What makes performance tasks motivating: Influences of task characteristics, gender and ethnicity

Jeffrey K. Smith, Alison Gilmore, David Berg, Lisa F. Smith and Madgerie Jameson-Charles

Abstract

The characteristics of performance tasks that make them appealing to students in a low-stakes environment were investigated with data from the National Education Monitoring Project (NEMP) of New Zealand. Random samples of Year 4 (8-year-old) and Year 8 (12-year-old) students were assessed using a set of performance tasks in the areas of mathematics, information skills and social studies. Students indicated whether they particularly liked or disliked each task, or whether they were neutral in their reactions to them. Each task was scored on a set of task characteristics generated from the literature and from experienced assessment experts. The characteristics of tasks were then related to students’ liking of the tasks. Additionally, the effects of year in school, gender and ethnicity were examined, along with the influence of the individuals in charge of administering the tasks. Students’ liking of tasks was also related to their performances on those tasks.

One of the fundamental assumptions of the interpretation of an assessment performance is that the student has made a sincere effort to do well (Wise, 2006; Wise & DeMars, 2005). Without such an assumption, it becomes nearly impossible to make reasonable inferences about the student’s abilities. Wolf, Smith and Birnbaum (1995) demonstrated that students typically do not work as hard or perform as well when the assessment is of little or no consequence to them. Furthermore, they showed that when there is little or no consequence, this decline in effort increases as tasks become more mentally demanding. In a meta-analysis of motivation and testing studies, Wise and DeMars (2005) found an effect size (Hedges’g)
of .59 for the effect of comparing motivated and nonmotivated groups on test performance. In two studies using Trends in International Mathematics and Science Study (TIMSS) data with Swedish students, Eklof (2007, 2010) found that Grade 8 students had a moderate to high level of motivation for performance, but that Grade 12 students had poor motivation for performance. Grade 8 students said that they felt it was important to do well to represent their country; hence, they actually did feel that the assessment was of consequence to them. Working with university students, Barry, Horst, Finney, Brown and Kopp (2010) demonstrated that in order to make valid inferences in low-stakes tests, it is critical to understand the level of effort that students make. Although the findings from the literature are not uniform, it is fairly clear that when students view an assessment as being of no consequence to them, there is the potential for their performance to be impaired, and, with that impairment, issues of validity in interpretation of results arise.

The literature in this area uses four closely related terms: stakes, consequence, motivation and effort. In this research, we introduce a fifth concept: likeability. Let us start by clarifying the relationships among stakes, consequence, motivation and effort. To begin, stakes and consequence are essentially the same idea. They both have to do with what the results of an assessment mean to a student (Porter, Linn, & Trimble, 2005). For example, if a student will either receive credit in a course or not depending upon the assessment result, then the result has high consequence for the student. Assessment can have consequences for other groups as well (e.g., teachers, schools), but our concern here is with students. Motivation is a well-known psychological construct that “involves any force that energizes and directs behavior” (O’Donnell, Reeve, & Smith, 2012, p. 341). If motivation is a force that energises and directs behaviour, then effort might best be conceptualised as the energy and direction that is has been engendered by the motivational force. With regard to assessment activities, effort can be defined as the mental engagement and work that is done in an attempt to address the task. Thus, stakes or consequence lead to motivation, which in turn leads to effort (mental energy expended in addressing the requirements of the assessment). In a low-stakes assessment environment, there is no extrinsic motivation to do well, so the motivation for an effortful performance must
be intrinsic. We argue that when students enjoy working on a task—when they like the task—they will put forth the requisite effort to be successful. We develop that idea in the following paragraphs.

Hidi and Harackiewicz (2000), using terms first described by Dewey (1913), differentiated between school activities that “catch” a student’s attention and those that “hold” such attention. They argued that triggering interest (catching) is not as important for long-term educational purposes as maintaining (holding) levels of interest. We agree that for activities that require sustained, long-term engagement, holding is more important than catching. But for assessment activities, which are typically short-term activities, what may be essential is engaging short-term interest, or the interest that is sufficient to engage the student productively through to the completion of the assessment item or task. This interest might be generated by the consequence of the assessment for the students (Wise, 2006; Wise & Smith, 2007; Wolf & Smith, 1995), or it might be generated by the items or tasks being inherently interesting, especially in a low-stakes situation. Thus, the motivation to perform well may be extrinsic or intrinsic in nature, or a combination of the two. Our interest here is in the low-stakes setting where intrinsic motivation is key.

If it is important that the required activities of an assessment capture the attention and interest of students, the question arises as to what are the characteristics of assessment activities that make them interesting? That is the question of interest in this research. If we need to develop assessments that engage students and hold their interest through to completion, can we know what characteristics engender such engagement? This would be useful not only for the development of tasks for use in national assessment programmes, but for classroom use as well.

Powers and Fowles (1999) looked at writing prompts that were motivating for postgraduate students. Their list of characteristics that students found engaging included prompts that let students draw on their personal experiences, prompts that students had strong feelings about or felt that they could relate to and prompts that were clearly stated in terms of what was expected. However, as their work was conducted only in the area of writing and they were working with university students, this
list may not generalise to other age groups or areas. In New Zealand, Eley and Caygill (2001) found that Year 4 students (roughly 8 years old) and Year 8 students (roughly 12 years old) enjoyed tasks that involved the use of equipment or manipulable materials, and that involved one-to-one participation with an interviewer. They did not enjoy tasks that were multiple choice or short answer in nature.

Approaching the issue of motivation and assessment performance from a different perspective, Stiggins and Conklin (1992) developed the notion of a classroom assessment environment that they argued influences how students view assessment. Building on this work, Brookhart and her colleagues (e.g., Brookhart & Durkin, 2003; Brookhart, Walsh, & Zientarski, 2006) examined a variety of issues related to the motivation and expended mental effort that students put forward on assessments and concluded, among other findings, that the items, or tasks, that comprise assessments need to engage students’ interest. They also presented insights on the “assessment climate” that exists within the classroom. They argued that some classrooms develop a positive attitude toward the use of assessments while others develop a negative attitude, and, in turn, the assessment climate influences learning. Thus, not only the tasks themselves might hold interest for students, but the context in which they are given might also influence the effort a student puts forth.

Our goal in this study was to build on the literature and examine how characteristics of assessment activities relate to primary students’ levels of liking of those activities. Our fundamental argument is that, in a low-stakes setting, if students like tasks, they will be motivated to invest more effort in them than if they do not. Thus, we argue that the likeability of a task (if students report that they liked a task) is a reasonable surrogate for the notion of interest. If students like a task, they invest more effort in working on the task, and a better picture of their abilities emerges. The data set that we used came from New Zealand’s national educational monitoring programme at Year 4 and Year 8 (the same programme as in the 2001 Eley and Caygill study). In that programme, the assessment activities are referred to as “tasks” and we will use that terminology throughout the remainder of the paper. In particular, we were interested in investigating the following questions:
1. What characteristics of tasks are associated with students’ liking of the tasks?
2. Are there differences in students’ liking by gender, year in school or ethnicity?
3. Are there differences by subject matter among mathematics, social studies and information skills?
4. Does the environment (assessment climate) influence student liking?
5. What is the relationship between liking a task and performing well on it?

**Method**

This study used assessment tasks that were completed by a sample of Year 4 (roughly 8-year-old) and Year 8 (roughly 12-year-old) students participating in New Zealand’s 2005 National Education Monitoring Project (NEMP). As part of NEMP, participating students indicated the level to which they liked each task in which they participated. A variety of analyses are presented here, using different subsets of the NEMP data set.

We first focus on the tasks that were particularly well liked or disliked in an effort to uncover the underlying characteristics that were liked or not liked by these students. Second, we conduct a study to test whether those characteristics were actually predictive of likeability. Third, we look at differences in liking according to year in school, gender, ethnicity (Pākehā, Māori or Pasifika) and subject matter. Fourth, we look at the issue of assessment climate (albeit only tangentially) by examining whether there are differences in student liking and student performance when broken down by the administrators of the assessment. And fifth, we look to see if performance on the tasks was related to liking of the tasks.

**Background**

NEMP assessed student performance for Year 4 and Year 8 students across 12 different curricular areas, primarily using performance assessment. The NEMP assessment was conducted annually between 1995 and 2009, and provided New Zealand educators with a picture of performance in three to four curricular areas per year. The programme was concluded in
2009, and a successor programme is under development as of the time of writing. The areas studied in any particular year changed based on a 4-year cycle that covered most of the New Zealand curriculum. The data presented in this study are from the 2005 administration, for which the three curriculum areas were mathematics, social studies and information skills.

**Participants**

The sample for NEMP was drawn from a two-stage procedure. Students from both public and private schools were included in the study. First, random samples of schools were selected for participation (separate samples for Year 4 and Year 8). Schools were then invited to participate in the programme. Over 95 percent of schools invited agreed to participate. Participating schools provided student lists and random samples of students were drawn from those lists. Twelve students were selected from each participating school. When a school did not have 12 students in a given year level, that school was paired with a neighbouring small school, and the sample of 12 was taken from a combination of those school rolls. Students, with parental consent, were then invited to participate. It should be noted than any students with limitations that would have severely affected their ability to successfully participate in the project were not included in the final selection. In total, there were 1,440 students at Year 4 and the same number at Year 8 in the 2005 samples. Almost all students invited to participate agreed to do so, yielding an overall participation rate of those invited of over 95 percent.

Once the samples were selected, the participating students were randomly assigned to one of three “teams” labelled A, B and C. The total sample had 480 students in each of these three teams at each year level. In any given school (or paired schools), each team (A, B and C) was made up of four students, and each team completed a set of tasks unique to that group. Individual tasks usually had 3–10 percent fewer students participating due to illness, occasional equipment failure, etc. To keep the analyses from becoming overly complicated, we focus our attention here on the 60 tasks that were given to both Year 4 and Year 8 students in the A team of the administration. These included 31 maths tasks, 13 information skills tasks
and 16 social studies tasks. Hence, the sample for this research consisted of 480 students at Year 4 and another 480 students at Year 8, all taking the same set of 60 tasks, described below.

**Materials**

The tasks that the participants completed were developed by NEMP staff, all experienced educators, in collaboration with national advisory committees of subject matter experts. They were pilot tested before their final administration. A variety of task formats was used. Tasks were administered in one of four modes: a one-to-one interview with a teacher-administrator; independent work involving manipulable materials or computer-based administration; independent work involving just paper and pencil; and team tasks involving teams of four students working together. Some of the tasks took as little as a minute to complete; team tasks took up to 10 minutes. Tasks typically involved a topic or theme, with a number of questions and prompts on that theme. Questions were typically open-ended, and some called for student opinions on issues, or asked students how they would go about trying to solve a problem.

Student responses were recorded on videotape, computer or on paper, and were scored either by experienced teachers or by upper-level university students studying to become teachers.

**Procedure**

The NEMP tasks were administered by two-person teams of experienced teachers. All administrators were trained in a week-long session held the week before they began administration of the tasks.

There were 24 teams for the Year 4 administration and another 24 teams for Year 8. Each pair of teacher-administrators worked in five schools, with a total of 60 students.

The student participants completed the tasks in 1-hour blocks of time with four blocks over the course of a week comprising their total involvement. At the end of each session, the students were asked to assess their reactions to each task. They were asked to mark each task as being one that they
“particularly liked”, “particularly did not like” or “didn’t like or dislike”.
These three options were represented to students with a “happy face”,
a face that was neutral and an “unhappy face”; these were explained to
the students. This approach for determining student reaction to the tasks
had been used in the programme effectively since its inception. The
term like instead of how motivated was used in the NEMP evaluation
for several reasons. One reason was historical; that was the way the
question was always phrased in the NEMP assessments. Additionally,
like was considered to be a word that students at Year 4 and Year 8 would
comprehend and relate to easily. For the purposes of this research, we feel
that the inference from students liking a task to being motivated to do well
on it is reasonable and conservative.

Measures
We organised the data in two fashions, once with the participants as
the unit of analysis and once with the tasks as the unit of analysis. For
participants, the following information was available: how well they
performed on each of 60 tasks; how much they liked each task; year in
school; age; ethnicity; and gender. For tasks, we have measures of how
well the task was liked by the participants, and five characteristics of
the task: content relevance; modality; clarity; structure/freedom; and task
demand. These characteristics are explained in the Results section below.

Results
Overall likeability, and likeability by year in school and subject
matter
The first set of analyses examined students’ overall liking of tasks, and
addressed the differences by year in school and by subject matter. To
begin, all students tended to regard the tasks favourably. Table 1 presents
the percentages of students reporting “particularly liked”, “neutral” or
“particularly didn’t like” summed over all tasks and all students. Year 4
students were more positive overall than Year 8 students in each subject
area; that is, for almost every individual task, the Year 4 rating was higher
than that of Year 8. Year 4 students liked the information skills tasks best,
while Year 8 ratings were roughly equal across the three subject areas. Social studies tasks were the least favoured by both groups, although the ratings were still substantially positive. Year 4 students were much less likely to utilise the neutral category in the ratings as compared to Year 8 students.

Table 1 Percentages of likeability by subject matter and year of student

<table>
<thead>
<tr>
<th>Rating</th>
<th>Information skills (13 tasks)</th>
<th>Social studies (16 tasks)</th>
<th>Mathematics (31 tasks)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Year 4</td>
<td>Year 8</td>
<td>Year 4</td>
</tr>
<tr>
<td>Particularly liked</td>
<td>83.0</td>
<td>63.6</td>
<td>68.2</td>
</tr>
<tr>
<td>Neutral</td>
<td>4.7</td>
<td>21.3</td>
<td>14.8</td>
</tr>
<tr>
<td>Particularly didn’t like</td>
<td>12.3</td>
<td>15.1</td>
<td>17.0</td>
</tr>
</tbody>
</table>

Characteristics related to liking

Having looked at overall likeability, and the relationship to year in school and to subject matter, the next challenge was to try to determine the characteristics of tasks that made them appealing to students. We used three approaches in generating a list of characteristics that we thought would make tasks likeable to students. First, we started with the work of Powers and Fowles (1999) and Eley and Caygill (2001), described above. Second, we interviewed four NEMP team members involved in the development and marking of the tasks about what they felt were the characteristics of tasks that made them popular or unpopular with students. These NEMP team members had a long tenure on the programme; their years of working directly with thousands of children on NEMP tasks provided them with clear ideas about what leads tasks to being liked or not liked.

Finally, we took a subset of the 60 tasks for more careful examination by the authors of this study. We wanted to have a sample of tasks that had equal numbers in each subject area to examine. We began by taking a sample of 36 tasks, 12 in each subject area, and identified the highest and lowest scoring four tasks in terms of likeability in each of the three subject areas. These were then analysed by three members of the author
team to determine what the underlying characteristics of the tasks were. After examining the tasks that were listed least and most favoured, and discussing results with the experienced NEMP staff mentioned above, we devised the following five categories that seemed to be the best determinants of task likeability. The characteristics are not dichotomous in nature; tasks may have more or less of each of the characteristics. It should be noted that these characteristics comprise a list generated through expert judgement. The degree to which they actually predict likeability is tested empirically below. The characteristics have been categorised as content relevance, modality, clarity, structure/freedom and task demand.

**Content relevance.** The content of the tasks, as opposed to the skills required, task demands etc., appeared to be an important factor in likeability. The argument here is that students like those things that are relevant and/or of interest to them. This is fundamentally the same thing that Powers and Fowles (1999) found with students taking the Graduate Record Examination, only related to younger students. In looking at the sample of highly and lowly rated tasks, the tasks that got the highest ratings involved sports, pets, space, money, food, comics and popular youth culture. The tasks with the lowest ratings tended to involve current events of a political nature, famous people not of the students’ era and cultures other than their own.

**Modality.** Modality refers to physical activities involved in a task. It appeared that tasks that involve “doing something”, such as in computer-based tasks or tasks that involve working with objects, building things and experimenting with things, are more likeable than those that involve less activity. Tasks that only involved paper and pencil seemed to be less popular. Additionally, tasks that involved students doing things that might be viewed as embarrassing in front of other students (e.g., speaking publicly) appeared to be not liked.

**Clarity.** Students appeared to like clarity in instruction (again, similar to Powers and Fowles, 1999). From looking at their likes and dislikes, it seemed that when they were not certain what was expected of them, they did not like the activity. We acknowledge that this required an inference on the part of the research team in terms of what would and would not be clear to students, but, as will be shown, we test this contention in the
study. In some instances, lack of clarity might have been attributable to characteristics of the student rather than the task; but in either case, it appeared that students liked to know what was expected of them.

*Structure/freedom.* This is a somewhat subtle characteristic. Tasks that focused on student opinion, without necessarily having a right or wrong answer, or those that allowed for students to feel in control of the task appeared to be liked by students. However, tasks were not liked if they were open-ended in a fashion that students found to be confusing. Choice in response also appeared to lead to students liking a task, although the set of tasks examined here did not include many opportunities for choice.

*Task demand.* Task demand relates to what the students are required to do—the underlying cognitive or physical component of the task. Tasks that required students to call on skills that they believed they have, even to solve a novel problem, appeared to be liked. Tasks that called on students to tap a knowledge base they thought was weak appeared to be less popular. Students seemed to like to feel that they were being successful and enjoyed demonstrating knowledge or skills in which they had a degree of confidence. Because this characteristic of student likeability was determined through interviews with the NEMP staff, not from analysis of the student likeability data, it is not possible to know whether students actually felt they had the necessary skills on a given task. This characteristic is probably best conceptualised as an interaction between the student and the task, rather than a uniform characteristic of the task itself. Due to that complexity, we ultimately decided that we could not provide ratings for this concept, and do not include it as part of our analyses. We present it here because it appears to be an important, if somewhat intractable, notion.

Once we had developed a set of characteristics that we felt were related to the likeability of the tasks to the students, we conducted a validation analysis to determine how closely these characteristics were related to the students’ liking. We had two professional educators independently rate each of the 36 tasks according to the first four of the characteristics thought to be related to student likeability (content relevance, structure, modality and clarity), using a scale of 1 = low to 5 = high. One of these educators has extensive background in educational measurement, and
the other has extensive classroom experience. Attempts at rating task demand seemed too subjective, so we did not include it in the analyses. The means, standard deviations and inter-rater reliabilities (based on Pearson correlations) for the four scales are presented in Table 2. After the independent ratings, the raters discussed their ratings and resolved discrepancies into a single rating for each task characteristic that was used in subsequent analyses.

Table 2 **Means, standard deviations and inter-rater reliabilities of task characteristics**

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Content relevance</th>
<th>Structure/freedom</th>
<th>Modality</th>
<th>Clarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (SD)</td>
<td>2.72 (1.25)</td>
<td>3.19 (0.86)</td>
<td>2.92 (0.87)</td>
<td>3.14 (.96)</td>
</tr>
<tr>
<td>Reliability</td>
<td>.77</td>
<td>.61</td>
<td>.62</td>
<td>.70</td>
</tr>
</tbody>
</table>

Note: Reliabilities based on correlations between two independent raters.

The reliabilities were modest for structure/freedom and modality, but stronger for clarity and content relevance. The structure/freedom characteristic was difficult to rate, as it involved a degree of understanding of how children would perceive a task. The low reliability for modality appeared to be due to situations where a task appeared on the surface to be interactive, but upon closer examination, was not. For example, one task appeared to use the computer, but in fact it merely had a “screen shot” of a computer page that was used by students for information. Thus, it wasn’t a computer interactive task, but one that had a computer screen for illustration. Different raters might see this task differently. All reliabilities were acceptable for general research purposes, although it must be acknowledged that the structure/freedom and modality reliabilities are low.

Our next step was to estimate the intercorrelations among these variables and their correlations with the students’ liking of the various tasks used in this analysis. We used the mean likeability for the Year 4 and Year 8 students combined as the likeability measure. These results are presented in Table 3.
Table 3 Intercorrelations among characteristics and student liking for 36 tasks

<table>
<thead>
<tr>
<th></th>
<th>Liking</th>
<th>Content</th>
<th>Structure/relevance</th>
<th>Modality freedom</th>
<th>Clarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Liking</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Content relevance</td>
<td>.655</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Structure/freedom</td>
<td>.132</td>
<td>.185</td>
<td>.328</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Modality</td>
<td>.398</td>
<td>.551</td>
<td>.328</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Clarity</td>
<td>.423</td>
<td>.175</td>
<td>-.034</td>
<td>.116</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Note: Bolded correlations are significant at \( p < .05 \), with Bonferroni correction.

As can be seen from the correlation matrix, three of the four characteristics investigated showed moderate to strong relationships to the student likeability data. Only structure/freedom did not show a significant correlation. It is particularly interesting that content relevance showed such a strong relationship to students liking the tasks. Although concepts such as clarity, modality and structure/freedom require a fair amount of work to build into a task (or to ensure that the characteristic exists), it is less difficult to make certain that the content of a task is pertinent to students. For example, it is easy enough to change a maths task from sections of rods to sections of a chocolate bar, or a social studies task from a debate in a town hall to a debate in a school.

To see how these characteristics related to students’ liking as a whole, we ran a four variable multiple regression analysis, with liking as the dependent variable, and the four characteristics under consideration as independent variables. We found that the regression was significant overall, with \( F(4, 31) = 8.71, p < .001 \), adjusted \( R \)-square = .529. The characteristics of content relevance and clarity were significant predictors \((b = .925, t = 3.86, p = .001\) for content relevance, and \(b = .061, t = 2.537, p = .016\) for clarity). Structure/freedom and modality were not significant predictors in the regression. We examined residuals to look for evidence of nonlinearity, but found none.
Gender and ethnicity differences

Gender and ethnic differences were examined by conducting a series of chi-squared tests on contingency tables for likeability responses for each task. We used chi-squared tests as there were only three levels of response on the likeability tasks, and for many tasks there were very few “didn’t particularly like” responses. Only Year 8 data were used for these analyses, as many of the Year 4 tasks did not have enough variability to conduct statistical analyses with sufficient power. We used an alpha = .05 level of significance without a Bonferroni adjustment as we wanted to be able to detect patterns across tasks showing significance and hence did not want the analysis to be highly conservative, as would be the case given the number of comparisons made. We were mindful of the number of differences that might come out by chance (an expected value of 6 given 120 comparisons). Ethnicity data were taken from students’ school records, using the categories Māori, Pākehā and Pasifika. Māori are the indigenous people of New Zealand; Pasifika is a term that refers to peoples of the Pacific Islands (e.g., Samoa, Tonga); Pākehā is a Māori term used to refer to New Zealanders of European ancestry. There were 60 analyses conducted for gender and ethnicity, 120 in total. We found 14 tasks to be significant for gender and a different set of 14 tasks for ethnicity. Māori and Pasifika students were generally more positive about tasks with a Māori or Pasifika theme than were Pākehā students. Girls were more interested than boys in tasks that featured girls, and vice versa. There were many more nonsignificant findings than significant ones in these analyses, suggesting that, overall, tasks were either liked or disliked across the board.

Differences by teacher-administrator team

The next analysis examined the issue of the potential influence of an “assessment environment” in terms of whether liking tasks was related to the teachers who administered the tasks. In working with the teacher-administrator teams, and reviewing the videotapes of the students taking the tasks, it appeared that some teacher-administrator teams were more enthusiastic and positive in their administration of the assessments than were other teacher-administrator teams. We reasoned that this might show up in terms of how positively the students viewed the tasks. We
should note that even though we think this is a worthwhile analysis, we do not believe that it provides a strong test of Stiggins and Conklin’s (1992) notion of classroom assessment environments, as that idea encompasses a host of characteristics not present here. What we wanted to explore in this analysis was whether the different dispositions of the administration teams would be reflected in the students’ liking of tasks.

To that end, we generated two scales, one comprised of the likeability scores for all maths tasks (31) and a second comprised of all social studies and information skills tasks (29). The decision to organise the tasks into these two scales was based on an exploratory factor analysis on the 60 tasks. A scree plot showed two strong factors, one consisting of the maths tasks and a second consisting of the social studies and information skills tasks. After these two factors, eigenvalues dropped substantially and then remained fairly flat, indicating that a two-factor solution was the best to use. We then calculated coefficient alpha reliabilities on the two scales. Reliability for the maths tasks was .81 and for the social studies and information skills was .86.

Next, a multivariate analysis of variance was used to look for differences among the 24 teacher-administrator teams (for Year 8 students only, for reasons of having sufficient variability, as described above). The dependent measures were the total likeability score for the maths tasks and the total likeability scores for the information skills and social studies (combined) tasks; the independent variable was the teacher-administrator team that administered the tasks to the students. Using Wilks’ Lambda, the differences among teacher-administrator teams fell short of statistical significance, $F(46, 758) = 1.07, p = .325$. Thus, we have no evidence here of teacher-administrators influencing how well the students liked the tasks. As mentioned above, we acknowledge that this is not a strong examination of the argument for an assessment environment. Tasks were administered one-on-one, so it might have been the case that one administrator was more effective than the other, and the administration of the tasks is different from a classroom setting. We decided to include it as it shows that likeability was not strongly related to who was administering the tasks.
The relationship between liking a task and doing well on it

The final analysis was to examine the relationship between liking a task and doing well on it. We ran 60 one-way analyses of variance, using the liking score as the independent variable, and the task performance score as the dependent variable. We then conducted post hoc tests to compare performance between students who said they particularly liked a task and those who said they particularly did not like the task. A total of 42 of the 60 tasks showed a significant difference at alpha = .05. We did not use a Bonferroni correction here, but point out that with an alpha = .05, we would have expected three of these comparisons to be significant simply due to random variation. We did this because we wanted to see how many of the tasks would be significant at .05, rather than restricting the alpha for each individual test. In each of these cases, the students who liked the task outperformed those who did not. Thus, we have fairly strong evidence of the relationship between liking the task and performing well on the task.

Discussion

Perhaps the strength of the analyses presented here is that these were responses to a large set of tasks in three distinct subject areas given to a national random sample of students at two different year levels. Thus, it allowed for a fairly broad look at how students responded to performance-based assessment tasks that were low stakes for the students. What we found is that students appeared to respond favourably to tasks that showed some mix of the following characteristics:

- personal relevance for them in terms of content
- clear instructions as to what is required
- an active response modality.

There was no evidence to support the argument on freedom and structure. We think that the characteristics of freedom and structure, and task demand, are areas that would benefit from more careful conceptualisation and study. We think this is an important area, but are not confident we have it “right” at this point.
Looking at the data from the perspective of year in school, gender and ethnicity, we found that younger students were more generally positive than older students. We also found that there were some gender and ethnic differences. Boys and girls tended to favour tasks that were of particular interest to their gender. Some ethnic differences were found as well, again with tasks being more popular when they pertained to the specific group. For both gender and ethnicity, many more tasks were found with no differences than with differences. Differences among subject areas were small, although other subject areas may show larger differences. There were no differences found according to which team of teachers administered the tasks. Generally speaking, these findings suggest that the nature of the task itself is the most important factor in determining how students will respond to it from an affective perspective. And, finally, most tasks showed that students who liked the task outperformed those who did not.

These findings raise some interesting questions, both in terms of interpretation, and in terms of practical consequence. If students report that they do not like tasks on which they did not perform well, what is the causal factor? That is, do they not like them because they felt they were not doing well and were upset over that fact; or, did they not like them and consequently did not put forth sufficient effort to do well on them? Clearly, these analyses cannot answer the issue of causality here. However, hours of watching videotapes of students responding to these tasks leads us to believe that the answer is not simply that one is causal of the other. Motivation regarding a task evolves throughout the student’s efforts and self-perception of success, but what is clear is that the nature of the task influences how the student approaches it from the outset. Tasks that are not engaging initially rarely grab a student’s attention halfway through the task. On the other hand, some tasks seem appealing at the outset to students who then become frustrated with their efforts and disengage. As mentioned above, the motivation of students, and how they view a task in terms of what is required of them and their potential for success, is an area where more work is needed.

Finally, there is clearly a link between liking a task and performing well on it. As has been pointed out consistently in research over the past 15 years, motivation is a factor that should not be ignored when looking at
test performance, particularly on assessments that are low stakes in nature to the students who are taking them.

References


**The authors**

Jeffrey K. Smith is Professor and Associate Dean—Research of the College of Education at the University of Otago. He is the former Chair of the Department of Educational Psychology at Rutgers University and Head of the Office of Research and Evaluation at the Metropolitan Museum of Art. He did his undergraduate work at Princeton University and his PhD at the University of Chicago. His research interests include assessment, learning and the psychology of aesthetics.

Email: jeffrey.smith@otago.ac.nz

David Berg is a lecturer at the University of Otago College of Education. He is an experienced primary teacher and former deputy principal with international experience. His research interests include initial teacher education, assessment, psychology of education and international education. David holds an EdD awarded by the University of Otago.

Email: david.berg@otago.ac.nz

Lisa F. Smith is Professor of Education and Dean at the University of Otago College of Education. Lisa’s research focuses on assessment issues related to both standardised and classroom testing, preservice teacher efficacy and the psychology of aesthetics. She is a foundation member of the New Zealand Assessment Academy. Lisa received her doctorate in Educational Statistics and Measurement from Rutgers University in New Jersey and has won teaching awards in both hemispheres.

Email: lisa.smith@otago.ac.nz

Alison Gilmore is Associate Professor at the University of Otago and Director of the National Monitoring Study of Student Achievement. She has been associated with the National Education Monitoring Project since 1993. Alison has worked in the educational assessment and evaluation fields for nearly 30 years both in New Zealand and internationally. A current research focus is on building assessment capability.

Email: alison.gilmore@otago.ac.nz

Madgerie Jameson-Charles, PhD, is a lecturer in Fundamentals of Education Research; Assessment and Evaluation; and Health and Family Life Education at the School of Education, The University of the West Indies, St. Augustine Campus, Trinidad. Her research interests are high-stakes testing; alternative assessment; education for employment; making transitions; and learning and instruction in higher education. She received her PhD from the University of Otago.

Email: madgeriej@gmail.com