New rules, new tools

Niessen, Anna Susanna Maria

Chapter 2

Predicting performance in higher education using content-matched predictors

A version of this chapter was published as:
Niessen, A. S. M., Meijer, R. R., & Tendeiro, J. N. (2016).
Predicting performance in higher education using
proximal predictors. *PLoS ONE,* 11(4): e0153663.
doi:10.1371/journal.pone.0153663

Abstract

We studied the validity of two methods for predicting academic achievement and student-program fit that were matched to the study content. Applicants to an undergraduate psychology program participated in a selection procedure consisting of a curriculum-sampling test based on a performance-sampling approach, and specific skills tests in English and math. Test scores were used to predict academic performance and progress after the first year, performance in specific course types, enrollment, and dropout after the first year. All tests showed positive significant correlations with the criteria. The curriculum-sampling test was consistently the best predictor in the admission procedure. We found no significant differences between the predictive validity of the curriculum-sampling test and prior educational performance, and substantial shared explained variance between the two predictors. Only applicants with lower curriculum-sampling test scores were significantly less likely to enroll in the program. In conclusion, the curriculum-sampling test yielded predictive validities similar to that of prior educational performance and possibly enabled self-selection. In admissions aimed at student-program fit, or in admissions in which past educational performance is difficult to use, a curriculum-sampling test may be a good instrument to predict academic achievement.

2.1 Introduction

There is an increasing interest in the content validity of instruments used for prediction and selection in higher education (e.g., Schmitt, 2012). Especially in many European countries where students apply to a specific study program rather than to a college, there is a trend towards selecting students based on admission tests that show correspondence to the program content. This trend is opposed to selecting students on the basis of more general admission criteria such as scores on general cognitive tests, personality questionnaires, or prior educational performance.

Content-matched predictors for academic success consist of tasks that require similar skills for success as the criterion measures. Content-matched tests have been extensively studied in predicting job performance and were found to be among the most valid predictors (Ployhart, Schneider, & Schmitt, 2006). Examples are job-knowledge tests, assessment centers, and work samples. In their meta-analysis, Schmidt and Hunter (1998) found that work sample tests were among the most valid test for predicting future job performance. However, despite the good results obtained in predicting job performance and the current use of such methods to select students for higher education in, for example, the Netherlands (Visser, van der Maas, Engels-Freeke, & Vorst, 2012) and Finland (Häkkinen, 2004) they have hardly been studied empirically within the context of higher education.

The aim of this study was to fill this gap in the literature and to investigate the predictive validity of content-matched tests for predicting academic achievement and student-program fit in an actual academic selection context. Most studies that investigate new methods to predict academic achievement use data collected in low-stakes conditions (e.g., Schmitt, 2012; Shultz & Zedeck, 2012). We investigated the predictive validity of a curriculum-sampling test, based on a performance-sampling approach analogous to work samples, and two specific skills tests for predicting academic achievement in high-stakes selection procedure for a psychology program. Doing so, we provide empirical evidence that is badly needed to justify the use of these selection methods in institutes of higher education. The curriculum-sampling test was designed to mimic a representative course in the program and the specific skills tests were designed to measure skills that were relevant for successful performance in specific courses.

2.1.1 Content-matched Predictors for Academic Achievement

Specific skills tests

A limited amount of studies have been conducted in which the predictive validity of specific skills tests was investigated for predicting academic outcomes. Most studies were conducted in the context of predicting graduate school performance. Kuncel, Hezlett, and Ones (2001) performed a meta-analysis across multiple disciplines and found that the specific subject tests of the Graduate Record Examinations were the best predictors for graduate school GPA in a study that also included verbal, quantitative and analytic ability, and undergraduate GPA. Furthermore, the specific subject tests alone predicted academic outcomes almost as well as composite scores of several general and subject-specific predictors. Kuncel, Hezlett, and Ones (2001) explained these results through the similarity of the subject tests with the criteria used. Additionally, Kuncel and Hezlett (2007) reviewed several studies and meta-analyses in predicting graduate school success and concluded that the strongest predictors were tests that were specifically linked to the discipline of interest.

Work sample tests

In behavioral prediction, a distinction can be made between signs and samples as predictors of future behavior. Sign-based tests measure a theoretical construct (e.g., intelligence, personality) that is conceptually related to the criterion. Sample-based tests aim to sample behavior or performance that is representative for the criterion of interest, based on the notion that current behavior is a good predictor for future behavior (Wernimont & Campbell, 1968). Tests for predicting educational performance have been mostly sign-based, measuring constructs such as cognitive abilities (Eva, 2003; Lievens & Coetsier, 2002). However, Wernimont and Campbell (1998) discussed that using behavior- or performance sampling in prediction resulted in greater predictive validity than using signs of behavior. Also, Asher and Sciarrino (1974) stated that the more a predictor and a criterion are alike, the higher the correlation is expected to be; “Information with the highest validity seems to have a point-to-point correspondence with the criterion” (p. 519).

Work sample tests are “high-fidelity assessment techniques that present conditions that are highly similar to essential challenges and situations on an actual job” (Thornton & Kedharnath, 2003, p. 533) and meet the criteria of performance sampling and point-to-point correspondence. As discussed above, Schmidt and Hunter (1998) also found in their meta-analysis that work sample tests were the best predictors of job performance. Callinan and Robertson (2000) suggested that work samples perform well in predicting future performance because they measure a complex combination of individual abilities and skills that yield a higher validity than when these abilities and skills are measured separately. They also suggested that work samples contain a motivational component that is related to future performance. Some studies also suggested that work samples could enhance self-selection of applicants, both with respect to interests and abilities (Breaugh, 2008; Downs, Farr, & Colbeck, 1978), and could therefore potentially reduce turnover. These characteristics also make the work sample approach appealing to use in admission to higher education. Curriculum-sampling tests are based on the work sample approach applied in the context of higher education.

Curriculum-sampling tests

Curriculum-sampling tests are performance samples that are constructed as simulations of academic programs or representative parts of academic programs. We are aware of two studies that used curriculum-sampling tests to predict performance in higher education (Lievens & Coetsier, 2002; Visser et al., 2012). Besides these two studies, there are a few studies about admission procedures for medical school that included similar methods (Schripsema, van Trigt, Borleffs, & Cohen-Schotanus, 2014; Urlings-Strop, Stegers-Jager, Stijnen, & Themmen, 2013), but they did not report validity coefficients for separate sections of the procedure, so we do not discuss them here.

Lievens and Coetsier (2002) studied a cohort of medical students and dentistry students who participated in an admission exam consisting of several cognitive tests, two curriculum-sampling tests, and two situational judgment tests. They found that a cognitive reasoning test showed the largest relationship with first year mean grade, followed by the curriculum-sampling tests, with medium-sized relationships. However, the reliabilities of the curriculum-sampling tests were low, which likely had a negative influence on the estimated correlation coefficients. Visser, van der Maas, Engels-Freeke, and Vorst (2012) studied a curriculum-sampling test administered to select applicants for an undergraduate psychology program. The curriculum-sampling test mimicked the first course in the program because results showed that the first grade obtained in higher education was a very good predictor for later academic performance (Busato, Prins, Elshout, & Hamaker, 2000). Applicants who were rejected based on the test or had not participated in the selection procedure could still get admitted through a lottery procedure. Visser et al. (2012) found that applicants admitted on the basis of the curriculum-sampling test dropped out less often, earned higher grades, and obtained more course credit in the first year than applicants who were rejected by the test.

2.1.2 Educational Context

Content-matched methods are particularly suitable when students apply directly to a program in a specific discipline, such as professional schools and graduate schools in the USA (like medical school or law school), and undergraduate programs and master programs in Europe. There are a number of reasons why especially the European higher education system is suitable for using content-matched methods for selecting students. First, students often choose a specific program in which they major before starting undergraduate education, and they often apply directly to the educational program (e.g., medicine, psychology, or law). Second, many European countries have a certain degree of stratification in secondary education, with the best performing students attending the highest level of education. Only students that finished the appropriate secondary education program are eligible to apply to a university. In addition, graduation often depends on nationally or centrally organized final exams based on a national curriculum. Thus, there is a well-controlled central system and there is severe pre-selection on learning abilities for admission to higher education. This limits the utility of traditional predictors that measure general cognitive skills. Therefore, general cognitive tests are not often used. Finally, there is an increasing amount of international applicants (e.g., Schwager, Hülsheger, Bridgeman, & Lang, 2015), which makes it difficult to use previous educational performance as a selection criterion in practice.

2.1.3 Aims of the Present Study

The aim of the present study was to investigate the use of specific skills tests and the curriculum-sampling test to predict performance in higher education and student-program fit. The curriculum-sampling test was constructed to mimic the first courses in the program, so that the test had a high similarity to tasks that students are expected to perform. The specific skills tests were not designed to mimic the program, but covered specific subjects that were considered important for successful performance in specific courses. The tests were administered in an actual admission procedure. We examined the predictive validity of these tests for first year academic achievement and performance in types of specific course. In addition, we compared the predictive validity of these tests to that of prior educational performance, one of the best general predictors for academic achievement in higher education (e.g., Atkinson & Geiser, 2009; Peers & Johnston, 1994; Westrick, Le, Robbins, Radunzel, & Schmidt, 2015). Furthermore, we explored the relationship between admission test scores and enrollment decisions to explore the presence of a self-selection effect.

2.2 Method

2.2.1 Participants

The sample consisted of 851 applicants for an undergraduate psychology program in the academic year 2013-2014 at a Dutch university. All applicants participated in the selection procedure containing two specific skills tests and a curriculum-sampling test. Of all applicants, 652 started the psychology program and 199 did not. The selection committee eventually rejected none of the applicants because the number of enrollments did not exceed the number of available places. Note that the students did not know this beforehand and thus the selection was perceived as high stakes and the applicants were likely to be very motivated to perform well. Sixty-nine percent of the applicants were female and the mean age was 20 for the entire applicant group (*SD* = 2.3) and also 20 (*SD* = 2.0) for the group that enrolled in the program. The students followed their courses in English or in Dutch, with similar content. The English program consisted of mainly international students. Fifty-seven percent of the applicants followed the English program. Forty-three percent of all applicants were Dutch, 43 percent were German, 10 percent had another European nationality, and four percent had a non-European nationality.

2.2.2 Materials and Procedure

Curriculum-sampling test

The curriculum-sampling test was designed to simulate a representative course in the first year. The psychology program requires a substantial amount of self-study and the students' main tasks are studying books and syllabi, and attending lectures. However, attending the lectures is not mandatory. At the end of most courses, a multiple-choice exam is administered. To trigger future student-behavior, the curriculum-sampling test mimicked the first course in the program: *Introduction to Psychology*. This course covered general psychological principles and theories. The applicants received two chapters from the book used in this course and were instructed to study them. One chapter was about research methodology, an important topic in this program, and one chapter was about more general psychological theories. The test consisted of 40 multiple-choice items and was constructed by a faculty-member who teaches first-year courses.

Skills tests

The applicants also completed specific skills tests in English reading comprehension and mathematics. English reading comprehension was included because most study material is in English, even in the Dutch program. The test consisted of 20 items and was constructed by a faculty member who is a professional translator. The items consisted of fill-in-the gap exercises and questions about the meaning of texts. Mathematical skills were tested because the psychology curriculum includes a number of courses in statistics. The math skills included in the test were selected for their relevance to the statistics courses in the program. The test consisted of 30 items and was constructed by a faculty member who teaches first-year statistics courses. The applicants did not receive specific material to prepare for the skills tests, but useful webpages and example items were provided for the math test.

Selection procedure

After applying to the program, all applicants were invited to visit the university to take the admission tests. Each test had to be completed within 45 minutes with 15-minute breaks in between the tests. Proctors were present to prevent cheating. Applicants who had a valid reason for not being able to attend (living or working outside of Europe) could complete the admission tests online. Thirteen percent of the applicants used this option (10% of the enrolled students). Each test score was the sum of the number of items answered correctly. All applicants received feedback after a few weeks, including their scores on each test and a rank based on a composite score of the individual test scores. Students that held the lowest 165 ranks were contacted by phone and encouraged to rethink their enrollment, but this advice was not binding.

High school grades

In addition, high school grades were collected through the university administration for research purposes. High school grades were only available for students who completed the highest level of Dutch secondary education (vwo). Table 2.1 shows the sample sizes for each variable and for each combination of variables. The grades were self-reported but verified by the central education administration. Grades were on a scale of one to ten with ten being the highest score. We calculated high school GPA (HSGPA) using the grades on all courses taken by a student, except courses that only provided a pass/fail result. The grade on a national final exam made up 50% of most final grades, the other 50% of the final grade was accounted for by exams administered by the schools in the last three years of secondary education.

Academic achievement

Three measures of first-year academic achievement were used: the first year mean grade for academic performance (FYGPA), the number of obtained credits (FYECT) for academic progress, and dropout. Academic achievement data were collected through the university administration after one academic year. Grades were on a scale of one to ten, with ten being the highest grade and a 6 or higher representing a pass. FYGPA was computed for each student, using the highest grade for each course after two exam opportunities (exam and resit) had taken place. One course resulted in a pass/fail decision and was not taken into account. The FYGPA consisted of 10 exam results when a student participated in all courses. Some students did start the program but did not participate in any exams. The resulting sample size for FYGPA and combinations with other variables are shown in Table 2.1. Credit was granted after a course was passed and for most courses students earned five credit points, with a maximum of 60 credits in the first year, resulting in the first-year degree. Dropout records were also obtained from the administration.

Since the specific skills tests were designed to predict performance for certain types of courses, we also computed a composite mean grade for statistics courses (SGPA) and theoretical courses (TGPA). The SGPA is the mean of the final grade for two statistics courses and the TGPA is the mean final grade for seven courses that were concerned with psychological theory and required studying literature and completing an exam. Sample sizes for the number of students are also shown in Table 2.1. Because we only used data available at the university, there were no manipulations in this study, and no identifiable information was presented, informed consent was not obtained. This study was approved by and in accordance with the rules of the Ethical Committee Psychology from the University of Groningen.

Table 2.1
Sample sizes for each variable and combinations of variables in the study, for applicants who enrolled.

Variable	1.	2.	3.	4.	5.	6.
1. Admission tests	851					
2. HSGPA	203	203				
3. FC grade	626	198	626			
4. FYGPA	638	201	626	638		
5. FYECT	652	203	626	638	652	
6. Dropout	652	203	626	638	652	652
7. SGPA	590					
8. TGPA	635					

Note. Sample sizes for each variable are on the diagonal. HSGPA = high school mean grade, FC grade = first course grade, FYGPA, first year mean grade, FYECT = number of credits obtained in the first year, SGPA = statistics courses mean grade, TGPA = theoretical courses mean grade.

2.2.3 Procedure

Correlations were computed between the test scores and the academic achievement measures. For significance tests we used $\alpha = .05$. Before conducting the analyses, we conducted t-tests to check if there were test score differences between the applicants who took the tests online and those who took the tests proctored. We assumed that if the online applicants had cheated this would result in higher scores for these applicants as compared to those in the proctored group. For predictive validity we expected that scores on all tests would show significant positive relationships with all performance criteria, but that the curriculum-sampling test would be the best predictor because it showed the most correspondence to the program.

To assess the validity of the curriculum-sampling test, we assessed the relationships between the first course grade (*Introduction to Psychology*), the curriculum-sampling test, and academic achievement in the first year. For these analyses results from the first course were excluded from the FYGPA and the number of obtained credits. In addition, we assessed relationships between the test scores and performance in specific course types, that is, the mean grade on the statistics courses, and the mean grade on the theoretical courses. For this purpose, multiple regression analyses were conducted with the test scores as independent variables and achievement in the courses as dependent variables. Squared semi-partial correlations were inspected to assess the unique contributions of the predictors. We expected that scores on the math test would be the strongest unique contributor to predicting the mean statistics grade, and that the curriculum-sampling test score would show the largest unique contribution to the mean theoretical grade, followed by the score on the English test.

To assess if the curriculum-sampling test was a good alternative to using high school grades for applicants who completed Dutch secondary education, we compared the correlations of the curriculum-sampling test scores and academic achievement with the correlations between HSGPA and academic achievement, using Williams test for differences between two dependent correlations (Steiger, 1980). We had no a priori expectation about the direction of these differences. In addition, we assessed the unique contributions of HSGPA and the curriculum-sampling test score to predict academic achievement. For FYGPA and FYECT as dependent variables, multiple regression analyses were conducted with the curriculum-sampling test score and high school grades as predictors. Squared semi-partial correlations were inspected to assess the unique contributions of both predictors. For dropout, a logistic regression analysis was conducted with, again, the curriculum-sampling test scores and HSGPA as predictors. As a proxy to semi-partial correlation in least-squared regression, pseudo-partial correlations, also known as Atkinson's *R*, were computed and inspected (Hox, Moerbeek, & van der Schoot, 2010). While these coefficients cannot be directly compared to results obtained in least-squares regression, they do provide an indication of the contribution of each variable to the model.

Finally, we investigated whether the admission tests may have resulted in self-selection using logistic regression analyses with enrollment as the dependent variable and the test scores as independent variables, while controlling for receiving a phone call to encourage reconsidering enrollment. High school grades were not assessed for a self-selection effect, since they were not part of the admission procedure, the students received no feedback with respect to high school grades, and they were collected for research purposes only.

2.3 Results

2.3.1 Predictive Validity

Before computing correlations between the test scores and academic achievement, t-tests were conducted to check for differences in tests completed online or proctored. The applicants in the online test group obtained a lower mean score than the applicants in the proctored group for the curriculum-sampling test and the English test and a higher mean score for the math test, but the latter difference was not significant ($t_{(849)}$ = 1.81, *p* = .07, Cohen's *d* = 0.18). Based on these results there was no evidence that cheating seriously raised scores in the online group, and we merged the results for both groups together for all analyses. Descriptive statistics for the admission test scores, HSGPA, academic achievement, and the correlations between these variables are shown in Table 2.2. The reliability estimates of the admission tests were satisfactory and all admission tests showed significant correlations in the expected direction with all academic-performance criteria.

The curriculum-sampling test was the best predictor for all performance measures and showed a large correlation with FYGPA (*r* = .49) and moderate correlations with for obtained credits and dropout (*r* = .39 and *r* = -.32). The math test and the English test showed moderate correlations with FYGPA (*r* = .29 and *r* = .25) and small correlations with obtained credits and dropout (*r* ranging between -.13 and .20). Note that, as intended, the first course *Introduction to psychology* was strongly positively related to the curriculum-sampling test (*r* = .56). Also, the grade in the first course was strongly related to all academic-performance criteria in the first year.

Table 2.2

Descriptive statistics and correlations between the predictors and academic achievement measures.

Variable	*M*	*SD*	1	2	3	4	5	6	7
Admission tests									
1. Curriculum-sampling test	29.7	5.2	.77						
2. Math test	16.6	4.7	.30 [.23, .37]	.76					
3. English test	13.7	3.3	.43 [.37, .49]	.21 [.14, .28]	.69				
Prior educational performance									
4. HSGPA	6.7	0.4	.45 [.33, .55]	.36 [.23, .47]	.29 [16, .41]				
Academic achievement									
5.First course grade	6.6	1.4	.56 [.50, .61]	.20 [.12, .27]	.34 [.27, .41]	.55 [.45, .64]			
6. FYGPA	6.6	1.3	.49 [.43, .55]	.29 [.22, .36]	.25 [.18, .32]	.52 [.41, .61]	.75[c] [.71, .78]		
7. FYECT	46.0	20.2	.39 [.32, .45]	.20 [.13, .27]	.16 [.09, .23]	.30 [.17, .42]	.62[c] [.57, .67]	.82 [.79, .84]	
8. Dropout[a]	0.20[b]		-.32 [-.39, -.25]	-.15 [-.22, -.08]	-.13 [-.20, -.05]	-.22 [-.35, -.09]	-.47 [-.53, -.41]	-.64 [-.70, -.58]	-.83 [-.85, -.81]

Note. HSGPA, high school mean grade; FYGPA, first-year mean grade. Internal consistency coefficients are on the diagonal (Cronbach's alfa). 95% confidence intervals for the population correlation ρ are between brackets. [a] Point-biserial correlations. [b] Proportion. [c] For these correlations, results on the first course were not included in the calculation of FYGPA and credits. All correlations were significant with $p < .01$.

2.3.2 Predictive Validity for Specific Course Performance

Results of multiple regression analyses with the scores on the admission tests as independent variables and mean grades for statistics courses and theoretical courses as the dependent variables are shown in Table 2.3. The correlation between the mean grade on the statistics courses and the mean grade on the theoretical courses was $r = .67$, 95% CI [.62, .71], showing that they are strongly related but can be distinguished. Zero-order correlations between the admission test scores and specific course performance were all positive and statistically significant. For both specific course types, scores on the English test did not significantly contribute to the explained variance of the model when the curriculum sample scores and the math scores were included.

Table 2.3
Multiple regression results predicting specific course performance with the admission test scores

Predictor	SGPA			TGPA		
	β	*r*	sr^2	β	*r*	sr^2
Curriculum-sampling score	.29*	.34*	.07*	.45*	.51*	.16*
Math score	.27*	.34*	.07*	.10*	.25*	.01*
English score	-.07	.11*	< .01	.06	.27*	< .01
F	44.22*			78.31*		
R^2	.19			.27		

* $p < .05$

The curriculum-sampling test scores and the math scores predicted the mean statistics grade equally well with moderate effect sizes ($r = .34$, for both tests), and showed equal unique contributions to the model ($sr^2 = .07$ for both tests). This only partly confirmed our expectations because we hypothesized that the math test would be the strongest predictor for statistics performance.

The curriculum-sampling test score showed a large positive relationship with the mean theoretical grade ($r = .51$) and the math score and the English score showed small to moderate positive relationships ($r = .25$ and $r = .27$). The unique contribution was the largest for the curriculum-sampling scores ($sr^2 = .16$) and very small to non-existent for the math scores and the English scores. This also partly confirmed our expectations, since a unique contribution of the English scores was expected.

2.3.3 Comparing Curriculum Sampling to Prior Educational Performance

For applicants who completed Dutch secondary education, the mean high school grade also showed significant correlations with all academic-performance criteria, with a large effect size for FYGPA (r = .52), a moderate effect size for obtained credits (r = .30), and a small effect size for dropout (r = -.22). To compare the predictive validities of HSGPA and the curriculum-sampling test we computed the correlations again for only the students with available data for HSGPA, the curriculum-sampling test, and the academic achievement measures. The correlations between the curriculum-sampling score and the academic achievement measures were slightly lower within this group than for the entire sample (FYGPA, r = .41, FYECT, r = .27, and dropout, r = -.26). Taking into account the correlation between the curriculum-sampling score and HSGPA (r = .45), the results of William's test showed no significant difference between the predictive validity of the curriculum-sampling score and HSGPA for FYGPA, with $t_{(198)}$ = -1.75, p = .08. There was also no significant difference in predictive validity for FYECT, with $t_{(200)}$ = -0.14, p = .89, and no significant difference in predictive validity for dropout, with $t_{(200)}$ = -0.56, p = .58.

To assess the unique contributions and overlap between these two predictors for academic performance (FYGPA) and progress (FYECT) in the first year, multiple regression analyses were conducted and semi-partial correlations were assessed. For predicting FYGPA, the unique contribution for HSGPA was sr^2 = .15, and for the curriculum-sampling test it was sr^2 = .04 (with $F_{(2,198)}$ = 44.45, p <.01 and R^2 = .31). Hence, the shared explained variance for FYGPA by HSGPA and the curriculum-sampling score equaled .12. Thus, for applicants with Dutch secondary education, HSGPA uniquely explained more variance in the FYGPA than the curriculum-sampling score, whereas they also shared a substantial part of explained variance. For predicting obtained credits, the unique contribution of HSGPA was sr^2 = .04, and the unique contribution for the curriculum-sampling score was sr^2 = .03 (with $F_{(2,200)}$ = 14.02, p <.01 and R^2 = .12). The shared explained variance for obtained credits by HSGPA and the curriculum-sampling score equaled .05. The uniquely explained variance for each predictor and the shared explained variance for obtained credits were of similar magnitude.

For dropout as the dependent variable, the logistic regression model with HSGPA and the curriculum-sampling score as independent variables was significant ($\chi^2_{(2)}$ = 17.02, p < .01, and Nagelkerke's pseudo R^2 = .13). Pseudo-partial correlations equaled *pr* = .09 for HSGPA and *pr* = .13 for the curriculum-sampling score. Thus, the unique contribution of the curriculum-sampling test when taking HSGPA into account was slightly larger than vice versa.

2.3.4 Self-selection

Descriptive statistics for enrolled and not-enrolled applicants are presented in Table 2.4. Enrollment was predicted based on the admission test scores using logistic regression analysis, controlling for receiving a phone call to reconsider enrollment after scoring among the 165 lowest ranks. Results for the logistic regression analysis are in Table 2.5. The model was statistically significant for predicting enrollment. The curriculum-sampling score was the only significant predictor in the model. A one-unit increase in the curriculum-sampling score increased the odds of enrolling by a factor of 1.02 to 1.09, when the other test scores and receiving a discouraging phone call were held constant.

Table 2.4

Means and standard deviations for applicants who did enroll and did not enroll in the program

Variable	Enrolled	Not enrolled
Phone call[a]	.14	.35
Curriculum-sampling score	29.7 (5.2)	27.0 (6.5)
Math test	16.6 (4.7)	14.9 (4.7)
English test	13.7 (3.3)	12.6 (3.6)

Note. Standard deviations are between brackets. [a] Proportion of students who received a discouraging phone call within the enrolled and non-enrolled group.

2.4 Discussion

The results of this study showed that all content-matched tests predicted academic achievement in the first year. The predictive validity of the curriculum-sampling test was moderate to large for academic achievement in the first year, whereas the predictive validities for the specific skills tests were small to moderate. The results also showed that the first course in the program was a very good predictor for performance in the rest of the first year, replicating results by Busato et al. (2000). Furthermore, scores on the curriculum-sampling test were related to student-program fit, as shown by a moderate relationship with dropout and a small but significant relationship with enrollment decisions.

Table 2.5

Logistic Regression Results for Predicting Enrollment Based on Selection-test Scores

Variables		*B*	*SE(B)*	Wald X^2	*df*	*p*	e^B	95% CI e^B
Phone call		.53	.29	3.38	1	.07	1.70	0.97, 2.99
Curriculum-sampling score		.05	.02	7.67	1	.01	1.05	1.02, 1.09
Math score		.03	.02	1.67	1	.20	1.03	0.77, 1.07
English score		.01	.03	0.06	1	.80	1.01	0.95, 1.07
Model X^2	46.44				4	<.01		
n	851							

The specific skills tests did not predict performance in specific related course types better than the other tests. An interesting result was that the curriculum-sampling test predicted performance in statistics courses equally well as the math test, and that the English test was not a better predictor than the math test for the grades in theoretical courses. A possible explanation for these results is that the curriculum-sampling test, following a performance-sampling approach, measures both ability and motivation to perform well (e.g., Callinan & Robertson, 2000). Such an implicit behavioral measurement of motivation may explain the relationships between the curriculum-sampling test and academic performance, even when the course content was different from the curriculum-sampling test. After all, motivation and effort are necessary for successful performance in any course. As de Raad and Schouwenburg (1996) stated 'achievement through ability alone is the exception rather than the rule'. Lievens and Coetsier (2002) found lower predictive validities using curriculum-sampling tests, but their tests had relatively low reliability and they were not specifically designed to mimic relevant parts of the programs. Furthermore, the predictive validities of the curriculum-sampling tests scores did not significantly differ from the predictive validities of high school GPA, one of the most established predictors of academic achievement in higher education. Additionally, the regression results obtained using this subsample showed that the curriculum-sampling score and HSGPA shared a substantial proportion of explained variance in academic performance. The HSGPA uniquely explained more variance in performance, and the curriculum-sampling test had a slightly larger unique contribution to predicting dropout.

It is important to note that although HSGPA is a good predictor for academic achievement, a drawback in practice is that these grades are not always available and/or are difficult to compare across applicants, as we explained above. Furthermore, an advantage of using content-matching tests is that these tests could help provide insight in what the study program is like, and what is expected of the applicants when they are accepted as students. This could result in a self-selection effect, and our results showed that applicants with lower scores on the curriculum-sampling test were significantly less likely to enroll in the program, even after controlling for actively discouraging low-scoring applicants to enroll. However, the effect was small and we do not know if the decision to enroll or not was based on the experience in or results of the admission procedure. It is possible that applicants who decided not to enroll were already less motivated or uncertain about their choice, and did not prepare well for the tests as a result. Another advantage of content-matched predictors is that applicants are not 'haunted by their past'. In contrast to HSGPA, which are fixed and cannot be altered by the applicants, content-matched tests provide applicants an opportunity to show their ability and motivation for the study program.

2.4.1 Limitations

In this study we used a sample of applicants from one cohort of students in one discipline, and obtained criterion measures after one academic year. Although previous studies found strong relationships between academic performance in the first year and in later years (e.g., Busato et al., 2000) data that provides insight in predictor-criterion relations collected after a longer period of time is needed. The predictive validity is expected to decrease somewhat when academic achievement is measured with a larger time interval. In addition, prior educational performance could only be studied for applicants who applied to the program after completing Dutch secondary education at the level that traditionally allows admission to universities. Approximately two-thirds of the students had a different educational background. However, this also illustrates that using prior educational performance, as an admission tool, is difficult to realize in practice.

Furthermore, constructing content-matched tests for programs like psychology is relatively straightforward. Many academic undergraduate programs, like psychology, require mostly independent studying, attending lectures, and completing exams. However, it may be more challenging to develop such tests for programs that are more directed towards the mastery of practical skills. For example in medical school, teacher training, or vocational education, skills such as motor skills or communication skills may have predictive value. These skills are more complicated to assess. In addition to a curriculum-sampling test measuring ‘classic’ student behavior like studying literature, content-matching methods can also be used to measure nonacademic skills. An example is the multiple mini-interview (MMI) used to assess applicants to medical school. The MMI can be used to assess applicants on moral reasoning, communication skills, and social skills. MMI scores predicted clerkship performance, and performance on clinical skills examinations (Eva, Reiter, Rosenfeld, & Norman, 2004; Eva et al., 2009; Reiter, Eva, Rosenfeld, & Norman, 2007). Lievens (2013) found that scores on SJT’s used to select applicants for medical school predicted especially the more practically and interpersonally oriented outcomes, whereas cognitive (skills) tests predicted those outcomes to a lesser or no extent. However, effect sizes were mostly small to moderate and based on data obtained in low-stakes conditions (e.g., Niessen & Meijer, 2016).

Also, curriculum-sampling tests have to be constructed for each program, preferably with new items each time the test is administered. Standardized tests are usually carefully constructed, analyzed, and checked with respect to difficulty level and psychometric quality. Constructing this curriculum-sampling test was not more time-consuming than constructing a typical exam. However, a potential drawback is the risk of unsatisfactory test-quality. Close attention should be paid to characteristics such as difficulty, item quality, and reliability.

Finally, our results showed that the predictors in this study explained roughly up to 25% of the variance depending on the outcome measure and predictor. This may seem low to some critics. However, it is good to remember that, as for example, Dawes (1979) argued, many critics implicitly assume that the remaining 90% through 75% of the variance can be explained. Considering the complex nature of the outcomes that we want to predict (that is, student performance in the future), we may not expect much better results. Indeed, in the context of predicting academic achievement, the highest predictive validities found in many studies are around r = .60 after correcting for range restriction and unreliability (Kuncel et al., 2001; Kuncel & Hezlett, 2007).

2.4.2 Conclusion

This study showed that a performance-sampling approach can be implemented successfully in the context of higher education. A question that can be addressed in the future is whether the favorable characteristics of content-matching and performance-sampling approaches found in research in personnel selection, such as perceived fairness and face validity (Anderson, Salgado, & Hülsheger, 2010), also extend to an educational context.

In our study, both prior educational performance and the curriculum-sampling test yielded moderate to large predictive validities, whereas the specific skills test showed smaller effect sizes. When information about prior educational performance is available, comparable, and verifiable for the majority of applicants, this information may be the most effective and efficient approach to select applicants. When this is not the case, using a curriculum-sampling test is a good alternative and may be preferred over specific skills tests. Contexts in which content-matched tests could be preferred over traditional admission criteria are admission procedures with an emphasis on assessing student-program fit that aim for high content validity. An example is the mandatory matching procedure in the Netherlands. Applicants to open-admission programs are required to participate in a 'matching' procedure organized by the individual study programs. The result is a non-binding advice about enrollment based on student-program fit. When constructing curriculum-sampling tests for other programs, we recommend to start with an analysis of the study program and to identify representative courses for the program that show a high relationship with performance in the rest of the program.