected proportions and so the samples do not contain a representative cross-section of students from each district.
[1]G(5) = 40.2, *p* = .000.
[2]G(5) = 354.9, *p* = .000.
[3]G(5) = 343.5, *p* = .000.
[4]G(5) = 268.6, *p* = .000.

Instruments

Each student's potential (P) for mathematical achievement was estimated using the Ravens Standard Progressive Matrices (RPM), a widely used non-verbal test of intellectual ability, consisting of five groups of 12 items in order of increasing difficulty (Raven, Court, & Raven, 1983). Responses were scored as correct or incorrect, corresponding to a Rasch model for dichotomous data.

Each student's achievement (A) in mathematics was estimated using abridged versions of the Hong Kong Attainment Tests (HKAT) that were currently in use at the time of testing. These tests were used by the (then) Education and ManPower Bureau (EMB) of Hong Kong SAR to assess student mastery of mathematical knowledge and skills. Four different achievement tests were used, corresponding to the four different grade levels. At the time of testing, however, Primary 3 and Primary 5 students were either studying an old or new version of mathematical curriculum because of a policy of allowing schools to gradually phase in the new curriculum over a period of time.

For Primary 3 students, the mathematical achievement test (HKATP3) consisted of 32 questions and each was graded as correct or incorrect. For Primary 5 students, the achievement test (HKATP5) consisted of 37 questions and each were given 0, 1, 2, 3 or 4 marks, depending on the correctness of the answer. However, not all Primary 3 and Primary 5 students completed all questions in their respective tests because of differences in the use of old and new curricula. The responses to the HKATP3 and HKATP5 were analyzed using the Rasch model for dichotomous data and partial credit model respectively.

Students in Secondary 1 were assessed using a 26 question test (HKATS1), with each question scored as correct or incorrect. Secondary 3 students were assessed using an 18 question test (HKATS3), consisting of 12 questions that were scored as correct or incorrect and six questions that were scored as 0, 1, 2, 3, 4 and 5 depending on the question and ability of the student. The subsequent analysis required a Rasch model that concurrently handled both dichotomous and partial credit responses.

Responses to the HKAT were double marked by student-teachers majoring in mathematics and specifically trained to a common marking scheme. The marking schemes were based on recommendations published by the Education and ManPower Bureau (2003a, 2003b).

Procedures

Data was collected during the latter half of 2005, corresponding to the first half of the Hong Kong school year. Participating schools were asked to set aside 1 1/2 hours for the specific purpose of completing both the RPM and the HKAT, with students sitting singly at a desk. Approximately half the schools administered the HKAT first. The RPM was administered according to the standard procedures described in the test manual (Raven, Court, & Raven, 1983). The HKAT was administered as a test booklet and answer book. As appropriate, students were advised to show working because it would contribute to the overall score for any one question.

Analysis

Student responses to the RPM and HKAT were subjected to Rasch analysis using Quest 90 software programme (PISA Version: August 4 1999) (Adams & Khoo, 1992) for dichotomous data and the partial credit model for responses with intermediate levels of success. Rasch modeling produces estimates of item difficulty and person maps, together with their respective error estimates, and reported as logits. Fit statistics are also produced in unstandardized (mean square) and standardized (*t* statistic) forms (Adams & Khoo, 1992; Bond & Fox, 2001; 2007). Perfect (and zero scores) on both the RPM and HKAT were ignored by the software because Rasch estimates are based on probabilities of success (and failure).

In order to equate the interval level scales from each test for each grade level, the relative length of the HKAT logit scale was adjusted to match the RPM logit scale (Bond & Fox, 2007, pp. 84-90). This was achieved according to the following relationhip: $HKAT_{adj} = HKAT.S_{\bar{X}}/S_{\bar{Y}} + \bar{X} - Y.S_{\bar{X}}/S_{\bar{Y}}$, where $\bar{X}$ and $S_{\bar{X}}$, and $\bar{Y}$ and $S_{\bar{Y}}$ are the sample mean and standard deviation for the RPM and the HKAT logit scores respectively. The HKAT_{adj} logit scores are used in all subsequent analysis.

Scatterplots

Person estimates from both the RPM and the HKAT_{adj} for each grade level were used to create scatterplots together with 95 % confidence bands, according to the guidelines in Bond and Fox (2001, pp. 54-60). Separate scatterplots were also created showing the relationship between RPM person logits and HKAT_{adj} person logits versus age of students to investigate the relationship between RPM scores and age, and mathematical achievement versus age respectively.

Six RPM percentile bands (50-59, 60-69, 70-79, 80-89, 90-94, >95) were estimated from the RPM person estimates at each grade level, and the distribution of HKAT_{adj} scores at each percentile was determined. Estimates of the proportion of students who underachieve were based on comparisons of the values of I_A at each percentile with critical values of $I_{A,\ RPM}$. Underachievement was judged when a student's I_A was less than $I_{A,\ RPM}$.

Results

Fit statistics

Table 2 presents a summary of the Rasch analysis of the RPM and HKAT at each grade level. All of the reliability indices, interpreted similarly to Cronbach alphas, are greater than .97, indicative of their use in high stakes testing. The infit mean squares for each test across all grade levels are close to the expected value of 1, indicating the data has an excellent fit to the Rasch model. Similarly, the outfit mean squares are all close to the expected value of 1, except for the person outfit mean square. In this isolated case, the value indicates that there is 21 % more variation between the Primary 3 student responses to the RPM than is expected according to the Rasch model.

Table 2:
Summary fit statistics following Rasch analysis of student responses to RPM and HKAT

			Primary 3		**[1]Primary 5**		**Secondary 1**		**Secondary 3**	
			RPM	HKAT P3	RPM	HKAT P5	RPM	HKAT S1	RPM	HKAT S3
		$\bar{x}$	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.02
		s	2.27	1.18	2.14	1.20	2.12	0.62	1.75	0.92
		[2]Reliability	1.00	0.99	0.99	0.97	0.99	0.98	0.98	0.98
Items	[3]Infit Mean SQ	$\bar{x}$	0.99	1.00	0.98	1.00	0.99	1.00	0.98	1.00
		s	0.08	0.09	0.09	0.12	0.08	0.09	0.10	0.06
	[4]Infit t	$\bar{x}$	-0.30	-0.10	-0.26	-0.10	-0.08	-0.08	-0.02	0.04
		s	2.40	2.70	1.77	2.28	1.37	2.42	1.38	1.34
	[5]Outfit Mean SQ	$\bar{x}$	1.21	1.02	1.12	1.04	1.04	1.00	0.98	1.01
		s	0.54	0.22	0.56	0.43	0.39	0.16	0.51	0.04
	[6]Outfit t	$\bar{x}$	0.08	0.09	0.49	0.07	0.27	0.00	0.14	0.01
		s	2.53	2.35	2.02	1.42	1.54	2.05	1.70	1.13
		$\bar{x}$	1.55	0.69	2.22	0.31	2.84	0.04	3.30	0.56
		s	1.13	1.11	1.24	0.83	1.22	1.17	1.26	0.65
		Reliability	0.87	0.75	0.87	0.83	0.83	0.82	0.78	0.71
Persons	Infit Mean SQ	$\bar{x}$	0.98	1.00	0.98	1.01	0.99	1.00	0.99	0.98
		s	0.25	0.14	0.26	0.38	0.29	0.12	0.28	0.52
	Infit t	$\bar{x}$	-0.08	0.07	-0.06	0.08	-0.02	0.05	0.04	0.02
		s	1.11	0.68	1.03	0.90	1.02	0.74	0.90	0.87
	Outfit Mean SQ	$\bar{x}$	1.21	1.01	1.12	1.07	1.04	1.00	0.98	1.01
		s	1.77	0.37	1.53	2.81	1.48	0.23	1.26	0.51
	Outfit t	$\bar{x}$	0.22	0.11	0.25	0.29	0.37	0.09	0.24	0.11
		s	0.99	0.66	0.95	1.53	0.77	0.59	0.84	0.82
Number of students tested		n =	1433		957		805		595	
Number of students with perfect or score of "0".		n =	1	26	1	3	4(+2)	15(+2)	13	1
[7]Number of students with invalid responses.		n =	0		0		28		3	
Final number of valid responses.		n =	1406		953		756		578	

[1]Data from Phillipson & Tse (2007).
[2]Item and person reliabilities reflect the replicability of item placements and the person ordering from sample to sample respectively and are interpreted similarly to Cronbach alphas.
[3]Infit mean square is the information weighted sum of squared standard residuals in the form of a χ^2 ratio. The expected value is 1.
[4]The infit t is the transformation of the infit mean square to a normalized t distribution. Expected mean values of t are $\bar{x}$ = 0, s = 1 and acceptable values are ± 2.0.
[5]Outfit mean square is the sum of the squared standardized residuals in the form of a χ^2. The expected value is 1.
[6] The outfit t is the transformation of the outfit mean square to a normalized t distribution. Expected mean values of t are $\bar{x}$ =1, s = 1 and acceptable values are ± 2.0.
[7]Students with missing data, such as date-of-birth were excluded from further analysis.
Note: Information regarding fit statistics from Adams & Khoo (1992), and Bond & Fox (2001).

The mean infit and outfit *t* statistics are all close to 0 as expected, but the standard deviations exceed 2 for the Primary 3 student responses to both the RPM and HKATP3, showing that there are some items that are misfitting. Other misfitting items occur in the HKATP5 and HKATS1 on the other hand, there is no indication of misfitting persons, with all *t* values within ± 2.0. Although the fit statistics for some items and persons are of concern, the overall impression is that both the responses to the RPM and the HKAT generally fit the Rasch model.

Item and person maps

When looking at the mean person logit scores for the RPM across the four grade levels in Table 2, it is clear that the mean score is increasing. This shows that, on average, this test is becoming increasingly easier. This change is obvious when examining the item and person maps for each grade level (Figures 3, 4, 5 and 6). These figures show the distribution of student responses to both the RPM and the HKAT on the logit scale. The most capable students and most difficult items are located at the top end of each scale. As well as distribution of students and items, the maps also show the "relationship between item difficulty and person estimates" (Bond & Fox, 2001, p. 45). A test must not be too easy nor too difficult, with each item contributing to the tests ability to distinguish between persons. A good relationship between test and person is evident when there is an adequate spread of items and there is a close alignment between items and persons at both ends of the scale.

For Primary 3 students, the most difficult RPM items at the top of the logit scale (items 58, 48, 60, 36 and 59) are aligned with the most capable students (Figure 3). In other words, the most difficult items are as difficult as students are capable. At the lower end of the logit scale, however, the easiest items of the RPM are unable to distinguish between any of these students. There is also evidence of item redundancy, with some items showing the same degree of difficulty (i.e. items 11, 46 and 49). On the other hand, the HKATP3 appears perfectly matched with the capabilities of these students, although there is again some evidence of item redundancy.

For Primary 5 students, there is a close match between the most difficult items in the RPM and the most capable students (Figure 4). Again, the least difficult items do not contribute to the tests ability to distinguish between students. Although there is a good match between item difficulty and student capabilities in the HKATP5, there also appears to be a number of redundant items.

The increasing capability of students, together with the decreasing facility of the RPM to distinguish students at both ends of ability is evident in the item and person maps for Secondary 1 students (Figure 5). Likewise, the HKATS1 does not easily distinguish between the most able students, but is too difficult for the least able students.

For Secondary 3 students, the RPM is increasingly less able to distinguish between the more capable students (Figure 6). In contrast, the HKATS3 does appear to distinguish students although a greater spread of item difficulty and student ability would have been desirable.

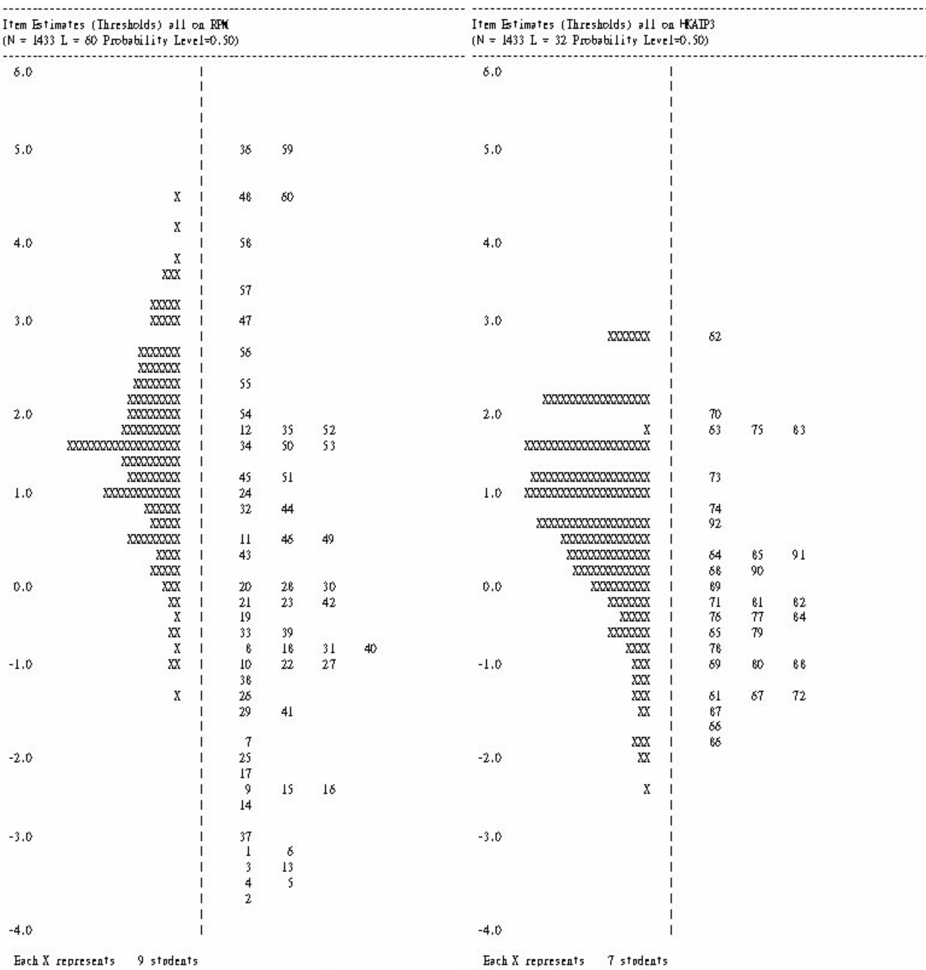


Figure 3:
Item and person maps for RPM (LHS) and HKATP3 following Rasch transformation

```
------------------------------------------------------
Item Estimates (Thresholds) all on RPM
(N = 957 L = 60 Probability Level=0.50)
------------------------------------------------------
7.0                              |
                                 |
                                 |
                                 |
                                 |
6.0                              |
                                 |
                                 |
                              X  |
                                 |
5.0                              |  60
                              X  |  59
                                 |
                          XXXXX  |  48
                                 |
4.0                       XXXXX  |  36
                          XXXXX  |  58
                                 |
                         XXXXXX  |  47
                     XXXXXXXXXX  |  57
3.0                    XXXXXXXX  |
                      XXXXXXXXX  |
                                 |
             XXXXXXXXXXXXXXXXXXX |  56
                      XXXXXXXXX  |  55
2.0                      XXXXXX  |
                         XXXXXX  |  54
                XXXXXXXXXXXXXXX  |  12    35    52
                        XXXXXXX  |  32    34
                           XXXX  |  24    50    53
1.0                     XXXXXXX  |  45    51
                            XXX  |
                             XX  |  11    44
                            XXX  |  43    46    49
                              X  |  20    30    42
0.0                           X  |  19    23    33
                              X  |  21    28
                                 |  39
                              X  |   8
                                 |  10    18    38
-1.0                          X  |  40
                                 |  22    31
                                 |  26    27    41
                                 |  29
                                 |  25
-2.0                             |   7    37
                                 |   9    16
                                 |   6    17
                                 |  13    15
                                 |   3     5    14
-3.0                             |   1
                                 |   4
                                 |
                                 |   2
                                 |
-4.0                             |
------------------------------------------------------
 Each X represents    7 students
======================================================
```

```
----------------------------------------------------------------------------
Item Estimates (Thresholds) all on HKATPS
(N = 957 L = 37 Probability Level=0.50)
----------------------------------------------------------------------------
7.0                              |
                                 |   67.3
                                 |
                                 |
                                 |
6.0                              |
                                 |
                                 |
                                 |
                                 |
5.0                              |
                                 |   73.3
                              X  |
                                 |
                                 |
4.0                              |
                                 |   72.3
                                 |
                                 |   65.3
                                 |
3.0                              |
                                 |
                              X  |
                                 |
                                 |
2.0                              |
                              X  |
                            XXX  |
                           XXXX  |
                           XXXX  |   80.4   82.4   84.4
1.0                     XXXXXXX  |
                   XXXXXXXXXXXX  |   72.1   72.2   84.3   87.2   96.2
                  XXXXXXXXXXXXX  |   81.3   84.1   84.2   85.4
             XXXXXXXXXXXXXXXXXX  |   63.1   63.2   63.3   70.1   70.2
            XXXXXXXXXXXXXXXXXXX  |   62.1   62.2   62.3   68.1   68.2
0.0             XXXXXXXXXXXXXXX  |   65.1   65.2   71.1   71.2   71.3
                     XXXXXXXXXX  |   61.2   64.1   64.2   79.1   79.2
                        XXXXXXX  |   66.1   66.2   67.1   67.2   85.3
                          XXXXX  |   89.1   89.2
                            XXX  |   74.1   74.2   77.1   85.1   85.2
-1.0                        XXX  |   90.1   90.2
                              X  |   91.1   91.2
                              X  |   61.1   76.2   87.1   92.1   92.2
                                 |   76.1   93.1   93.2
                                 |
-2.0                          X  |   96.1
                                 |
                                 |
                                 |
                                 |
-3.0                             |
                                 |
                                 |
                                 |
                                 |
-4.0                             |
----------------------------------------------------------------------------
Each X represents    7 students
============================================================================
```

Figure 4:
Item and person maps for RPM (LHS) and HKATP5 following Rasch transformation
(Data from Phillipson & Tse, 2007)

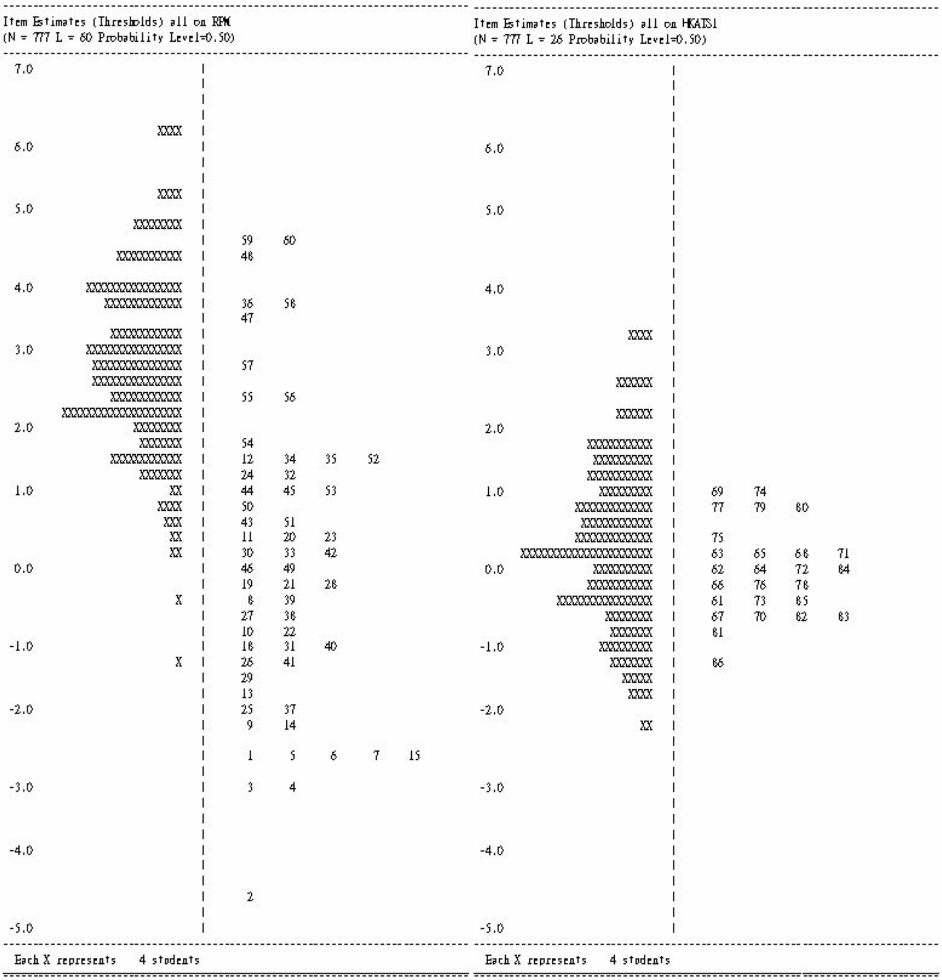


Figure 5:
Item and person maps for RPM (LHS) and HKATS1 following Rasch transformation

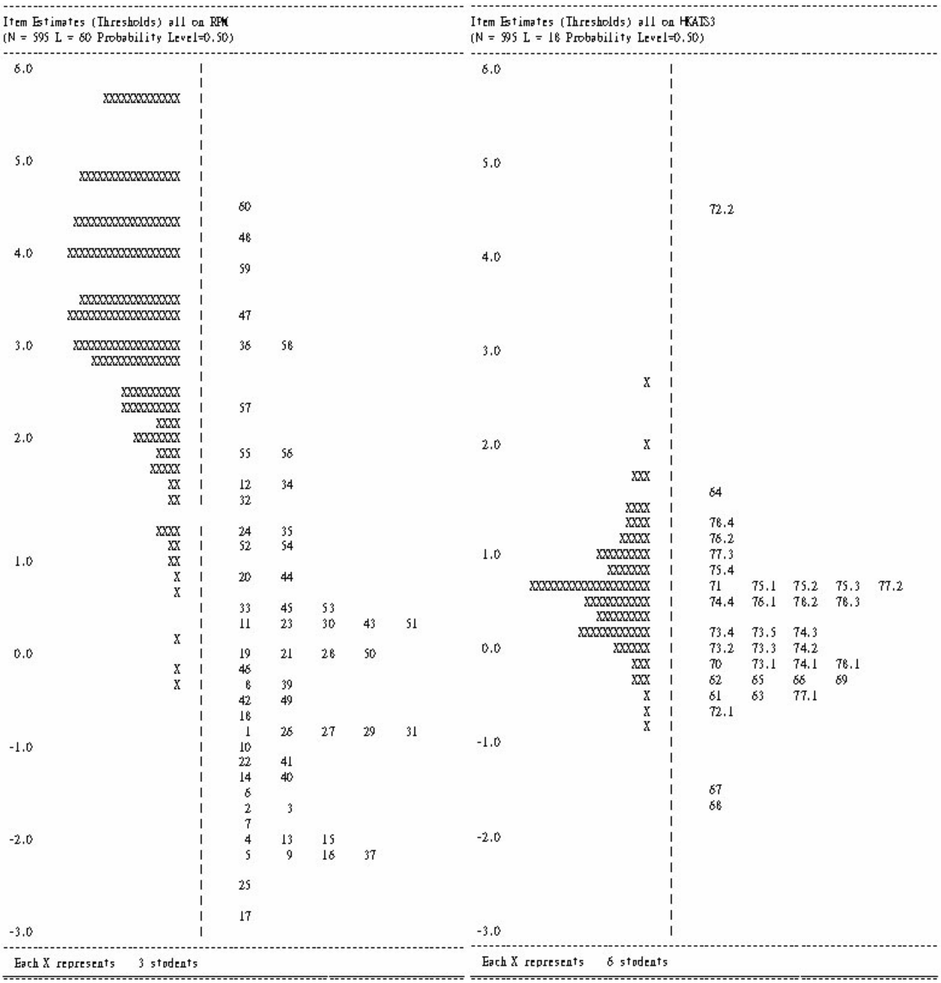


Figure 6:
Item and person maps for RPM (LHS) and HKATS3 following Rasch transformation

Item and person estimates

Individual person and item estimates and fit statistics are also produced by the Rasch analysis. These estimates show that over 90 % of student responses are within the acceptable infit and outfit mean square values of 1 ± 3, and infit and outfit *t* values of ± 2.0. Item fit maps show that 95 % of all items are centred on a value of 1 (range .84 – 1.2), indicating that both the RPM and all versions of the HKAT have a high degree of unidimensionality.

Scatterplots of HKAT and RPM

Scatterplots showing the relationship between the RPM and $HKAT_{adj}$ logit scores and frequency distributions of I_A for Primary 3, Secondary 1 and Secondary 3 are shown in Figures 7A, B and C. The correlations between the RPM and $HKAT_{adj}$ logit scores are .47, .54 and .43 for the Primary 3, Secondary 1 and Secondary 3 responses respectively. In agreement with the responses by Primary 5 students, the correlations between RPM and $HKAT_{adj}$ and age were zero. The frequency distributions of I_A all appear to be normal as indicated by non-significant values of the Kolmogorov-Smirnov *Z* statistic.

Estimates of underachievement

Figure 8 shows the proportion of students whose I_A was less than the than $I_{A,\ RPM}$ at the six percentile bands. Together with the data from Primary 5 students, the results show that proportion of students who are underachieving increases from around 10 % at the 50-59 percentile band to between 26-32 % at the >95 percentile band. For Secondary 3 students at this band, the proportion of students who are underachieving is 53 %.

Discussion

This study constituted a simultaneous cross-sectional survey, designed to estimate the proportion of students in Hong Kong who were underachieving in mathematics. It also argued that some of the difficulties associated with discrepancy models of underachievement can be overcome through the use of objective measurement techniques. Rasch modelling enables student responses to the RPM and tests of mathematical achievement to be transformed onto a linear level scale (logit). When the responses by individual students to both tests are directly compared using scatterplots, the patterns of achievement, including unexpected underachievement can be observed. More importantly, however, is the conceptual basis of the comparison, whereby an achievement index, I_A, based on the difference between achievement (*A*) and potential (*P*) is described mathematically as $A_n = k.P_n e^E$, where $E = \left(\sum_{i=1}^{N} e_i - \sum_{j=1}^{N} i_j \right)$, the difference between the sum of all of the enhancing components and inhibiting components in the environment. While scatterplots have been used in this way

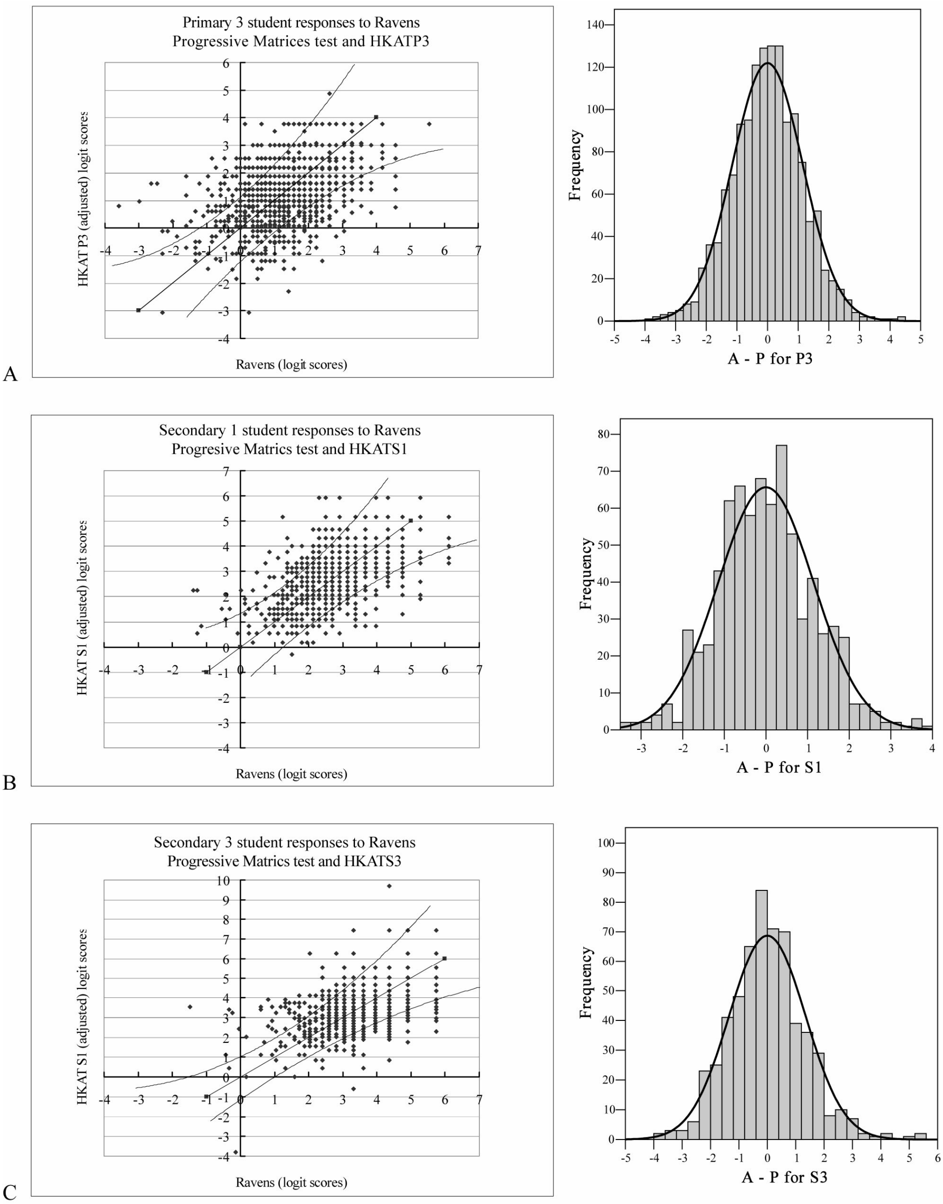


Figure 7:
Scatterplots of RPM and $HKAT_{adj}$ logits (left-hand-side) and frequency distributions of I_A (A – P) for Primary 3 (A) , Secondary 1 (B) and Secondary 3 (C) (Note that each point may represent more than one student.)

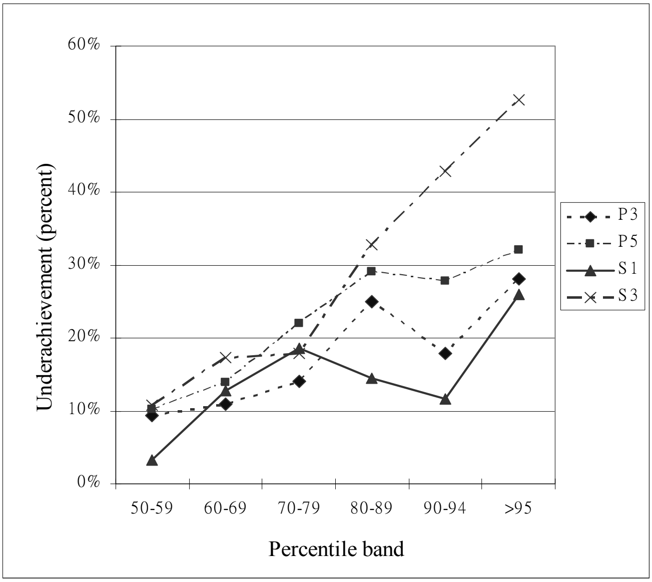


Figure 8:
Proportion of students in Primary 3, Primary 5, Secondary 1 and Secondary 3 who are underachieving in Mathematics

		Percentile band						
		50-59	60-69	70-79	80-89	90-94	>95	Total
Primary 3	*n* below 95 % confidence band	16	10	22	35	18	23	124
	Total *n*	171	91	156	140	101	82	741
Primary 5	*n* below 95 % confidence band	8	16	13	32	20	16	105
	Total *n*	78	115	59	110	72	50	484
Secondary 1	*n* below 95 % confidence band	2	8	18	9	5	14	56
	Total *n*	61	63	97	62	43	54	380
Secondary 3	*n* below 95 % confidence band	6	9	10	18	21	20	84
	Total *n*	56	52	56	55	49	38	306

to measure the degree of agreement between two tests (see Bond & Fox, 2007), the present study is based on the premise that both the RPM and Mathematical tests of achievement already share a common psychological construct (Jensen, 1998). Hence, we are interested in the lack of concordance between the two tests and argue that the difference between the test scores reflects the relative influence of the competing environmental factors, some of which inhibit achievement. Included in the inhibiting factors are personality traits that contribute to motivation such as *ability beliefs*, *effort beliefs*, *value placed on tasks* and *characteristics of task* (Legault, Green-Demers & Pelletier, 2006). As the current study highlights, the patterns of achievement can be studied at the individual level where I_A is derived directly from estimates of person ability and achievement and not sample dependent parameters. Furthermore, definitions of underachievement are not based on the arbitrary use of cut-off points but on the 95 % confidence bands. These confidence bands are also objectively derived and not based on sample parameters.

In the optimal achievement model (OAM), $A_n = k.P_n e^E$, it is clear that there appears to be upper and lower limits of achievement in this cohort of students. Hence, achievement reaches potential at these upper levels, corresponding to optimal environmental conditions, thereby explaining that overachievement is a misnomer. The midpoint of the scatterplots, corresponding to $A = Pe^0$ describes an environmentally neutral context and implies that these students are nowhere near their upper limit and, therefore, also "underachieving".

By directly comparing the environmental conditions of students with different values of I_A for a given P, it may be possible to identify both the important enhancing and inhibiting environmental variables hat contribute to A as a function of P. Understanding these different environmental contexts makes it possible for education authorities to take appropriate intervention strategies to maximize A. It is this fine-grained characterization of students that overcomes many of the problems associated with both the conceptual basis and operationalization of underachievement.

The OAM also highlights the problems associated with the use of mean scores of student achievement as the basis for deciding who is underachieving. At all grade levels, it is possible to calculate the mean HKAT score and to draw a vertical line from this value on the corresponding scatterplots. At each end of this line will be students who unexpectedly overachieve and unexpectedly underachieve. The practical difficulty for a teacher, of course, will be to decide if a student with an "average" performance is at an acceptable level of performance.

Hong Kong is one of a number of East Asian countries that are recognised for the outstanding academic achievements of its students (Leung, 2002, Phillipson & Tse, 2007). This study also shows that the worldwide phenomenon of scholastic underachievement, particularly for students with high ability (see Stoeger & Ziegler, 2005) also includes students from Hong Kong. Indeed, the data from this study suggests that mathematical underachievement also increases with student potential. Furthermore, the proportion of students who underachieve is greatest in the highest grade level (Secondary 3). The reasons for this are as yet undetermined but could be associated with the highly pragmatic nature of Hong Kong culture (Phillipson, Shi, & Zhang, et al). The study of mathematics per se may not be as important for these students and the high levels of underachievement may be offset by high levels of interest and achievement in other subjects.

Acknowledgements

I would like to thank Bonnie Yuen-yan MAK for her management of this project and Pelly Pui-yee NG, Gigi Mei-chi CHEUNG, Ami Tsz-mei LAI, Winnie Wan-ha WONG, and Wai-huen LAU from the Fung Hon Chu Gifted Education Center for their cheerful assistance with data collection. I also thank Trevor G. BOND and Dr. Moritz Haene for a number of helpful discussions and comments on an early draft of this paper. This study was supported by an Internal Research Grant from The Hong Kong Institute of Education.

References

Adams, R. J., & Khoo, S. T. (1992). *Quest: The interactive test analysis system*. Hawthorn, AUSTRALIA: ACER.

Andersen, E. B. (1973). A goodness of fit for the Rasch model. *Psychometrika, 38*(1), 123-140.

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage.

Andrich, D., & Styles, I. (1994). Psychometric evidence of intellectual growth spurts in early adolescence. *Journal of Early Adolescence, 14*(3), 328-344.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Erlbaum Associates.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences (2nd ed.)*. Mahwah, NJ: Erlbaum Associates.

Chan, D. W. (1999). Reversing underachievement: Can we tap unfulfilled talents in Hong Kong? *Educational Research Journal, 14*(2), 177-190.

Deary, I. J., Egan, V., Gibson, G. J., Austin, E., Brand, C. R., & Kellaghan, T. (1996). Intelligence and the differentiation hypothesis. *Intelligence, 23*, 105-132.

Dixon, R. M., Craven, R. G., & Martin, A. J. (2006). Underachievement in a whole city cohort of academically gifted children: What does it look like? *The Australasian Journal of Gifted Education, 15*(2), 9-15.

Education and Manpower Bureau (2003a). *Hong Kong Attainment Tests Series 7 Mathematics Primary 5*. Hong Kong: Printing Department of the Government of the Hong Kong SAR.

Education and Manpower Bureau (2003b). *Teacher Handbook of Hong Kong Attainment Tests Series 7 Mathematics Primary 5*. Hong Kong: Printing Department of the Government of the Hong Kong SAR.

Fletcher, J. M., Denton, C., & Francis, D. J. (2005). Validity of Alternative Approaches for the Identification of Learning Disabilities: Operationalizing Unexpected Underachievement. *Journal of Learning Disabilities, 38*(6), 545-552.

Gagné, F. (2005). Transforming gifts into talents: The DMGT as a developmental theory. *High Ability Studies, 15*(2), 119-147.

Jacobs, N., & Harvey, D. (2005). Do parents make a difference to children's academic achievement? Differences between parents of higher and lower achieving students. *Educational Studies, 31*(4), 431-448.

Jensen, A. R. (1998). *The g factor: The Science of mental ability*. Westport, CT: Praeger.

Kubinger, K. D. (2005). Psychological test calibration using the Rasch model - Some critical suggestions on traditional approaches. *International Journal of Testing, 5*(4), 377-394.

Kubinger, K. D., Formann, A. K., & Farkas, M. G. (1991). Psychometric shortcomings of Raven's Standard Progressive Matrices (SPM) in particular for computerized testing. *European Review of Applied Psychology, 41*, 295-300.

Lane, K. L., Gresham, F. M., & O'Shaughnessy, T. E. (2002). Serving Students With or At-Risk for Emotional and Behavior Disorders: Future Challenges. *Education & Treatment of Children, 25*(4), 507.

Lau, K.-L., & Chan, D. W. (2001). Identification of Underachievers in Hong Kong: do different methods select different underachievers? *Educational Studies, 27*(2), 187-200.

Legault, L., Green-Demers, I., & Pelletier, L. (2006) Why do high school students lack motivation in the classroom? Toward an understanding of academic amotivation and the role of social support. *Journal of Educational Psychology, 98*(3), 567-582.

Leung, F. K. S. (2002). Behind the high achievement of East Asian students. *Educational Research and Evaluation, 8*(1), 87-108.

Linacre, J. M., & Wright B. D. (2004). *WINSTEPS: Multi-choice, rating scale, and partial credit Rasch analysis* [computer software]. Chicago: MESA Press.

Lissitz, R. W. (Ed.) (2005). *Value added models in education: Theory and applications.* Maple Grove, MN: JAM Press.

Liu, X., & Boone, W. J. (Eds.) (2006). *Applications of Rasch measurement in Science education and school achievement.* Maple Grove, MN: JAM Press.

Lynn, R., Allik, J., & Irwing, P. (2004). Sex differences on three factors identified in Raven's Standard Progressive Matrices. *Intelligence, 32*, 411-424.

Mackintosh, N. J., & Bennett, E. S. (2005). What do Raven's Matrices measure? An analysis in terms of sex differences. *Intelligence, 33*, 663-674.

Phillipson, S. N., & Callingham, R. (*in press*). Understanding mathematical giftedness: Integrating self, action repertoires and the environment. In L. V. Shavinina (Ed.). *The International Handbook on Giftedness*. Amsterdam: Springer Science and Business Media.

Phillipson, S. N., & Tse, K.-o. A. (*2007*). Discovering patterns of achievement in Hong Kong students: An application of the Rasch measurement model. *High Ability Studies, 18*(2), 173-190.

Phillipson, S. N., Shi, J., Zhang, G., Tsai, D.-M., Quek, C. G., Matsumura, N., & Cho, S. (*in press*). Recent developments in gifted education in East Asia. In L. V. Shavinina (Ed.). *The International Handbook on Giftedness*. Amsterdam: Springer Science and Business Media.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. (Expanded edition). Chicago: University of Chicago Press.

Raven, J. C., Court, J. H., & Raven, J. (1983). Manual for Raven progressive matrices and vocabulary scale. London: Lewis.

Smith, E., & Smith, R. (Eds.) (2004). *Introduction to Rasch measurement.* Maple Grove, MN: JAM Press.

Sokal, R. R., & Rohlf, F. J. (1981). *Biometry (2nd ed.)*. San Francisco, CA: W. H. Freeman and Co.

Staudt, B., & Neubauer, A. C. (2006). Achievement, underachievement and cortical activation: A comparative EEG study of adolescents of average and above-average intelligence. *High Ability Studies, 17*(1), 3-16.

Stoeger, H., & Ziegler, A. (2005). Evaluation of an elementary classroom self-regulated learning program for gifted mathematics underachievers. *International Education Journal, 6*(2), 261-271.

Styles, I. (1999). The study of intelligence-The interplay between theory and intelligence. In M. Anderson (Ed.), *The development of intelligence* (pp. 19-42), East Sussex, UK: Psychology Press Inc.

van der Linden, W. J., & Hambleton, R. K. (1987). Handbook of modern item-response theory. New York: Springer.

van der Ven, A. H. G. S., & Ellis, J. L. (2000). A Rasch analysis of Raven's standard progressive matrices. *Personality and Individual Differences, 29*(1), 45-64.

Vigneau, F. & Bors, D. A. (2005). Items in context: Assessing the dimensionality of Raven's advanced progressive matrices. *Educational and Psychological Measurement, 65*(1), 109-123.

Wilson, M. (2004). On choosing a model for measuring. In E. V. Smith & R. M. Smith (Eds.) *Introduction to Rasch measurement: Theory, models and applications*. (pp. 123-142). Maple Grove, MN: JAM Press

Ziegler, A. (2005). The Actiotope Model of Giftedness. In R. J. Sternberg & J. E. Davidson (Eds.), *Conceptions of Giftedness* (pp. 411-436). New York: Cambridge University Press.

Ziegler, A., & Stoeger, H. (2003). Identification of underachievement: An empirical study on the agreement among various diagnostic sources. *Gifted and Talented International, 18*(2), 87-94.

Ziegler, A., & Stoeger, H. (2004). Differential effects of motivational orientation on self-confidence and helplessness among high achievers and underachievers. *Gifted and Talented International, 19*(2), 61-68.