Chapter 3
A multi-cohort study on the predictive validity and construct saturation of high-fidelity curriculum-sampling tests

This chapter was based on the manuscript:
Abstract

We investigated the validity of curriculum-sampling tests for admission to higher education in two studies. Curriculum-sampling tests mimic representative parts of an academic program to predict future academic achievement. In the first study, we investigated the predictive validity of a curriculum-sampling test for first year academic achievement across three cohorts of undergraduate psychology applicants and for academic achievement after three years in one cohort. We also studied the relationship between the test scores and enrollment decisions. In the second study, we examined the cognitive and noncognitive construct saturation of curriculum-sampling tests in a sample of psychology students. The curriculum-sampling tests showed high predictive validity for first year and third year academic achievement, mostly comparable to the validity of high school GPA. In addition, curriculum-sampling scores showed incremental validity over high school GPA. Applicants who scored low on the curriculum-sampling tests decided not to enroll in the program more often, indicating that curriculum-sampling admission tests may also promote self-selection. Contrary to expectations, the curriculum-sampling tests scores did not show any relationships with cognitive ability, but there were some indications for noncognitive saturation, mostly for perceived test competence. So, curriculum-sampling tests can serve as efficient admission tests that yield high predictive validity. Furthermore, when self-selection or student-program fit are major objectives of admission procedures, curriculum-sampling test may be preferred over or may be used in addition to high school GPA.
3.1 Introduction

Curriculum-sampling tests are increasingly used in admission procedures for higher education across Europe. For example, in Finland, Belgium, the Netherlands, and Austria these tests are used across various academic disciplines such as medicine (de Visser, et al., 2017; Lievens & Coetsier, 2002; Reibnegger et al., 2010), psychology (Niessen, Meijer, & Tendeiro, 2016; Visser, van der Maas, Engels-Freeke, & Vorst, 2012), teacher education (Valli & Johnson, 2007), economics and business (Booij & van Klaveren, 2017), and computer science (Vihavainen, Luukkainen, & Kurhila, 2013). The rationale behind these tests is to mimic later behavior that is expected during an academic study. Thus, these tests often mimic representative parts of academic programs such as studying domain-specific literature or watching a video-lecture, followed by an exam. However, they can also consist of giving a demonstration lesson in the case of teacher education (Valli & Johnson, 2007), and they can take different forms, such as a massive online open course (MOOC) for programming skills (Vihavainen et al., 2013) or a ‘virtual semester’ in medical school (Reibnegger et al., 2010).

There are several arguments for using curriculum samples in admission to higher education in addition to or instead of traditional admission criterion used in higher education such as high school GPA. High school GPA is a good predictor of academic performance in higher education (e.g., Peers & Johnston, 1994; Westrick, Le, Robbins, Radunzel, & Schmidt, 2015; Zwick, 2013). However, due to the increasing internationalization and different educational ‘routes’ to higher education (Schwager, Hülsheger, Bridgeman, & Lang, 2015), these grades are often difficult to compare across applicants. Also, Sackett, Walmsley, Koch, Beatty, and Kuncel (2016) found that matching the content of the predictor to the criterion was beneficial to the predictive validity.

There are, however, few studies in which the predictive validity of curriculum sampling tests has been investigated and the existing studies (Lievens & Coetsier, 2002; Niessen et al., 2016) used single cohorts and only used first year academic achievement as a criterion measure. To our knowledge, there are no predictive validity studies using outcomes such as graduation rates and long-term college GPA. Furthermore, there is only one study (Lievens & Coetsier, 2002) in which the “construct saturation” (i.e., the degree to which the score variance reflects construct variance; e.g., Roth, Bobko, & McFarland, & Buster, 2008) of curriculum-sampling tests was investigated. Investigating the construct saturation of curriculum-sampling tests may help to explain their predictive value.

To address these shortcomings in the literature we conducted two studies. The aim of the first study was to investigate the predictive validity of curriculum-sampling
tests in different cohorts, using not only first year GPA but also third year GPA and bachelor-degree attainment. The aim of the second study was to investigate the construct saturation of curriculum sampling tests to gain more insight into how these test scores are related to psychological constructs.

### 3.1.1 Signs, Samples, and Construct Saturation

Most traditional assessments for performance prediction, like cognitive ability tests and personality inventories, are based on a signs approach. Signs are psychological constructs that are theoretically linked to the performance or behavior of interest. In contrast, samples are based on the theory of behavioral consistency: Representative past or current behavior is the best predictor for future performance or behavior (Callinan & Robertson, 2000, Wernimont & Campbell, 1968). The samples approach originated in the context of personnel selection, where high-fidelity simulations like work sample tests and assessment centers were good predictors of future job-performance (Hermelin, Lievens, & Robertson, 2007, Schmidt & Hunter, 1998; Roth et al., 2008). An explanation for the high predictive validity of sample-based assessments is the ‘point-to-point correspondence’ of the predictor and the criterion (Asher & Sciarrino, 1974).

It is often assumed that sample-based assessments are multifaceted compound measures that are saturated with cognitive- and noncognitive constructs that also underlie performance on the criterion task (Callinan & Robertson, 2000; Thornton & Kedharnath, 2013). The construct saturation of such an assessment represents the degree to which score variance reflects variance on different constructs. The concept of construct saturation may seem in conflict with the underlying idea of a samples approach (Lievens & De Soete, 2012). However, construct saturation can affect the predictive validity of sample-based assessments and the size of subgroup score differences (Dahlke & Sackett, 2017; Lievens & Sackett, 2017; Roth et al., 2008). Furthermore, investigating construct saturation is helpful to gain an understanding of what sample-based assessments scores represent in relation to certain psychological variables. Finally, noncognitive saturation may provide great benefits in high-stakes assessment. Noncognitive constructs like personality traits, self-efficacy, and self-regulation are good predictors of future job-performance and academic performance, and show incremental validity over cognitive abilities (Richardson, Abrahams, & Bond, 2012). However, they are difficult to measure validly in high-stakes assessment due to faking (Niessen, Meijer, & Tendeiro, 2017b; Peterson, Griffith, Isaacson, O’Connell, & Mangos, 2011). A performance-based assessment method that is able to capture noncognitive traits and skills may provide a solution to that problem.
3.1.2 Curriculum Sampling: Existing Research

Curriculum-sampling tests follow the same rationale as work sample tests in personnel selection (Callinan & Robertson, 2000; Wernimont & Campbell, 1968); they mimic representative parts of an academic program. Often, they mimic an introductory course of a program, because performance in such courses is a good indicator for later academic performance (e.g., Bacon & Bean, 2006; Busato, Prins, Elshout, & Hamaker, 2000; Niessen et al., 2016).

Most studies on curriculum sampling compared the academic performance of students admitted through curriculum-sampling procedures with students admitted via other methods. Results showed that students admitted through a curriculum-sampling procedure earned higher grades, progressed through their studies faster, and dropped out less often compared to students admitted via lottery (de Visser et al., 2017; Visser et al., 2012) or via traditional entrance tests or matriculation exam grades (Vihavainen et al., 2013). However, participation in the curriculum-sampling admission procedures was voluntary in these studies, and admission could also be achieved through other procedures. Thus, these differences may also be caused by, for example, motivation; that is, highly motivated applicants may have chosen to participate in the curriculum-sampling procedure that requires effort, whereas less motivated applicants may have chosen an alternative route (Schripsema, van Trigt, Borleffs, & Cohen-Schotanus, 2014).

Reibnegger et al. (2010) compared the dropout rate and time to completion of the first part of medical school for three cohorts of students admitted through open admission, a ‘virtual semester’ in medicine followed by a two-day examination, or secondary school-level knowledge tests about relevant subjects. The best results were found for the cohort admitted through the ‘virtual semester’. However, the selection ratio was also the lowest for that cohort, which may have influenced the results.

Lievens and Coetsier (2002) examined the predictive validity of two curriculum-sampling tests for medical school for first year GPA, one using a video-based lecture ($r = .20$) and one using written medical material ($r = .21$). However, these tests had relatively low reliability ($\alpha = .55$ and .56, respectively). There were moderate relationships between the scores on these tests and scores on a cognitive ability test ($r = .30$ and $r = .31$, respectively), indicating at least some cognitive ability saturation, but they found no relationships with scores on the Big Five personality scales. In addition, Niessen et al. (2016) found that the predictive validity of a curriculum-sampling test for undergraduate psychology applicants was $r = .49$ for first year GPA, $r = .39$ for credits obtained in the first year, and
Chapter 3

\[ r = -0.32 \] for dropping out of the program in the first year. Furthermore, the curriculum-sampling test scores were related to subsequent self-chosen enrollment, indicating that it may serve as a self-selection tool. Booij and van Klaveren (2017) found similar results based on an experiment in which applicants to an economics- and business program were randomly assigned to a non-binding placement procedure consisting of an intake interview or a curriculum-sampling day without a formal exam or assignment. Compared to the interview condition, fewer students from the curriculum-sampling condition enrolled and more students passed the first year. Enrolling students who participated in the curriculum-sampling procedure also reported that the program met their expectations more often than enrollees from the interview condition.

### 3.1.3 Aim of the Current Studies

In the present chapter, we extended the results from chapter 2. In two studies we investigated the validity of curriculum-sampling tests designed for the admission procedure of a psychology undergraduate program at a Dutch university. In the first study we investigated the predictive validity of curriculum-sampling tests for first year academic achievement using data from three cohorts of applicants, and for academic achievement after three years using data from one cohort of applicants. The academic outcomes were academic performance, study progress, and college retention. In addition, we studied (1) the incremental validity of the curriculum-sampling scores over high school GPA, (2) the predictive validity of curriculum-sampling tests for achievement in specific courses as compared to specific skills tests designed to predict performance in those courses (e.g., Sackett et al., 2016), and (3) the relationship between curriculum-sampling tests scores and self-chosen enrollment decisions.

We studied two types of curriculum-sampling tests: A literature-based test and a video-lecture test. Given the correspondence between criterion and the predictors, we expected to replicate the high predictive validity results from Niessen et al. (2016) for first year academic achievement and we expected somewhat lower predictive validity for later academic achievement. In addition, we expected that the skills tests would predict performance in courses that they were designed to predict (the math test for statistics courses and the English test for theoretical courses), and that the curriculum-sampling tests would predict performance in both types of courses, but that the correlation with statistics course performance would be lower compared to the math test.

In the second study we investigated the hypothesis that the curriculum-sampling tests are saturated with both cognitive and noncognitive constructs. Assuming that
the curriculum-sampling test scores represent a ‘sample’ of future academic performance, we expected that the curriculum-sampling test scores would be saturated with variables that also predict academic performance, such as cognitive ability (e.g., Kuncel & Hezlett, 2007, 2010) and several noncognitive constructs and behavioral tendencies (Credé & Kuncel, 2008; Richardson et al., 2012). To investigate this hypothesis, we studied if and to what extent the scores on the curriculum-sampling test can be explained by cognitive ability, conscientiousness, procrastination tendencies, study-related cognitions, and study strategies (Credé & Kuncel, 2008; Richardson et al., 2012).

**Study 1: Predictive Validity**

**3.2 Method**

**3.2.1 Procedure**

All applicants to an undergraduate psychology program at a Dutch university were required to participate in an admission procedure. In 2013 and 2014, the admission procedure consisted of the administration of a literature-based curriculum-sampling test and two skills tests: A math test and an English reading comprehension test. In 2015, a math test and two curriculum-sampling tests (a literature-based test and a video-lecture test) were administered. Administration time for each test was 45 minutes, with 15-minute breaks in between. Each test score was the sum of the number of items answered correctly. Applicants were ranked based on a composite score with different weights for the individual tests in each cohort. The highest weight was always assigned to the literature-based curriculum-sampling test. All applicants received feedback after a few weeks, including their scores on each test and their rank. In addition, the lowest ranking applicants (20% in 2013 and 15% in 2014 and 2015) received a phone call to discuss their results with an advice to rethink their application. However, the selection committee did not reject applicants because the number of applicants willing to enroll did not exceed the number of available places. The applicants did not know this beforehand, thus the applicants perceived the admission procedure as high stakes. The study program and the procedure could be followed in English or in Dutch. Applicants to the English program were mostly international students.

**3.2.2 Materials and Measures**

*Curriculum-sampling tests*

The literature-based curriculum-sampling test was used in each cohort, and was designed to mimic the first-year course *Introduction to Psychology*. The applicants were instructed to study two chapters of the book used in that course. The second curriculum-sampling test, only administered in 2015, required applicants to watch a twenty-minute video lecture on the topic *Psychology and the Brain*. A lecturer
who taught the course in the first year provided the lecture. At the selection day, the applicants completed multiple-choice exams about the material. The exams were similar to the exams administered in the first year of the program and were designed by faculty members who taught in the first year. The first curriculum-sampling test consisted of 40 items in 2013 and 2014, and of 39 items in 2015. The second curriculum-sampling test consisted of 25 items. The exams consisted of different items each year. Cronbach’s alpha for each test is displayed in Table A3.1 in the Appendix.

Skills tests
The English test was included in the procedure because most of the study material was in English, also in the Dutch-taught program, and the math test was included because statistics courses are a substantial part of the program. The English test consisted of 20 multiple choice items on the meaning of different English texts. The math test consisted of 30 multiple choice items in 2013 and 2014, and 27 multiple choice items in 2015, testing high-school level math knowledge. The applicants did not receive specific material to prepare for these tests, but example items were provided for the math test. The tests consisted of different items each year.

High school performance
High school grades of applicants who completed the highest level of Dutch secondary education (Dutch: vwo) were collected through the university administration. The grades were not part of the admission procedure. The mean high school grade (HSGPA) was the mean of the final grades in all high school courses, except courses that only resulted in a pass/fail grade. For most courses, 50% of the final course grade was based on a national final exam. The other 50% consisted of the grades obtained in the last three years of secondary education.

Academic achievement
Outcomes on academic achievement were collected through the university administration. For all cohorts, the grade on the first course, the mean grade obtained in the first year (FYGPA, representing academic performance), the number of credits obtained in the first year (representing study progress), and records of dropout in the first year (representing retention) were obtained. For the 2013 cohort we also collected the mean grade obtained after three years (TYGPA, representing academic performance) and bachelor degree attainment after three years (representing study progress). The bachelor program can be completed in three years. All grades were on a scale of 1 to 10, with 10 being the highest grade and a six or higher representing a pass. Mean grades were computed for each student using the highest obtained grade for each course (including resits).
Courses only resulting in a pass/fail decision were not taken into account. Credit was granted after a course was passed; most courses earned five credit points, with a maximum of 60 credits per year. The first and second year courses were mostly the same for all students; the third year consisted largely of elective courses. Since the skills tests were designed to predict performance in particular courses, we also computed a composite for statistics courses (SGPA) and theoretical courses (all courses that required studying literature and completing an exam about psychological theories; TGPA) in the first year. The SGPA is the mean of the final grades on all statistics courses and the TGPA is the mean final grades on courses about psychological theory. In addition, we also obtained information on whether students chose to enroll after participating in the admission procedure. This study was approved by - and in accordance with the rules of the Ethical Committee Psychology from the university.

### 3.2.3 Applicants

The 2013 cohort consisted of 851 applicants, of whom 652 (77%) enrolled in the program, and 638 participated in at least one course. For enrollees the mean age was 20 (SD = 2.0), 69% was female, 46% were Dutch, 42% were German, 9% had another European nationality, 3% had a non-European nationality, and 57% followed the program in English. A high school GPA obtained at the highest level of Dutch secondary education was available for 201 enrollees. Third year academic performance was available for 492 students, the others dropped out of the program in the first or second year. A high school GPA was available for 159 of these students.

The 2014 cohort consisted of 823 applicants, of whom 650 enrolled in the program (79%) and 635 participated in at least one course. For the enrollees the mean age was 20 (SD = 2.1), 66% was female, 44% were Dutch, 46% were German, 7% had another European nationality, 3% had a non-European nationality, and 59% followed the program in English. The high school GPA obtained at the highest level of Dutch secondary education was available for 217 enrollees.

The 2015 cohort consisted of 654 applicants, of whom 541 (83%) enrolled in the program, and 531 participated in at least one course. For enrollees the mean age was 20 (SD = 2.0), 70% was female, 43% were Dutch, 46% were German, 9% had another European nationality, 2% had a non-European nationality, and 62% followed the program in English. The high school GPA obtained at the highest level of Dutch secondary education was available for 188 enrollees.

4 This is the same sample that was used in chapter 2.
Chapter 3

3.2.4 Analyses
To assess the predictive validity of the curriculum-sampling tests, correlations were computed for each cohort between the different predictor scores and FYGPA, obtained credits in the first year, dropout, TYGPA, and degree attainment after three years. Because the literature-based curriculum-sampling tests was designed to mimic the first course, and because the first course was previously found to be a good predictor of subsequent first year performance (Busato et al., 2000; Niessen, et al., 2016), correlations between the curriculum-sampling test scores and the first course grade and correlations between the first course grade and subsequent first year performance were also computed. For these analyses, results from the first course were excluded from the FYGPA and the number of obtained credits. In addition, to compare the predictive validity of the curriculum-sampling tests to the predictive validity of HSGPA, correlations between HSGPA and the academic performance outcomes were also computed.

Although applicants were not rejected by the admission committee, indirect range restriction occurred due to self-selection through self-chosen non-enrollment for first year results, and through dropout in earlier years of the program for third year results, which may result in underestimation of operational validities. Therefore, the correlations (r) were corrected for indirect range restriction (IRR) using the Case IV method (Hunter, Schmidt, & Le, 2006) resulting in an estimate of the true score correlation (ρ), corrected for unreliability and IRR, and the operational validity (rc), only corrected for IRR. The correlations were aggregated across cohorts when applicable (resulting in ℛ, ⌢ρ, and ℛc) 5. In the discussion of the results we focus on the operational validities (Hunter & Schmidt, 2004).

In addition, the incremental validity of the literature-based curriculum-sampling test over HSGPA was studied based on the observed and corrected aggregated correlations, including an IRR correction on the correlation between HSGPA and the curriculum-sampling test scores. Furthermore, the skills tests (math and

---

5 Individual correlations were corrected for indirect range restriction based on the case IV method (Hunter et al., 2006) using the selection package in R (Fife, 2016); we first corrected correlations for unreliability in the predictor and second for indirect range restriction (ρ). To obtain the operational validities (rc), the predictor unreliability was reintroduced after correcting for IRR, as recommended by Hunter et al. (2006). Because the number of cohorts was small, the validity estimates were aggregated (ℛ, ⌢ρ, and ℛc) using a fixed effects model, using the metafor package in R (Viechtbauer, 2010). It was not possible to obtain the operational validities for HSGPA and the first course grade for first year results, since only data of enrolled students were available for these variables. For third year results, the correlations for the first course grade and high school GPA could only be corrected for range restriction due to dropout in earlier years. Since high school GPA was obtained from the administration, the reliability was set to 1 (e.g., Richardson et al., 2012). The reliability of the first course grade was only known for the 2013 sample (α = .74) and was assumed constant across cohorts. Statistical significance (α < .05) was determined before corrections were applied.
English reading comprehension) were included in the admission procedure to predict performance in first year statistics courses and theoretical courses, respectively. We studied the predictive validity for these courses by computing correlations between the scores on the admission tests and the mean grade on these two types of courses in the first year, and the incremental validity of the skills tests over the literature-based curriculum-sampling test. The IRR correction and aggregation procedures described above were also applied to these analyses.

Lastly, the relationship between the admission test scores and enrollment was studied by computing point-biserial correlations between the scores on the admission tests and enrollment. No corrections were needed since test scores and enrollment decisions were available for all applicants. However, low ranking applicants were contacted by phone and were encouraged to reconsider their application. To further investigate the relationships between enrollment and the scores on the admission tests, logistic regression analyses controlling for receiving a phone call were conducted in each cohort, with the admission test scores as independent variables and enrollment as the dependent variable. To ease interpretation of the odds ratios, the admission test scores were standardized first.

## 3.3 Results

### 3.3.1 Short-term Predictive Validity

Tables A3.1 and A3.2 in the Appendix contain all descriptive statistics, and Table A3.3 shows the observed correlations between the predictors and the first year academic outcomes in each cohort. Table 3.1 shows the aggregated observed, true score, and operational validities of each predictor measure for the first year academic performance outcomes.

**Curriculum samples as predictors**

The validity of the literature-based curriculum-sampling test was consistent across cohorts and the aggregated operational validity was high for first year academic performance in terms of GPA ($\bar{r}_c = .53$) and moderate for obtained credits ($\bar{r}_c = .42$) and for dropout in the first year ($\bar{r}_c = -.32$). The video-lecture test (only administered in 2015) showed moderate predictive validity for FYGPA ($r_c = .34$) and obtained credits ($r_c = .29$), and a small negative correlation with dropout ($r_c = -.15$). In the entire applicant sample, the correlation between the scores on both curriculum-sampling tests equaled $r = .51$. In addition, the video-lecture test showed very small incremental validity for predicting FYGPA over the literature-based curriculum-sampling test ($\Delta R^2 = .01$, $R^2 = .20$, $\Delta F(1, 528) = 5.69$, $p = .02$, and based on the corrected correlations, $\Delta R^2_c = .01$, $R^2_c = .26$).
Table 3.1

Correlations between predictors and first year academic outcomes aggregated across cohorts

<table>
<thead>
<tr>
<th>Predictor</th>
<th>FYGPA</th>
<th>FYECT</th>
<th>FY dropout&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Enrollment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ρ</td>
<td>β</td>
<td>ρ&lt;sub&gt;c&lt;/sub&gt;</td>
<td>ρ</td>
</tr>
<tr>
<td>Cur. 1</td>
<td>.46</td>
<td>.59</td>
<td>.53</td>
<td>.36</td>
</tr>
<tr>
<td>Cur. 2&lt;sup&gt;b&lt;/sup&gt;</td>
<td>.29</td>
<td>.40</td>
<td>.34</td>
<td>.25</td>
</tr>
<tr>
<td>Math</td>
<td>.25</td>
<td>.30</td>
<td>.26</td>
<td>.18</td>
</tr>
<tr>
<td>English</td>
<td>.17</td>
<td>.25</td>
<td>.19</td>
<td>.12</td>
</tr>
<tr>
<td>HSGPA</td>
<td>.47</td>
<td>.28</td>
<td>-.20</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[.41,.53]</td>
<td>[.20,.35]</td>
<td>[-.27,-.12]</td>
<td></td>
</tr>
<tr>
<td>FCG&lt;sup&gt;c&lt;/sup&gt;</td>
<td>.72</td>
<td>.61</td>
<td>-.43</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[.69,.74]</td>
<td>[.58,.64]</td>
<td>[-.47,-.39]</td>
<td></td>
</tr>
</tbody>
</table>

Note. Cur. 1 = curriculum-sampling test based on literature, Cur. 2 = curriculum-sampling test based on a video lecture, Math = math test, English = English reading comprehension test, HSGPA = high school mean grade, FCG = first course grade, FYGPA = first year mean grade, FYECT = first year credits, FY dropout = first year dropout, Enrollment = first year enrollment, ρ = the aggregated correlation across cohorts, β = the aggregated true score correlation (corrected for unreliability and indirect range restriction), ρ<sub>c</sub> = the aggregated operational correlation across cohorts (corrected for indirect range restriction). <sup>a</sup> Point-biserial correlations. <sup>b</sup> Based the 2013 cohort. <sup>c</sup> For these correlations, results on the first course were not included in the calculation of FYGPA and credits. 95% confidence intervals are in brackets. All correlations were statistically significant with p < .05.
Specific skills tests and grades as predictors

The operational validities of the math and English skills tests were less consistent across cohorts (Table A3.3) and the aggregated operational validities were moderate to small for all outcome measures. The data needed to check for and, if needed, correct for range restriction in HSGPA were not available, so we only computed the aggregated observed correlations with the academic outcomes. High school GPA showed high predictive validity for FYGPA ($r_{c} = .47$), and moderate predictive validity for the obtained number of credits ($r_{c} = .28$) and dropout ($r_{c} = -.20$). The first course grade showed very high predictive validity for subsequent performance in the first year, with $r_{c} = .72$ for FYGPA, $r_{c} = .61$ for obtained credits, and $r_{c} = -.43$ for dropout. These correlations were substantially higher than those for the literature-based curriculum-sampling test, which was modeled after this first course. The literature-based curriculum-sampling test showed an aggregated correlation with the first course grade of $r_{c} = .49$ ($r_{c} = .56$ after correction for IRR).

3.3.2 Long-term Predictive Validity

Table 3.2 shows observed and corrected correlations between the predictors and academic performance after three years (only studied for the 2013 cohort). The literature-based curriculum-sampling test showed a high operational validity of $r_{c} = .55$ for third year GPA, and a moderate operational validity of $r_{c} = .32$ with bachelor’s degree attainment in three years. The math skills test showed small validities ($r_{c} = .27$ for TYGPA and $r_{c} = .19$ for TYBA), and the English test showed small validity for TYGPA of $r_{c} = .17$, and small, non-significant validity for TYBA ($r_{c} = .06$). High school GPA had a high correlation with TYGPA of $r_{c} = .63$ and a moderate correlation with TYBA of $r_{c} = .31$. Lastly, the first course grade on Introduction to Psychology obtained in the first year showed large correlations with TYGPA ($r_{c} = .71$) and with TYBA ($r_{c} = .48$). Thus, the curriculum-sampling test scores, high school GPA, and the first course grade were good predictors of academic performance after three years of studying in the Psychology program.

3.3.3 Incremental Validity

The incremental validity of the literature-based curriculum-sampling tests over high school GPA was computed based on the aggregated correlations across cohorts, both observed and corrected for indirect range restriction in the curriculum-sampling test. When the analyses were conducted for each cohort separately (results not shown but available upon request), the incremental validity of the literature-based curriculum-sampling test over HSGPA was statistically significant in each cohort and for each criterion. The aggregated results are shown in Table 3.3. The aggregated correlation between the curriculum-sampling test
score and high school GPA was $\bar{r} = .49$ and $r_c = .55$. The curriculum-sampling test showed a substantial increase in explained variance over high school GPA for predicting FYGPA ($\Delta R^2_c = .11$), and for predicting the number of obtained credits ($\Delta R^2_c = .10$). Together, high school GPA and the curriculum-sampling test scores explained a large percentage of the variance in FYGPA ($R^2_c = .33$) and in obtained credits in the first year ($R^2_c = .18$). For predicting TYGPA, the curriculum-sampling test and high school GPA combined explained 46% of the variance, but the incremental validity of the curriculum-sampling test over high school GPA was modest ($\Delta R^2_c = .06$).

Table 3.2

<table>
<thead>
<tr>
<th>Predictor</th>
<th>FYGPA</th>
<th>FYECT</th>
<th>TYGPA*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$r$</td>
<td>$\rho$</td>
<td>$r_c$</td>
</tr>
<tr>
<td>Cur. 1</td>
<td>.38*</td>
<td>.62</td>
<td>.55</td>
</tr>
<tr>
<td>Math</td>
<td>.24*</td>
<td>.31</td>
<td>.27</td>
</tr>
<tr>
<td>English</td>
<td>.17*</td>
<td>.20</td>
<td>.17</td>
</tr>
<tr>
<td>HSGPA</td>
<td>.61*</td>
<td>.63</td>
<td>.30*</td>
</tr>
<tr>
<td>FCG</td>
<td>.62*</td>
<td>.83</td>
<td>.71</td>
</tr>
</tbody>
</table>

Note. Cur. 1 = curriculum-sampling test based on literature, Math = math test, English = English reading comprehension test, HSGPA = high school mean grade, FCG = first course grade, FYGPA = third year mean grade, TYGPA = third year Bachelor's degree attainment, $r = $ uncorrected correlation, $\rho = $ the aggregated true score correlation (corrected for unreliability and indirect range restriction), $r_c = $ correlation corrected for indirect range restriction. $^*$Point-biserial correlations. 95% confidence intervals are in brackets. * $p < .05$

Table 3.3

Incremental validity of the literature-based curriculum-sampling test over high school GPA

<table>
<thead>
<tr>
<th>Data</th>
<th>FYGPA</th>
<th>FYECT</th>
<th>TYGPA*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\bar{R}$</td>
<td>$\bar{R}^2$</td>
<td>$\Delta R^2$</td>
</tr>
<tr>
<td>Observed</td>
<td>.54</td>
<td>.29</td>
<td>.07</td>
</tr>
<tr>
<td>Corrected</td>
<td>.57</td>
<td>.33</td>
<td>.11</td>
</tr>
</tbody>
</table>

Note. FYGPA = first year mean grade, FYECT = first year credits, TYGPA = third year mean grade, $\bar{R} = $ aggregated multiple correlation, $R^2 = $ aggregated variance explained based on HSGPA and the curriculum-sampling test scores, $\Delta R^2 = $ aggregated increase in explained variance based on curriculum-sampling test scores over HSGPA. Observed = based on observed correlations, Corrected = based on operational correlations (corrected for IRR). $^*$Based on the 2013 cohort.
3.3.4 Specific Course Achievement
The predictive validity of the admission tests for specific course achievement in the first year is shown in Table 3.4. The aggregated correlation between the mean grade on the statistics courses (SGPA) and the mean grade on theoretical courses (TGPA) was \( \bar{r} = .66 \). As expected, the scores on the English test only showed a small correlation with performance in the statistics courses (\( \bar{r}_c = .14 \)). The curriculum-sampling tests and the math test predicted performance in statistics course equally well (around \( \bar{r}_c = .40 \)). Their scores combined accounted for a large percentage of the variance in statistics GPA (\( \bar{R}_c = .26 \)), with an incremental validity of the math test over the curriculum-sampling test of \( \Delta R^2_c = .08 \). The curriculum-sampling tests were the strongest predictors of performance in theoretical courses, with \( \bar{r}_c = .53 \) for the literature-based test and \( r_c = .30 \) for the video-lecture test. The English test and the math test showed small correlations (both \( \bar{r}_c = .19 \)) with theoretical course performance. The scores on the English test and the literature-based curriculum-sampling test combined accounted for a large percentage of the variance in theoretical GPA (\( \bar{R}_c = .28 \)) and with an incremental validity of the English test over the curriculum-sampling test of \( \Delta R^2_c = < .01 \).

3.3.5 Enrollment
The aggregated operational and true score correlations between the admission test scores and enrollment are shown in Table 3.1, and the observed correlations per cohort are in Table A3.3 in the Appendix. The aggregated operational correlation between the scores on the literature-based curriculum-sampling tests and enrollment was \( \bar{r} = .25 \); for the video-lecture test this was \( r = .18 \). The math test scores (\( \bar{r} = .09 \)) and the English test scores (\( \bar{r} = .15 \)) showed small correlations with enrollment. So, the admission test scores showed small relationships with enrollment, with the largest correlation for the literature-based curriculum-sampling test.

To assess the relationship between the scores on the admission tests and enrollment further, enrollment was regressed on the admission test scores, controlling for receiving a discouraging phone call. To ease interpretation of the

---

6 The incremental validity of the math test over the curriculum-sampling test for SGPA based on the observed correlations equaled \( \Delta R^2 = .08 \) (\( R^2 = .21 \)). The aggregated correlation between the math test scores and the curriculum-sampling test scores was \( \bar{r}_c = .27 \). The incremental validity of the English test over the curriculum-sampling test for TGPA based on the observed correlations equaled \( \Delta R^2 = < .01 \) (\( R^2 = .21 \)). The aggregated correlation between scores on the English test and scores on the curriculum-sampling test was \( \bar{r}_c = .44 \). When the analyses were conducted for each cohort separately (results not shown but available upon request), the incremental validity of the math test over the literature-based curriculum-sampling test for statistics GPA was statistically significant in each cohort. The incremental validity of the English test over the literature-based test was only statically significant in the 2013 cohort.
odds ratios, the admission test scores were standardized first. The results of the logistic regression are shown in Table 3.5. In each cohort, the score on the literature-based curriculum-sampling test was significantly related to enrollment, with higher scores associated to higher odds for enrollment. With each standard deviation unit increase in the test score, the odds of enrollment increased with \( e^B = 1.33 \), \( e^B = 2.04 \), and \( e^B = 1.55 \) in 2013, 2014, and 2015, respectively. In the 2013 and 2014 cohorts, the curriculum-sampling test score was the only significant predictor of enrollment. The video-based curriculum-sampling test score and the math test score also showed statistically significant relationships with enrollment in the 2015 cohort (\( e^B = 1.34 \) and \( e^B = 0.78 \)). Notably, a higher score in the math test was associated with lower odds of enrollment.

Table 3.4

Predictive validity for specific course achievement in the first year

<table>
<thead>
<tr>
<th>Predictor</th>
<th>SGPA</th>
<th>TGPA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \bar{r} )</td>
<td>( \bar{\rho} )</td>
</tr>
<tr>
<td>Cur. 1</td>
<td>.36</td>
<td>.47</td>
</tr>
<tr>
<td></td>
<td>[.32, .40]</td>
<td>[.42, .52]</td>
</tr>
<tr>
<td>Cur. 2*</td>
<td>.33</td>
<td>.45</td>
</tr>
<tr>
<td>Math</td>
<td>.37</td>
<td>.44</td>
</tr>
<tr>
<td>English</td>
<td>.13</td>
<td>.17</td>
</tr>
<tr>
<td></td>
<td>[.07, .18]</td>
<td>[.09, .25]</td>
</tr>
</tbody>
</table>

Note. Cur. 1 = curriculum-sampling test based on literature, Cur. 2 = curriculum-sampling test based on a video lecture, Math = math test, English = English reading comprehension test, SGPA = statistics courses GPA, TGPA = theoretical courses GPA, \( \bar{r} \) = the aggregated correlation across cohorts, \( \bar{\rho} \) = the aggregated true score correlation (corrected for unreliability and indirect range restriction), \( \bar{r}_c \) = the aggregated operational correlation across cohorts (corrected for indirect range restriction). 95% confidence intervals are in brackets. * Based on the 2015 cohort. All correlations were statistically significant with \( p < .05 \).
The admission test scores were standardized first. The results of the logistic regression are shown in Table 3.5. In each cohort, the score on the literature-based curriculum-sampling test was significantly related to enrollment, with higher scores associated with higher odds for enrollment. With each standard deviation unit increase in the test score, the odds of enrollment increased with $e^B = 1.33$, $e^B = 2.04$, and $e^B = 1.55$ in 2013, 2014, and 2015, respectively. In the 2013 and 2014 cohorts, the curriculum-sampling test score was the only significant predictor of enrollment. The video-based curriculum-sampling test score and the math test score also showed statistically significant relationships with enrollment in the 2015 cohort ($e^B = 1.34$ and $e^B = 0.78$). Notably, a higher score in the math test was associated with lower odds of enrollment.

<table>
<thead>
<tr>
<th>Variable</th>
<th>2013a $e^B$ 95% CI</th>
<th>Wald</th>
<th>2014 $e^B$ 95% CI</th>
<th>Wald</th>
<th>2015 $e^B$ 95% CI</th>
<th>Wald</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phone call</td>
<td>1.70 0.97, 2.99</td>
<td>3.38</td>
<td>1.09 0.59, 2.02</td>
<td>0.79</td>
<td>1.09 0.49, 2.42</td>
<td>0.04</td>
</tr>
<tr>
<td>Cur. 1</td>
<td>1.33 1.09, 1.63</td>
<td>7.67*</td>
<td>2.04 1.64, 2.53</td>
<td>40.83*</td>
<td>1.55 1.15, 2.07</td>
<td>8.42*</td>
</tr>
<tr>
<td>Cur. 2</td>
<td>1.34 1.05, 1.71</td>
<td>5.51*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Math</td>
<td>1.14 0.93, 1.40</td>
<td>1.67</td>
<td>1.04 0.84, 1.30</td>
<td>0.13</td>
<td>0.78 0.61, 0.98</td>
<td>4.43*</td>
</tr>
<tr>
<td>English</td>
<td>1.03 0.85, 1.24</td>
<td>0.06</td>
<td>1.11 0.91, 1.36</td>
<td>1.15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model $X^2$</td>
<td>46.44*</td>
<td></td>
<td>81.28*</td>
<td></td>
<td>39.10*</td>
<td></td>
</tr>
</tbody>
</table>

Note. Cur. 1 = curriculum-sampling test based on literature, Cur. 2 = curriculum-sampling test based on a video lecture, Math = math test, English = English reading comprehension test. $e^B =$ odds ratio. * These results (using unstandardized data) were also shown in chapter 2. * $p < .05$. 

Table 3.5 Logistic regression results for predicting enrollment based on admission test scores.
3.4 Discussion

The results showed that the predictive validity of the literature-based curriculum-sampling test for first year academic achievement was consistent across cohorts, with high correlations with FYGPA and moderate correlations with the number of obtained credits and dropout. These results replicated the results shown in chapter 2. However, the validity of the curriculum-sampling test was lower than the predictive validity of the grade in the course that it was designed to mimic; this course grade showed very large correlations with later academic achievement. In addition, the video-lecture based curriculum-sampling test showed moderate predictive validity and little incremental validity over the literature-based test. Notably, the operational predictive validity of the literature-based curriculum-sampling test for academic performance after three years of bachelor courses was still high. The predictive validity of the curriculum-sampling test was mostly comparable to or slightly higher than the predictive validity of high school GPA; in comparison, high school GPA was a slightly better predictor for third year GPA. Furthermore, whereas the literature-based curriculum-sampling test scores and high school GPA were strongly related, the unique explanatory power of the curriculum test scores over high school GPA was substantial for first year academic achievement and modest for third year academic achievement.

For the prediction of specific course achievement, our expectations were partially confirmed. The math test scores did not predict statistical course performance better than the curriculum-sampling tests, but they did show some incremental validity over the curriculum-sampling test. The English test scores did not add incremental validity over the curriculum-sampling test for predicting theoretical course performance. A possible explanation may be that English reading comprehension was also implicitly assessed by the curriculum-sampling test, as the material to be studied was in English. Finally, in the regression analyses, relationships between test scores and enrollment decisions were only found for the curriculum-sampling test scores, indicating that this relationship may be an advantage of representative sample-based admission tests in particular.

Study 2: Construct Saturation

3.5 Method

3.5.1 Participants and Procedure

The sample consisted of 104 first year students who enrolled in the Dutch-taught psychology program in 2015 and participated in the study for course credit. The mean age was 19.1 (SD = 1.5), 71% was female, and all students had a Dutch nationality. The data were collected in two sessions scheduled three days apart. In the first session, participants completed self-report questionnaires about their
personality, procrastination tendencies, study skills, and study habits. In the second session they completed a cognitive ability test that took about an hour to complete. All measures were administered on a computer in the Dutch language. Both sessions were proctored and took place in a research lab at the university, with several participants completing the measures simultaneously. The participants received their scores on the cognitive test immediately after completion to encourage effortful participation. All participants provided informed consent and provided permission to link the data to their admission test scores and academic performance. Participation took place between October 2015 and April 2016. This study adhered to the rules of and was approved by the Ethics Committee Psychology.

3.5.2 Materials and Measures

Cognitive ability
Cognitive ability was measured by the Q1000 Capaciteiten Hoog (in Dutch; van Bebber, Lem, & van Zoelen, 2010), developed for personnel selection and career development purposes for adults who followed upper higher education. The test consisted of seven subscales across three domains: Analogies, syllogisms, and vocabulary for the verbal domain; digit series and sums for the numerical domain; and matrices and cubes for the figural domain. The test also yields a higher-order general cognitive ability score consisting of a combination of the three domain scores. Van Bebber et al. (2010) showed that the test has convergent validity with other measures of cognitive ability (the Dutch version of the General Aptitude Test Battery and Raven’s Progressive Matrices), and the Dutch Committee on Tests and Testing Affairs (Dutch: COTAN) evaluated the reliability and construct validity of the test as sufficient (Egberink, Holly-Middelkamp, & Vermeulen, 2010). The proportion of items answered correctly per scale was used to compute the scale scores, which were averaged across scales to obtain the total score representing general cognitive ability. These scores were converted to a scale of 0 to 10 for convenience. The estimated reliability of the full test score was $\alpha = .87$, which was similar to the reliability reported in the test manual (van Bebber et al., 2010).

Conscientiousness
Conscientiousness was measured using the Dutch version of the Big Five Inventory (BFI; Denissen, Geenen, van Aken, Gosling, & Potter, 2008). The entire BFI was administered, but the other scale scores were not used in this study. The conscientiousness scale of the BFI consisted of nine items answered on a five-point Likert scale (1 = strongly disagree through 5 = strongly agree), and Cronbach's $\alpha = .86$. 

57
**Procrastination**

Lay’s Procrastination Scale (Lay, 1986) was used to measure procrastination tendencies. This scale consisted of 20 items answered on a five-point Likert-scale (1 = never through 5 = all of the time) with Cronbach’s α = .84.

**Study skills and study habits**

The Study Management and Academic Results Test (SMART: Kleijn, van der Ploeg, & Topman, 1994) consisted of four scales measuring study-related cognitions (academic competence and test competence), and study management skills (time management and strategic studying). Academic competence was defined as being able to understand the study material and enjoying studying the material; test competence was defined as being able to separate essentials from details, managing the amount of study material, coping with tensions, and preparing for examinations; time management was defined as the ability to plan study time and combine studying with leisure activities; and strategic studying was defined as utilizing strategies such as summarizing and testing one’s own knowledge and deliberating how to approach the study material. The SMART contained 29 items, with four to six items per subscale. All items were answered on a four-point Likert scale (1 = almost never through 4 = nearly always). The scales yielded estimated reliabilities of $\alpha_{\text{academic competence}} = .69$, $\alpha_{\text{test competence}} = .76$, $\alpha_{\text{time management}} = .82$, and $\alpha_{\text{strategic studying}} = .66$.

### 3.5.3 Analyses

First, we computed observed correlations and correlations corrected for unreliability between the curriculum-sampling test scores and scores on the cognitive and noncognitive measures. Corrections for range restriction could not be applied because the necessary information was not available for the explanatory variables. We expected that the curriculum-sampling tests were saturated with cognitive- and noncognitive constructs, because these constructs also predict academic performance. To check our assumptions about the relationships between the explanatory variables and academic performance, the same analyses were conducted with FYGPA as the dependent variable. Second, we conducted multiple regression analyses with the cognitive and noncognitive predictors described above as independent variables and each of the curriculum-sampling test scores and FYGPA as dependent variables. The `SetCor` function in the psych R package (Revelle, 2015) was used to conduct regression analyses based on the correlation matrices corrected for unreliability.
3.6 Results

Descriptive statistics for all variables are in the Appendix in Table A3.4. Table 3.6 shows the correlations between the variables and the multiple regression results after correcting the correlations for unreliability. We will focus on the results after correcting for unreliability, because we were interested in the theoretical relationships between the variables. The results based on the uncorrected correlations are shown in Table A3.5 in the Appendix. The variance inflation factors (VIF) were all smaller than 4, (the largest VIF before correction was 2.04, and 3.66 after correction, both for the time management variable). Although this value indicates that there were no serious problems with multicollinearity (e.g., Miles & Shevlin, 2001) and we adopted a conservative rule of thumb (10 is another often-used threshold, e.g., Field, 2013), some of the results do suggest multicollinearity problems, as we discuss below. This makes some of the results difficult to interpret.

In the model with the literature-based curriculum-sampling test as the dependent variable, cognitive ability was not a statistically significant predictor, and the zero-order correlation was also small and not statistically significant. Only conscientiousness, test competence, and time management were significant predictors of the literature-based curriculum-sampling test score in the regression model. However, the independent variables did explain a substantial proportion of the variance in the curriculum-sampling tests scores ($\Delta R^2_c = .38$), and the regression coefficient for the time management score was negative, which contradicts theoretical expectations and the zero-order correlation. These somewhat contradictory results are signs of multicollinearity problems (O’Brien, 2007). The model with the video-based curriculum-sampling test scores as the dependent variable showed similar results.

For the model with FYGPA as the dependent variable there was also no statistically significant relationship with the cognitive ability scores. The zero-order correlations for the noncognitive independent variables were all statistically significant and in the expected direction, with conscientiousness, procrastination, and time management as significant predictors in the regression model. However, procrastination showed a positive regression coefficient in the model, which also does not coincide with theoretical exceptions and the zero-order correlations. So, the zero-order correlations confirmed our theoretical expectations, but the regression results, again, indicated problems with multicollinearity, which makes them difficult to interpret.
Table 3.6

Construct saturation multiple regression results based on correlations corrected for unreliability

<table>
<thead>
<tr>
<th>Variables</th>
<th>Cur. 1</th>
<th>Cur. 2</th>
<th>FYGPA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$r_c$</td>
<td>$\beta/R^2$</td>
<td>$r_c$</td>
</tr>
<tr>
<td>Cognitive ability</td>
<td>.11</td>
<td>.15</td>
<td>.10</td>
</tr>
<tr>
<td></td>
<td>[-.08, .30]</td>
<td>(.09)</td>
<td>[-.09, .29]</td>
</tr>
<tr>
<td>Conscientiousness</td>
<td>.09</td>
<td>.32*</td>
<td>.18</td>
</tr>
<tr>
<td></td>
<td>[-.10, .27]</td>
<td>(.14)</td>
<td>[-.01, .36]</td>
</tr>
<tr>
<td>Procrastination</td>
<td>-.01</td>
<td>.07</td>
<td>.03</td>
</tr>
<tr>
<td></td>
<td>[-.20, .18]</td>
<td>(.15)</td>
<td>[-.16, .22]</td>
</tr>
<tr>
<td>Academic Competence</td>
<td>.24*</td>
<td>-.28</td>
<td>.15</td>
</tr>
<tr>
<td></td>
<td>[.05, .41]</td>
<td>(.15)</td>
<td>[-.04, .33]</td>
</tr>
<tr>
<td>Test Competence</td>
<td>.49*</td>
<td>.92*</td>
<td>.21*</td>
</tr>
<tr>
<td></td>
<td>[.33, .62]</td>
<td>(.14)</td>
<td>[.02, .39]</td>
</tr>
<tr>
<td>Time Management</td>
<td>.12</td>
<td>-.41*</td>
<td>-.02</td>
</tr>
<tr>
<td></td>
<td>[.07, .31]</td>
<td>(.15)</td>
<td>[-.21, .17]</td>
</tr>
<tr>
<td>Strategic Studying</td>
<td>.10</td>
<td>.07</td>
<td>.08</td>
</tr>
<tr>
<td></td>
<td>[.09, .29]</td>
<td>(.10)</td>
<td>[.11, .27]</td>
</tr>
<tr>
<td>Model $R^2$ (adj. $R^2$)</td>
<td>.38*</td>
<td>.23*</td>
<td>.29*</td>
</tr>
<tr>
<td></td>
<td>(.33)</td>
<td></td>
<td>(.17)</td>
</tr>
</tbody>
</table>

Note. Cur. 1 = curriculum-sampling test based on literature, Cur. 2 = curriculum-sampling test based on a video lecture, FYGPA = first year mean grade. 95% CI’s are between brackets. * $p < .05$

3.7 Discussion

We hypothesized that, similar to academic achievement, the curriculum-sampling test scores were multifaceted compound measures, saturated with cognitive ability and several noncognitive constructs (Callinan & Robertson, 2000; Thornton & Kedharnath, 2013; Niessen & Meijer, 2017). Contrary to our expectations, the scores on the curriculum-sampling tests and first year GPA were not significantly related to cognitive ability as measured by the Q1000 test. One explanation may be restriction in range in the cognitive scores: The students in our sample were a selective group compared to the general population (Furnham, Chamorro-Premuzic, & McDougall, 2003). Another explanation may be that the participants did not put maximum effort in completing the cognitive test because the test was administered low-stakes, which could have impacted the validity of the test scores (e.g., Barry & Finney, 2016).

We did not find consistent results for the relationships between most noncognitive scale scores and the curriculum-sampling test scores. The unexpected switch of
signs for some regression coefficients, and the few statistically significant predictors in the models paired with the relatively high explained variance are indicative of problems with multicollinearity (O’Brien, 2007), and make it difficult to draw conclusions based on the results. However, the results did indicate a relationship between test competence and the scores on both curriculum-sampling tests. So, we cautiously conclude that there does seem to be some noncognitive saturation in the curriculum-sampling tests, especially of test competence, but the results on the explanatory power of the rest of the noncognitive variables were inconclusive.

### 3.8 General Discussion

The central question in this study was: Do curriculum-sampling test scores predict later academic achievement, and what do they measure in terms of cognitive and noncognitive constructs? The results of the first study showed that the literature-based curriculum-sampling test was a good predictor for short-term and long-term academic achievement. These results were consistent across cohorts and provided an important extension of earlier findings (Niessen et al., 2016).

An interesting and quite remarkable result was that the literature-based curriculum sample test, which is a simple 40-item test, yielded equally good or somewhat better predictions of first year academic achievement than high school GPA (they predicted FYGPA about equally well, but the curriculum-sampling tests was a slightly better predictor for progress and retention in the first year). High school GPA is commonly considered to be the best predictor of academic performance in higher education (Peers & Johnston, 1994; Westrick et al., 2015; Zwick, 2013) and is a very rich summary measure, containing grades on thoroughly developed national final exams and on secondary school exam performance aggregated over the last three years of high school. Furthermore, the literature-based curriculum-sampling test showed substantial incremental validity over high-school GPA, and a combination of the two explained 33% of the variance in first year GPA and 46% of the variance in third year GPA. Furthermore, the relationship between the curriculum-sampling test scores and enrollment decisions indicated that curriculum sampling may encourage self-selection and thus may help applicants to gain insight into their fit to the academic program, which can be a considerable advantage.

A video-lecture test may have the advantage that the study material is presented in a vivid and attractive format. However, this test explained very little unique variance in addition to the literature-based tests and showed lower predictive validity. In contrast, assessing math skills to predict performance in statistics
courses added to the prediction of performance in those courses over the curriculum-sampling test scores. This confirms that assessing specific skills that are needed for distinct components of the curriculum can add to the prediction of performance in those specific components. These findings agree with the content-matching approach of predictors and outcomes (Sackett et al., 2016) and this result may be further explored in the context of assessing nonacademic skills such as clinical skills in medicine (e.g., Lievens, 2013) or practical skills in vocational education. Note that in these cases, specific skills can also be measured using a samples approach. In future research it would be interesting to investigate if a samples approach to measuring specific skills would increase the predictive validity.

The aim of the second study was to investigate the cognitive and noncognitive construct saturation of the curriculum-sampling tests, in order to explore possible mechanisms underlying their predictive validity. The results did not completely coincide with our expectations. We did not find relationships between the cognitive ability test scores and the two curriculum-sampling tests scores or first year GPA. On the one hand, this was a surprising finding given the large body of literature that shows a relationship between cognitive abilities and academic performance (Kuncel & Hezlett, 2007; 2010), and given the rather obvious cognitive nature of the curriculum-sampling tests. On the other hand, our results are in line with findings based on restricted samples (e.g., Moneta-Koehler, Brown, Petrie, Evans, & Chalkley, 2017), and with the observation that correlations between cognitive ability and educational achievement tend to decline in every subsequent educational phase due to ongoing selectivity in the educational careers of students (Chamorro-Premuzic & Furnham, 2005; Furnham et al., 2003). In the Netherlands, students are selected for different levels of secondary education around age 12, largely based on a strongly cognitively-loaded test (Bartels, Rietveld, van Baal and Bouma, 2002). This degree of early selection restricts the relation between cognitive ability and educational achievement measured in later years (e.g., Crombag, Gaff & Chang, 1975; Resing & Drenth, 2007), although small positive relationships between cognitive ability and academic achievement were found in Dutch student samples in earlier studies (e.g., Busato et al., 2000; Kappe & van der Flier, 2012).

Although our results are based on a relatively small sample and we are, therefore, cautious to draw firm conclusions, our results corresponded to other recent publications that showed that it can be difficult to discriminate between applicants to higher education based on tests of general cognitive skills. For example, Moneta-Koehler et al. (2017) found no relationships between GRE scores and graduating in
a PhD program, passing the qualifying exam, time to defense, publishing first
author papers, or obtaining individual grants, and moderate relationships with
graduate GPA and some elements of faculty evaluations. They found that
undergraduate GPA was a better predictor of graduate GPA, first semester grades,
and graduating with a PhD. Thus, measures that were more similar to the criteria
were better predictors than general measures.

Whereas FYGPA was related to and saturated with the noncognitive variables
according to expectations based on previous studies (Credé & Kuncel, 2008;
Richardson et al., 2012), the relation between curriculum-sampling test scores and
the noncognitive variables was not straightforward. Our findings showed that
there was some noncognitive saturation in curriculum-sampling tests, mostly for
conscientiousness and test competence scores. Thus, this may indicate that the
predictive validity of curriculum-sampling tests may partly be explained by its
saturation with the competence to prepare for examinations, separating essentials
from details, managing the amount of study material, and coping with tensions,
which are important learning skills (Credé & Kuncel, 2008; Richardson, et al.,
2012). Furthermore, the same rationale applies to the high and robust predictive
validity of high school GPA, which can be explained by its multifaceted nature
including both cognitive abilities, personality traits, and study skills (Borghans,
Golsteyn, Heckman, & Humphries, 2016; Deary, Strand, Smith, & Fernandes, 2007;
Dumfart & Neubauer, 2016).

3.8.1 Limitations
A first limitation of this study was that the data in each cohort were collected
within a single academic program at a single university. Hence, the results do not
necessarily generalize to other academic disciplines or to programs focused less on
theoretical knowledge and more on practical skills. Future research may extend
our study to different programs where different skills may be important, such as
communication skills or ethical reasoning (e.g., Pau et al., 2013). A second
limitation was that when calculating the operational validities we did not correct
for unreliability in the outcomes measures, which could lead to an underestimation
of the operational validity coefficients. Second, the case IV method to correct for
indirect range restriction (Hunter et al., 2006) can result in over- or
underestimation of the operational validity, although it yields the most accurate
estimate of operational correlations compared to other correction methods
(Beatty, Barratt, Berry, & Sackett, 2014).

Furthermore, we cannot conclude that the positive relationships between the
curriculum-sampling test scores and enrollment were caused by the experience of
studying for or taking the tests, or by obtaining low scores on the tests. In future studies, a measure of enrollment intentions may be included before the admission tests are administered. In addition, the possibility to promote self-selection through curriculum sampling may also be used in procedures aimed at placement decisions or advising on student-program-fit. In future research it may be investigated whether the predictive validity of curriculum-sampling tests generalizes to lower-stakes procedures.

A possible explanation for the lower noncognitive saturation of the curriculum-sampling tests as compared to first year GPA is that the curriculum samples were not representative enough for first year courses, which may require more prolonged effort. This is supported by the finding that the grade on the first course in the program was a much better predictor of short-term and long-term academic performance than the scores on the curriculum-sampling tests. Whereas the literature-based curriculum-sampling test mimicked that course, the course required studying an entire book, while the curriculum-sampling test only required studying two chapters. A more comprehensive curriculum sample may be more saturated with constructs related to effortful behavior.

3.8.2 Conclusions
From existing research we know that high school GPA is a good predictor of future academic performance in higher education and is arguably the most efficient measure to use for admission decisions. In the present study we showed that the literature-based curriculum-sampling tests mostly showed similar or slightly higher predictive validity than high school GPA for first-year academic outcomes, and similar or slightly lower predictive validity for third year academic outcomes. In addition, this curriculum-sampling test showed incremental validity over high school GPA. One caveat is that the tests should have an acceptable psychometric quality and that extra resources are needed to construct and administer the tests. However, in the format adopted in this study, constructing and administering the tests took relatively little time and effort, and the reliability was sufficient. A final advantage is that curriculum sampling is perceived as a favorable admission method, whereas applicants disliked the use of high school GPA in admission procedures (Niessen, Meijer, & Tendeiro, 2017a). So, in cases where using high school GPA is not feasible, or when self-selection or applicant perceptions are of major interest, curriculum-sampling tests may be preferred over, or may be used in addition to high school GPA.