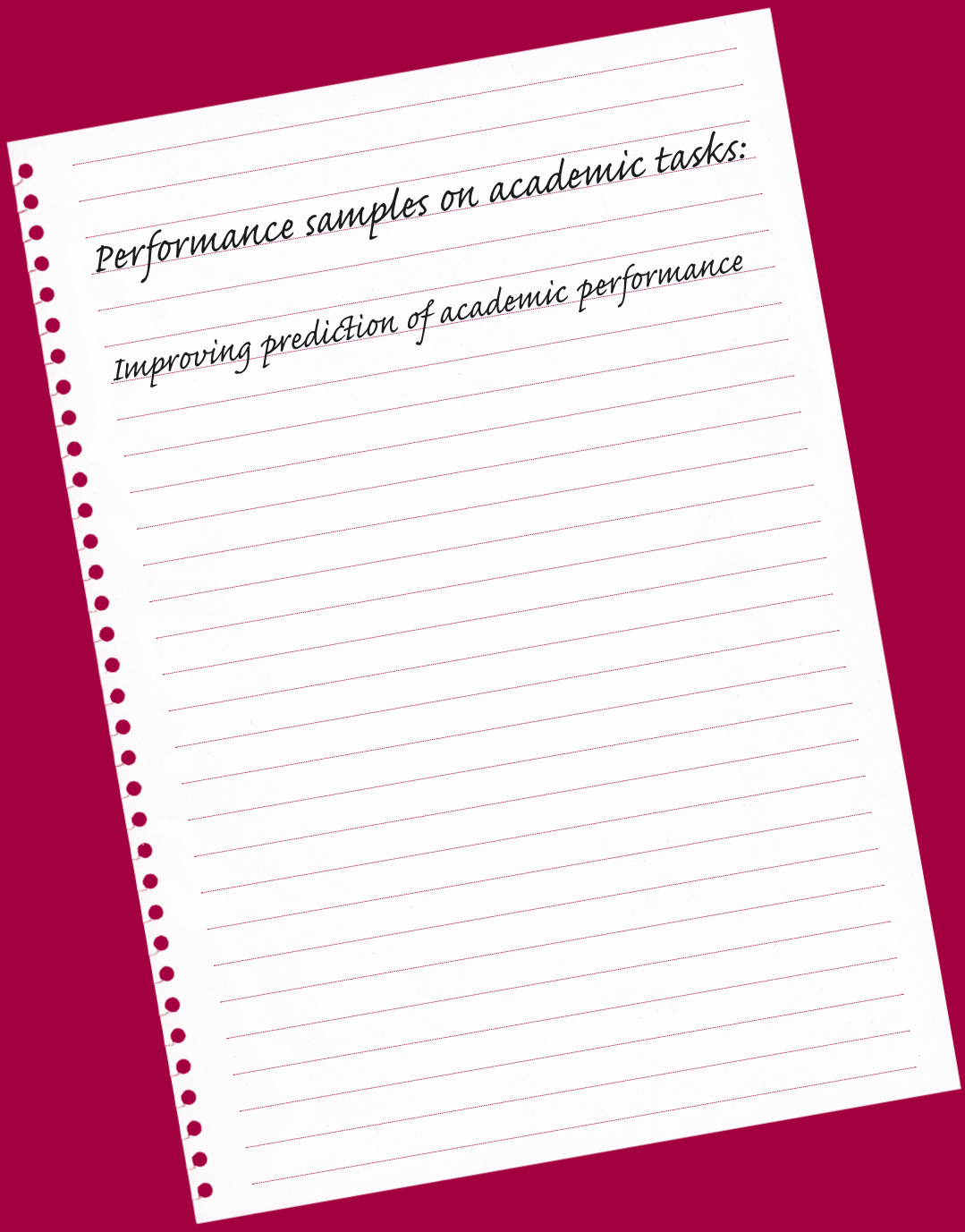
Performance samples on academic tasks : improving prediction of academic performance

Tanilon, J.



Jenny Tanilon

Performance samples on academic tasks: Improving prediction of academic performance

Jenny Tanilon

ISBN/EAN: 978-94-90858-00-1

Printed by Mostert & Van Onderen, Leiden

Performance samples on academic tasks:
Improving prediction of academic performance

PROEFSCHRIFT

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden
op gezag van Rector Magnificus Prof. mr. P.F. van der Heijden
volgens besluit van het College voor Promoties
te verdedigen op 4 oktober 2011
klokke 15.00 uur

door
Jenny Tanilon
geboren te Manilla, Filippijnen
in 1975

Promotiecommissie

Promotoren:

Prof. dr. M.S.R. Segers (Universiteit Maastricht)
Prof. dr. P.H. Vedder (Universiteit Leiden)

Overige leden:

Prof. dr. P.W. van den Broek (Universiteit Leiden)
Prof. dr. W.H. Gijselaers (Universiteit Maastricht)
Prof. dr. P. van Petegem (Universiteit Antwerpen)
Prof. dr. J.T. Swaab-Barneveld (Universiteit Leiden)

Aan Immanuel Gerardus van Geel
en Valerie Mya Burke

Contents

1 Introduction

Assessment designed to index individual differences in prespecified domains (e.g., mastery of prescribed content in educational and occupational contexts) will always be important, but, increasingly, skills in coping with novelty, generalizing and discriminating dynamic relationships, and making inferences that anticipate distal events are what modern society demands. – Lubinski, 2004

Student admission in higher education remains a controversial topic in the field of education. From an economical perspective, higher education contributes to the productivity of the labor force leading to the economic well-being of a country; and from a social standpoint, it provides opportunities for economic mobility (Kaiser & De Weert, 1995). Conversely, restrictive financial resources from governments reduce participation in higher education. Student selection is one way of regulating participation when the demand for higher education increases while the resources in it remain limited. Student selection aims to improve and maintain the quality of education by providing a balanced student-teacher ratio (Kaiser et al., 1995), and identifying students who would have an increased likelihood of completing the required academic work (Zwick, 2006).

Student admission may be categorized as non-restrictive or restrictive (see also The College Board, 1999). University institutions that consider higher education as an entitlement, or advancement from secondary education employ non-restrictive admission of students, that is, minimum qualifications are accepted such as a high school diploma. On the other hand, university institutions that consider higher education as a reward, or a platform to cultivate talent employ restricted admission of students. Certain admission criteria are required of students and these criteria vary among universities as well as among study programs within a university.

Admission criteria

Admission criteria usually include grade average in prior education and cognitive ability tests. Many empirical studies have shown the predictive validity of these measures (e.g., Kuncel, Hezlett, & Ones, 2001; Kuncel, Hezlett, & Ones, 2004). In employing grade average in prior education as a predictor, one assumes that prior performance of an individual is the best indicator of his or her future performance. On the other hand, this notion is valid in as far as no considerable change occurred in the individual and in the individual's environment (Guthke & Beckmann, 2003). Likewise, performance during prior education depends on the quality of the curriculum pursued. For this reason and with students having various interests and abilities, a common measure of students' abilities has to be developed, thereby the use of standardized admission tests. Generally, standardized admission tests are

cognitive ability measures anchored in trait psychology (see Mislevy, 1996). That is, scores on these measures are considered to be an indication of general intelligence, a rather stable psychological trait that differs among individuals and is largely independent of contextual variations (Barab & Plucker, 2002; Gardner, 2003; Snow, 1994).

In the continued search to improve prediction of academic performance, the use of performance-based tests has expanded the view on admission testing. Performance-based tests in higher education are comparable to work samples in personnel selection (Lievens & Coetsier, 2002). Work samples have demonstrated validity in predicting job performance (Schmidt & Hunter, 1998). Performance-based tests, also known as performance assessments, refer to measurements of behaviors and products carried out in conditions similar to those conditions in which the relevant abilities are actually applied (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education [AERA, APA, and NCME], 1999). Contrary to cognitive ability tests, performance-based tests are rooted in behavioral psychology (see Mislevy, 1996). That is, scores on these tests represent the level of proficiency of an individual in performing a set of tasks similar to those that he or she would eventually encounter during education or employment. In educational settings, this suggests a direct association with the criterion academic performance. In addition, this approach to student ability interprets comprehension, reasoning, and learning as interactive processes between individuals and contexts (Barab & Plucker, 2002; Snow, 1994).

The difference in conceptual paradigm between cognitive ability tests and performance-based tests does not exclude the possibility that both kinds of tests capture similar cognitive processes. However, performance-based tests tap on broader aspects of academic performance. To illustrate, academic work involves tasks such as demonstrating comprehension of theoretical frameworks and applying them to new situations. Performance-based tests capture not only components of cognitive ability such as numerical reasoning and verbal reasoning, but also abilities such as integration of new information with prior knowledge, sifting through relevant and irrelevant information, and formulating coherent arguments (see also Hedlund, Wilt, Nebel, Ashford, & Sternberg, 2006; Lindblom-Ylänne, Lonka, & Leskinen, 1999; Rothstein,

Paunonen, Rush, & King, 1994). With the addition of performance-based tests as an academic predictor, a wider net is cast in the predictor space of academic performance.

Student admission in the Netherlands

Admission testing plays an increasing role in universities in the Netherlands. Within the Dutch educational system, students in secondary education are stratified into a) preparatory vocational education (VMBO, voorbereidend middelbaar beroepsonderwijs); b) preparatory higher professional education (HAVO, hoger algemeen voortgezet onderwijs); or c) preparatory university education (VWO, voorbereidend wetenschappelijk onderwijs) (e.g., De Weert & Boezerooy, 2007). After completing secondary education, VMBO students continue to attend vocational education (MBO, middelbaar beroepsonderwijs). HAVO and VWO students proceed to tertiary education, which is a two-tier system. HAVO students continue to higher professional education (HBO, hoger beroepsonderwijs), and VWO students are directly admitted to university education (WO, wetenschappelijk onderwijs). Higher professional education focuses on practice-oriented education while university education is research-based. The stratification of students at the secondary level as well as at the tertiary level combined with the use of national school examinations at the end of the secondary education makes the implementation of an admission procedure at the tertiary level superfluous.

In recent years however, changes in higher education have put this system under pressure. For economic and cultural reasons, it is deemed increasingly important for students to be able to study abroad. To facilitate cross-country mobility of students, most countries within the European Union have decided to implement a common bachelor-master format in their universities similar to that of North American universities. Within the Netherlands however, student mobility is limited by the two-tier Dutch educational system. A Bachelor's degree in higher professional education does not grant direct admission to a Master's program in a university. For a smooth progress from higher professional education to university education, many universities implemented bridging programs that prepare students for eventual admission to Master's programs. Some universities set up admission

procedures to the bridging programs, mainly because students vary in acquired competencies, and because universities themselves have to largely cover the costs of the bridging programs. The challenge was then to develop admission procedures in universities that would allow students who have an educational background other than a Dutch university education to compete for the limited placement there is. As part of developing admission procedures, the current thesis investigates the utility of a performance-based test over and above traditional academic measures in predicting academic performance of students.

The current thesis

This thesis is on the development and validation of a performance-based test, labeled as Performance Samples on academic tasks in Education and Child Studies (PSEd). PSEd is designed to predict later academic performance through assessment of performance on academic tasks characteristic of those that would eventually be encountered by students in an Education and Child Studies bridging program. In line with the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999), commonly referred to as the *Standards*, sources of validity evidence, reliability, and item properties are addressed in the validation process of PSEd.

Evidence based on test content is a current term for construct validity, and is relevant for proper interpretation of test scores and to develop test items that fall within the relevant construct domain. Evidence based on test content is obtained by examining the relation between test content and the intended construct domain. In the current thesis, the relation between the content of PSEd and the intended construct domain of comprehension tasks as defined by Doyle (1983) is analyzed using confirmatory factor analysis.

Another source of validity evidence is based on the internal structure of a test. A test with high inter-item correlations is a test that is internally consistent (Ghiselli, Campbell, & Zedeck, 1981; Oosterveld & Vorst, 2003). An internally consistent test contains items that largely measure the same attribute. Such a test is likely to have limited predictive value if the criterion represents a substantively wider domain than covered by the test. To maximize

prediction, which is the primary objective of admission testing, a test with low inter-item correlations and simultaneously a high correlation with the criterion of interest is more likely to capture a broad range of abilities that reflect the criterion performance. In the current thesis, the coefficient alpha as a measure of internal consistency is reported to emphasize correlations between tasks included in PSEd.

Internal consistency is one of several estimates of reliability. The term reliability refers to the consistency of test scores if testing is to be repeated several times (e.g., Ghiselli et al., 1981). The classical test theory (CTT), the generalizability theory, and the item response theory (IRT) are test theories that could be applied to estimate reliability. In CTT, reliability may be expressed in terms of internal consistency, test-retest, or split-half. Generalizability theory expresses reliability in terms of sources of variance; IRT expresses reliability in terms of test information function. In addition to reporting internal consistency and test information function, the current thesis reports consistency of pass/fail classification expressed in terms of a dependability coefficient obtained using the generalizability theory. This coefficient indicates in how far an examinee's test score can be consistently classified as below or above a cutoff score (Haertel, 2006).

Evidence based on relations to other variables refers to test-criterion relations that include predictive validity. According to the *Standards*, 'When prediction is actually contemplated, as in education or employment settings, or in planning rehabilitation regimes, predictive studies can retain the temporal differences and other characteristics of the practical situation' (p.14). With PSEd designed for prediction, the academic tasks included in PSEd are characteristic of those that would eventually be encountered by students in an Education and Child Studies bridging program. In the current thesis, regression analysis is employed to examine the predictive validity of PSEd.

With regard to item properties, the *Standards* state the application of CTT or IRT in estimating these properties (p. 44-45). IRT offers several advantages over CTT such as the matching of test items to ability levels, the estimation of ability levels independent of test difficulty, and the estimation of item parameters independent of the sample population (see Hambleton & Jones, 1993; Hambleton, Swaminathan, & Rogers, 1991). Simulation studies on IRT show however, that a large sample size, of more than 500 examinees

depending on the IRT model being fitted, is required to obtain stable parameter estimates (e.g., Barnes & Wise, 1991; Hulin, Lissak, & Drasgow, 1982; Parshall, Kromrey, Chason, & Yi, 1997; Sireci, 1991). Accordingly, the *Standards* emphasize the use of adequate sample size when estimating item properties (p. 44-45). In the current thesis, the application of CTT to estimate item properties was more viable since the data from PSEd involved small sample sizes that range from 100 to 200 students. In CTT, item difficulty in performance-based tests that involve categorical scores is expressed as a ratio between item mean score and the maximum item score possible (Huynh, Meyer, & Barton, 2000 as cited in Johnson, Penny, & Gordon, 2009). Item discrimination is expressed in polyserial correlation between task score and total score (see also Johnson, Penny, & Gordon, 2009).

Having presented how validity evidence, reliability, and item properties have been addressed in the current thesis as indicators of the quality of PSEd as a test used for prediction purposes, specific contents of every chapter are now outlined. Initially, an overview of the relations between academic predictors that have adequately established their validity is provided in Chapter 2 (submitted as Tanilon, Vedder, Segers, & Van Geel, 2011). This chapter aims to advance understanding regarding relations between academic predictors. In Chapter 3 (published as Tanilon, Segers, Vedder, & Tillema, 2009), construct validity, predictive validity, and reliability estimates of PSEd are presented. In addition, this chapter provides an account on how the intended construct domain of comprehension tasks was established. Subsequently, Chapter 4 (submitted as Tanilon, Vedder, & Segers, 2011) examines the degree of similarity of the construct domain across three forms of PSEd using multigroup confirmatory factor analysis. Chapter 5 (published as Tanilon, Vedder, Segers, & Tillema, 2011) examines the incremental validity of PSEd over and above an academic achievement test and grade average in prior education. Newly developed instruments intended for admission decisions should demonstrate incremental validity over and above conventional academic predictors (see also Hunsley & Meyer, 2003). In conclusion, Chapter 6 discusses validation outcomes on PSEd and implications of its use in admission testing, specifically within the Dutch educational context. In view of the continuous development and validation of PSEd during the process of writing the current thesis, the instrument reported in Chapters 3

and 4 consisted of nine tasks, while that reported in Chapter 5 comprised of 12 tasks. In addition, as part of the continuous development and validation of PSEd, Chapter 7 is an appendix that presents findings from a preliminary attempt at applying item response theory using the data reported in Chapter 4. These data were analyzed using the graded response model (Ostini & Nering, 2006; Samejima, 1997). To sum up, this dissertation contains four chapters, a general discussion, and an appendix, all dealing with validation issues on PSEd. Inevitably, there is some overlap between these sections. This dissertation aims to contribute to research on alternative academic predictors to further improve prediction of academic performance.

2 Examining relations between academic predictors in higher education: An overview using meta-analytic path analysis

Submitted for publication

A meta-analytic path analysis was performed to model relations between academic predictors that include general cognitive ability, prior education, declarative and procedural knowledge, personality, and motivation. The criterion of interest is grade average. A regression model, a fully mediated, and a partially mediated model were tested for goodness of fit. Correlations between the academic predictors were obtained from eight meta-analytic studies and used as input data in structural equation modeling. In the absence of meta-analytic studies that examine relations between a few of the academic predictors, five primary studies were obtained to represent these relations. Structural equation modeling was performed using LISREL and results showed that a partially mediated model of academic predictors demonstrated model fit. This model may be used as a guideline in setting up admission procedures and may be expanded to include performance samples.

2.1 Introduction

Prediction of academic performance is one of the more comprehensively investigated topics in the fields of psychology and education. Specifically at the higher educational level, research on academic predictors has been summarized in several meta-analytic studies such as that of Kuncel and colleagues on cognitive ability tests, and study habits, skills, and attitudes (Kuncel, Hezlett, & Ones, 2001; Kuncel, Hezlett, & Ones, 2004; Credé & Kuncel, 2008); Robbins et al. (2004) on psychosocial and study skills; and Trapmann, Hell, Hirn, and Schuler (2007) on personality traits. With admission decisions being made based on these academic predictors, it is relevant to empirically establish the relations between them to serve not only as a guideline in setting up or expanding admission procedures but also to further improve prediction of academic performance. To illustrate, tests of general cognitive ability and grades on prior education are traditionally used as admission criteria. Since both criteria are cognitive measures, a moderate to high correlation between them cannot be ruled out (e.g., Kuncel et al., 2004). The inclusion of these measures in a regression analysis may fail to increase variance accounted for because of their limited contribution to the overall prediction (Smolkowski, 2004). Consequently, to improve prediction of academic performance, other academic predictors should be taken into account, and in doing so, relations between them should be mapped out. By examining models of academic predictors using meta-analytic path analysis, this study aims to advance understanding regarding relations between these predictors, which can lead to improved prediction of academic performance.

According to Credé et al. (2008), academic performance is a function of proximal determinants which in turn are related to distal determinants through mediating variables. Distal determinants refer to general conditions of academic performance such as general cognitive ability, prior training and experience, interests, and personality. Proximal determinants refer to constituents of actual task accomplishment and engagement such as declarative knowledge, procedural knowledge, and motivation. The mediating variables between distal and proximal determinants are study skills, study habits, and study attitudes. As an example, a high score on a general cognitive ability test is related to high grades in school, and this relation is mediated by acquired knowledge about school subjects and study skills. The current study examines

three models of academic predictors adapted from this framework proposed by Credé et al. (2008).

The current study

The criterion of interest is grade average and the academic predictors include general cognitive ability, prior education, declarative and procedural knowledge, personality, and motivation. These predictors have amply established their validity in predicting academic performance, hence their inclusion in the current study. Declarative and procedural knowledge as academic predictors were clustered to form one variable because both types of knowledge are associated with each other in so far as declarative knowledge precedes procedural knowledge (McCloy, Campbell, & Cudeck, 1994). Personality as an academic predictor is operationalized as one of the Big Five factors namely conscientiousness, which has been found to be a valid predictor of academic performance (e.g., Trapmann et al., 2007). Furthermore, motivation as an academic predictor is defined in terms of degree attainment, achievement motivation, study motivation, and performance motivation. These operational definitions of motivation are similar to the extent that they involve completion of academic tasks.

Three models of academic predictors are examined in the current study. The first model is a regression model wherein each of the academic predictors directly relates to academic performance (Figure 2.1). Such a model has been proposed by Trapmann et al. (2007) and is commonly employed in primary studies on the prediction of academic performance. However, with regression analysis, relations between predictors are not explicitly modeled, potentially leading to underprediction. As an example, conscientiousness and motivation as personality-oriented predictors are related such that highly conscientious individuals are likely to be persistent and disciplined, and these behaviors are beneficial when performing and completing tasks (Gellatly, 1996; Judge & Ilies, 2002).

The second model tested is a fully mediated model (Figure 2.2) wherein academic performance is related to general cognitive ability, prior education, and conscientiousness through the mediating factors declarative and procedural knowledge, as well as motivation. This model is in line with Credé et al.'s (2008) point of view that distal academic determinants are fully

mediated by proximal academic determinants. As an example, high general cognitive ability does not necessarily lead directly to successful academic performance. Rather, high general cognitive ability leads to increased understanding of domain-specific tasks that consequently leads to successful academic performance.

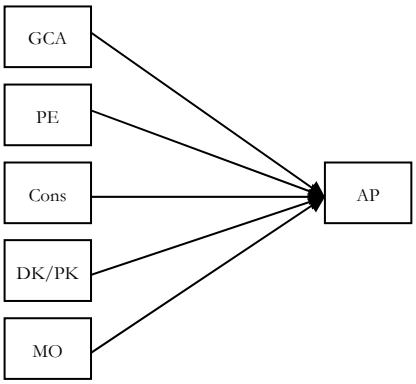


Figure 2.1. A regression model of academic performance.

Note. GCA=general cognitive ability; PE=prior education; Cons=Conscientiousness; DK=declarative knowledge; PK=procedural knowledge; MO=motivation; AP=academic performance.

The third model examined is a partially mediated model (Figure 2.3) wherein general cognitive ability, prior education, and conscientiousness are not only related to academic performance through the mediating factors declarative and procedural knowledge as well as motivation, but also directly linked to academic performance. To illustrate, the fluid component of general cognitive ability is independent of acquired knowledge (Valsiner & Leung, 1994) and may be directly related to academic performance, while the crystallized component of general cognitive ability relies on acquired knowledge (Valsiner et al., 1994) that could serve as a source of information when gaining declarative and procedural knowledge. Note that for the fully and partially mediated model, general cognitive ability and prior education were set to correlate because of their cognitive orientation (see Shavelson & Huang, 2003; Klein, Kuh, Chun, Hamilton, & Shavelson, 2005).

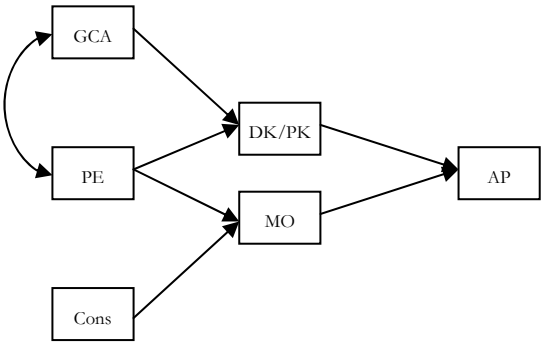


Figure 2.2. A fully mediated model of academic performance.

Note. GCA=general cognitive ability; PE=prior education; Cons=Conscientiousness; DK=declarative knowledge; PK=procedural knowledge; MO=motivation; AP=academic performance.

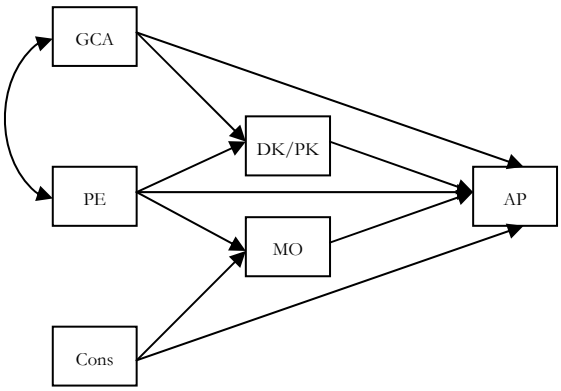


Figure 2.3. A partially mediated model of academic performance.

Note. GCA=general cognitive ability; PE=prior education; Cons=Conscientiousness; DK=declarative knowledge; PK=procedural knowledge; MO=motivation; AP=academic performance.

Correspondingly, the three models examined in the current study are parsimonious adaptations of the Credé et al. (2008) framework to the extent that the mediating factors as study skills, study habits, and study attitudes were left out. This was done for two reasons: (a) to maintain comparability of studies included in the data analysis; and (b) to limit factors that are least likely to be included when setting up admission procedures. In addition, parsimonious models are more likely to be indicative of actual admission procedures especially since there is a strong tendency to set up these procedures as efficiently and time effective as possible.

2.2 Method

Compilation of meta-analytic studies

Meta-analytic path analysis is a methodological approach that combines and re-analyzes studies using structural equation modeling (see Brown et al., 2008). To examine models of academic predictors described, eight meta-analytic studies on predictors of academic performance were identified. In the absence of meta-analytic studies that examine relations between conscientiousness and other predictors, five primary studies were obtained to represent these relations (see Premack & Hunter, 1988 for a comparable method). These meta-analytic and primary studies were published in the last 10 years, used similar samples of participants and comparable operational definitions of academic predictors. Table 2.1 provides an overview of the studies included.

Measures of constructs

Academic performance is operationalized as (graduate) grade point average (GPA), and prior education as undergraduate GPA. With regard to general cognitive ability, there were three measures included namely, the Miller Analogies Test, the Wonderlic Personnel Test, and the Otis-Lennon test of Mental Maturity. The Graduate Record Examinations (GRE; GRE-V, Verbal measure; GRE-Q, Quantitative measure; GRE-A, Analytical measure; GRE-S, Subject Tests) were used as a measure of declarative and procedural knowledge (Kuncel et al., 2001); Conscientiousness as defined by the Big Five personality factors characterizes the construct personality. Examples of measure of conscientiousness are the NEO Five Factor Inventory (NEO-FFI; Costa &

McCrae, 1989, 1992), NEO Personality Inventory (Costa et al., 1992), NEO Personality Inventory Revised (NEO-PI-R; Costa et al., 1992), International Personality Item Pool (IPIP; Goldberg et al., 2006), and the Big Five Inventory (BFI; John, Donahue, & Kentle, 1991). Operational definitions of motivation include degree attainment (Kuncel et al., 2004), achievement motivation characterized by various measures such as the Achievement Scale as reported in the meta-analytic study of Robbins et al. (2004), study motivation as measured by the Learning and Study Skills Inventory (LASSI; Credé et al., 2008), and performance motivation as described in the meta-analytic study of Judge and Ilies (2002).

Procedure

Correlations corrected for attenuation (ρ) were obtained from meta-analytic studies (see Table 2.1). Where there is more than one correlation coded for a particular relation, the mean correlation was calculated. In the absence of meta-analytic studies that support relations between conscientiousness and other academic predictors, primary studies were obtained to represent these relations. Correlations from primary studies are expressed in zero-order correlations. Subsequently, a correlation matrix was formed and used as input data in structural equation modeling.

2.3 Results

Given the lack of clear guidelines as to the sample size to be included in meta-analytic path analysis (Cheung & Chan, 2005), the use of harmonic mean has been recommended (Viswesvaran & Ones, 1995). The harmonic mean of the sample sizes of the studies included in this review is 738. A maximum likelihood procedure using LISREL (Jöreskog & Sörbom, 1996) was used to fit the models to the data. The Comparative Fit Index (CFI), Goodness of Fit Index (GFI), and Standardized Root Mean Square Residual (SRMR) fit indices were used to evaluate measure fit. These measures are robust against small sample size, and it was found in simulation studies that CFI and SRMR are best used for determining the adequacy of the model fit (Cheung & Rensvold, 2002; Hu & Bentler, 1999). Generally, CFI and GFI values of .90 and higher, and a SRMR value lower than .08 indicate acceptable fit.

Table 2.1

List of studies included in the data analysis

Relation	Measures		N	ρ	Study
GCA-AP	Miller Analogies Test	Graduate GPA	11368	0.39	Kuncel, Hezlett, & Ones, 2004
Cons-AP	e.g., NEO-PI-R; IPIP	GPA	10855	0.27	Trapmann, Hell, Hirn, & Schuler, 2007
	e.g., NEO-PI-R; NEO-FFI	Academic performance	5878	0.24	O'Connor & Paunonen, 2007
PE-AP	Undergraduate GPA	Graduate GPA	9748	0.30	Kuncel, Hezlett, & Ones, 2001
DK/PK-AP	GRE-V	Graduate GPA	14156	0.34	Kuncel, Hezlett, & Ones, 2001
	GRE-Q	Graduate GPA	14425	0.32	Kuncel, Hezlett, & Ones, 2001
	GRE-A	Graduate GPA	1928	0.36	Kuncel, Hezlett, & Ones, 2001
	GRE-S	Graduate GPA	2413	0.41	Kuncel, Hezlett, & Ones, 2001
MO-AP	e.g. Achievement Scale	GPA	9330	0.30	Robbins, Lauver, Le, Davis, Langley, & Carlstrom, 2004
	LASSI	GPA	3287	0.38	Credé & Kuncel, 2008
GCA-Cons	*Wonderlic Personnel Test*	*Conscientiousness*	*100*	*0.01*	*Furnham, Moutafi, & Chamorro-Premuzic, 2005*
	Otis-Lennon test of Mental Maturity	*Conscientiousness*	*175*	*0.01*	*Lounsbury, Sundstrom, Loveland, Gibson, 2003*
GCA-PE	Miller Analogies Test	Undergraduate GPA	2999	0.41	Kuncel, Hezlett, & Ones, 2004
GCA-DK/PK	Miller Analogies Test	GRE-V	8328	0.88	Kuncel, Hezlett, & Ones, 2004
	Miller Analogies Test	GRE-Q	7055	0.57	Kuncel, Hezlett, & Ones, 2004
GCA-MO	Miller Analogies Test	degree attainment	3963	0.21	Kuncel, Hezlett, & Ones, 2004

Note. GCA=general cognitive ability; PE=prior education; Cons=Conscientiousness; DK=declarative knowledge; PK=procedural knowledge; MO=motivation; AP=academic performance. Primary studies are in italics. [a]Based on combined sample size.

Table 2.1 (*continued*)

Relation	Measures		N	ρ	Study
Cons-PE	*BFI*	*Freshman GPA*	*131*	*0.17*	*Wagerman & Funder, 2007*
	NEO-FFI	*Freshman GPA*	*432*	*0.17*	*Farsides & Woodfield, 2003*
Cons-DK/PK	*IPIP*	*GRE-V*	*342*	*-0.12*	*Powers & Kaufman, 2004*
		GRE-Q	*342*	*-0.14*	*Powers & Kaufman, 2004*
		GRE-A	*342*	*-0.17*	*Powers & Kaufman, 2004*
Cons-MO	e.g. NEO-PI	Performance motivation (goal-setting)	2211[a]	0.26	Judge & Ilies, 2002
		Performance motivation (expectancy)	1487[a]	0.21	Judge & Ilies, 2002
		Performance motivation (self-efficacy)	3483[a]	0.21	Judge & Ilies, 2002
PE-DK/PK	Undergraduate GPA	GRE-V	6897	0.24	Kuncel, Hezlett, & Ones, 2001
		GRE-Q	6897	0.18	Kuncel, Hezlett, & Ones, 2001
		GRE-A	3888	0.24	Kuncel, Hezlett, & Ones, 2001
		GRE-S	892	0.20	Kuncel, Hezlett, & Ones, 2001
PE-MO	Undergraduate GPA	degree attainment	6315	0.12	Kuncel, Hezlett, & Ones, 2001
DK/PK-MO	GRE-V	degree attainment	6304	0.18	Kuncel, Hezlett, & Ones, 2001
	GRE-Q		6304	0.20	Kuncel, Hezlett, & Ones, 2001
	GRE-A		1233	0.11	Kuncel, Hezlett, & Ones, 2001
	GRE-S		2575	0.39	Kuncel, Hezlett, & Ones, 2001

Note. GCA=general cognitive ability; PE=prior education; Cons=Conscientiousness; DK=declarative knowledge; PK=procedural knowledge; MO=motivation; AP=academic performance. Primary studies are in italics. [a]Based on combined sample size.

Firstly, the regression model was tested (Figure 2.1), wherein all variables directly predict academic performance. This model did not show adequate fit (CFI=.38, GFI=.81, SRMR=.20; R^2=.22). Subsequently, the fully mediated model (Figure 2.2) was examined, with the predictors general cognitive ability and prior education set to correlate. This model too did not provide an adequate fit (CFI=.85, GFI=.93, SRMR=.10; R^2=.17). Finally, the partially mediated model depicted in Figure 2.3 was tested, with the predictors general cognitive ability and prior education set to correlate as well. This model showed acceptable fit of the data (CFI=.93, GFI=.97, SRMR=.07; R^2=.29). Standardized path coefficients in this partially mediated model were significant at .05 alpha level (Figure 2.4). Noticeably, the relation between prior education and declarative and procedural knowledge is negative, which could indicate a suppression effect. That is, prior education accounts for some of the error variance in declarative and procedural knowledge, leading to the latter being an improved predictor of academic performance (Tzelgov & Henik, 1991).

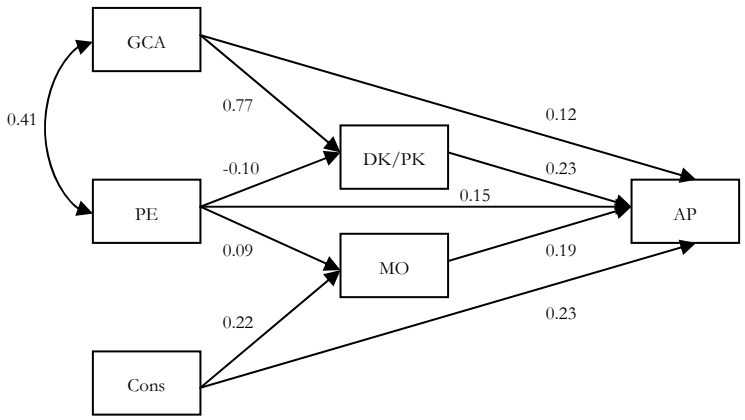


Figure 2.4. Partially mediated model with standardized path coefficients.

Note. GCA=general cognitive ability; PE=prior education; Cons=Conscientiousness; DK=declarative knowledge; PK=procedural knowledge; MO=motivation; AP=academic performance.

2.4 Discussion

This study examined three models of academic predictors using meta-analytic path analysis. The three models examined were regression model, fully mediated model, and partially mediated model. While the fully mediated model fit the data better than the regression model, i.e. the former provides a better description of the relations between academic predictors, the regression model explained more variance in academic performance. In view of this, the association between academic predictors and academic performance is possibly best understood in a partially mediated model, which integrates the fully mediated and the regression model.

The partially mediated model showed adequate fit wherein general cognitive ability, prior education, and conscientiousness are not only related to academic performance through the mediating factors declarative and procedural knowledge as well as motivation, but also directly linked to academic performance. As an example, prior education is directly related to academic performance in so far as prior knowledge serves as a resource that can aid in the completion of an academic task. At the same time, prior education is related to motivation. The association between these two variables, however slight but significant, is such that pursuing an academic career brings with it new challenges; given that past performance is a good indicator of future performance (Guthke & Beckmann, 2003), students with a higher grade average in prior education are more likely to be confident to take up these challenges and stay motivated.

The partially mediated model accounts for 29% of the variation in academic performance. This suggests that future studies will need to look at alternative measures to capture more of the variation in academic performance. Specifically, measures with minimal overlap with the predictors included in the partially mediated model may improve prediction. Some learning theories for example, suggest that context plays a role in academic performance (Anderson, Reder, & Simon, 1996; Bredo, 1994). In response to this, performance-based measures have caught up with the expanding view of admission testing. These measures are 'an attempt to emulate the context or conditions in which the intended knowledge or skills are actually applied' (Lane & Stone, 2006). Drawing on research in personnel selection wherein work samples have demonstrated validity in predicting job performance (Schmidt & Hunter,

1998), research on the use of performance samples in student selection continues to gain attention. Studies of Lievens and colleagues (Lievens, Buyse, & Sackett, 2005; Lievens & Coetsier, 2002) on situational judgment tests; Hedlund, Wilt, Nebel, Ashford, and Sternberg (2006) on the assessment of practical intelligence; and Tanilon and colleagues (Tanilon, Segers, Vedder, & Tillema, 2009; Tanilon, Vedder, Segers, & Tillema, 2011) on performance samples of academic tasks are examples of performance-based measures used as academic predictors.

The limitations of the current study are the restricted operationalizations of the predictors and the criterion, and the use of primary studies to represent relations between the construct conscientiousness and other predictors. The operational definition of the criterion academic performance is grade average. However, there are other aspects of academic performance, which when taken as a criterion, may or may not alter the relations between academic predictors (see also Credé et al., 2008). The same argument can be used if the operational definitions of the academic predictors applied in this study are to be expanded. As to the primary studies obtained to represent relations between the construct conscientiousness and other predictors, these associations are not customarily investigated, thus the absence of meta-analytic studies is to be expected.

The models proposed in this study provide an overview of the abundance of primary research on prediction of academic performance. In doing so, it advances understanding as to the relations of academic predictors and can serve as a guideline in setting up parsimonious but efficient assessment procedures for student admission in higher education.

3 Development and validation of an admission test designed to assess samples of performance on academic tasks

Studies in Educational Evaluation, *35*, 168-173

This study illustrates the development and validation of an admission test, labeled as Performance Samples on academic tasks in Education and Child Studies (PSEd), designed to assess samples of performance on academic tasks characteristic of those that would eventually be encountered by examinees in an Education and Child Studies program. The test was based on one of Doyle's (1983) categories of academic tasks namely comprehension tasks. There were 108 examinees who completed the test consisting of nine comprehension tasks. Factor analysis indicated that the test is basically unidimensional. Furthermore, generalizability analysis indicated adequate reliability of the pass/fail decisions. Regression analysis then showed that the test significantly predicted later academic performance. The implications of using performance assessments such as PSEd in admission procedures are discussed.

3.1 Introduction

The implementation of the internationally recognized Bachelor's and Master's degrees in European universities has increased student mobility, leading to heterogeneity in student populations with regard to prior educational background and previous encounters with various instructional and learning approaches. This has posed the challenge of identifying students who will successfully participate in and complete academic programs, particularly in graduate programs that are popular among students with various educational as well as cultural backgrounds. In response to this development, university officials are searching for ways to increase success rate in the graduate programs that these students intend to enroll in. Many universities require the completion of a bridging program wherein students pursue preparatory courses before they can enroll in the graduate program of their choice (Westerheijden et al., 2008). In addition, admission tests are implemented with the purpose of identifying students who are most able to perform the academic tasks in the bridging programs, thereby increasing success rate in these programs and simultaneously increasing the likelihood of students continuing to and successfully participating in the graduate program of their choice. Students who are most able to perform the academic tasks in the bridging programs are less likely to experience difficulty in coping with academic work and thus presumably obtain passing grades in the courses in these programs. Admission tests then serve as a source of information that predicts performance in the bridging programs. The present study illustrates the development and validation of such an admission test which differs from the traditional predictors of academic performance as grade average in prior education and cognitive ability tests.

Predictors of academic performance

Academic performance is usually operationalized as grade average. Consequently, the continuous use of grade average in prior education, that is, in high school and in the undergraduate level respectively, as a predictor of later academic performance is based on the assumption that prior academic performance is a good estimate of future academic performance (Guthke & Beckmann, 2003). However, as educational curricula and quality of teaching differ across disciplines and among universities and countries, grade average in

prior education does not suffice as a uniform measure of academic abilities (Whitney, 1989). The use of admission tests then becomes essential in as far as they provide standardized measures of students' academic abilities. Scores on these tests can be interpreted as *signs* of underlying cognitive processes or as *samples* of performance (Kane, Crooks, & Cohen, 1999; Messick, 1993; Mislevy, 1994).

Scores on cognitive ability tests are usually interpreted as signs of underlying cognitive processes. These underlying cognitive processes are considered to be rather stable characteristics of an individual independent of the environment he finds himself in (Messick, 1993). The emphasis on individual differences in these cognitive processes has been the focus of many cognitive ability tests used in admission procedures (cf. Gardner, 2003). Meta-analytic studies provide evidence that scores on cognitive ability tests are predictive of grade average in graduate programs (e.g., Kuncel, Crede, & Thomas, 2007; Kuncel, Hezlett, & Ones, 2001). However, a large part of variation in academic performance remains to be explained (Kaplan & Sacuzzo, 2005). Furthermore, cognitive ability tests as usually defined by verbal, spatial and quantitative reasoning (Snow, 1994) hardly represent actual academic performance from which grades are derived. As an example, if one wants to assess examinees' abilities to draw up a research plan, then one can ask them to do so and rate their performance, instead of administering a verbal reasoning test to find out the scope of the vocabulary they can use to draw up a research plan. Direct assessments such as in this example are in line with the framework of performance assessments in which scores are interpreted as samples of performance (Kane et al., 1999; Mislevy, 1994). That is, scores represent an individual's level of proficiency in executing certain tasks similar to that of the criterion of interest.

Using performance assessment as an admission instrument

Formally defined, performance assessments refer to measurements of behaviors and products carried out in conditions similar to those conditions in which the relevant abilities are actually applied (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999). Examples of performance assessments are learning-from-text (LFT) tests which measure critical thinking

skills of medical school applicants and have been found to be predictive of grades in medical courses (Lindblom-Ylänne, Lonka, & Leskinen, 1996, 1999), and objective structured clinical examinations (OSCE's) which assess competence of medical practitioners (e.g., Govaerts, Van der Vleuten, & Schuwirth, 2002; Schoonheim-Klein et al., 2008). Similar to these studies, the admission test described in the present study corresponds with the framework of performance assessment.

The purpose of the admission test, labeled as Performance Samples on academic tasks in Education and Child Studies (PSEd), is to assess samples of performance on academic tasks characteristic of those that would eventually be encountered by examinees in an Education and Child Studies bridging program, and thus identify examinees who are most able to perform the academic tasks involved in the program. PSEd is a criterion-referenced test that focuses on the proficiency level of an examinee to adequately perform a given set of tasks. This is clearly different from the more common approach of norm-referencing based on cognitive ability. Furthermore, where cognitive ability tests are *associated with* academic performance, PSEd is a *direct measure* of academic performance.

The current study contributes to empirical support for using performance assessments in admission procedures, specifically in Educational Sciences, a domain that thus far received little attention in this respect. This study also provides empirical evidence on Doyle's (1983) categories of academic tasks which attempt to define a broader set of abilities embedded in the academic work students encounter at a regular basis. With academic performance as the criterion of interest in admission testing, developing an admission test measuring performance on academic tasks similar to those that examinees would eventually encounter in the educational program of their choice represents an actual demonstration of academic performance. Such an actual demonstration of academic performance from students informs instructors regarding students' level of proficiency in relevant tasks at the beginning of an educational program. This information may eventually allow for an adaptation of instructional activities that is expected to be conducive to students' learning progress. In addition, the development of such a test can lead to the identification and inclusion of predictors of academic performance specific to some disciplines, such as Educational Sciences or Medicine.

Test development

Based on a survey among 17 lecturers and professors involved in a graduate program of Education and Child Studies (Van der Haar & Van Lakerveld, 2004), a list of tasks that students should be able to perform during the graduate program was made. Examples of tasks are applying theories and interpreting statistical results. These tasks were then categorized according to Doyle's (1983) four general types of academic tasks that employ specific cognitive operations necessary to perform the task adequately. Memory tasks are those that require recognition and reproduction of information previously encountered; procedural tasks entail the application of standard methods or formula in providing a response; comprehension tasks involve applying previously encountered information to new situations, recognizing previously encountered information, or formulating assumptions based on previously encountered information; while opinion tasks involve conveying a preference and providing arguments for and against the conveyed preference.

It can be argued that these academic tasks are embedded in the academic work in higher education. Moreover, these categories of academic tasks cover not a single construct but a broader set of abilities. To illustrate, in comprehension tasks, students are expected to apply previously encountered information to new contexts (application tasks), recognize previously encountered information (paraphrase tasks), or draw inferences based on previously encountered information (inference tasks) (Doyle, 1983). In this study, the PSEd contains comprehension tasks that emulate basic critical features of the criterion, that is, academic performance, in as far as these tasks are performed in the bridging program and the products that arise from these tasks are graded.

Validation of test scores

Construct validity and predictive validity are critical aspects of validation studies on admission tests. It is relevant to define what is being measured for a meaningful interpretation of a score (Cronbach, 1971), and it is essential as well that scores on an admission test can predict later academic performance, which is usually operationalized as grade average. Validity theories have influenced the views on validation studies on admission tests. Cognitive ability tests used in admission procedures are usually analyzed

according to the validity theory purported by Cronbach (1971) wherein content validity, construct validity, and criterion-related validity are critical aspects of measurement. While performance assessments are usually evaluated in light of the validity theory proposed by Messick (as cited in Abu-Alhija, 2007; Wolming, 1999) that expands on the critical aspects of validity measurement to include the utility, the social consequences and the value implications of a test (Lane & Stone, 2006; Miller & Linn, 2000). If the use of performance assessments in admission procedures is to be evaluated and compared with cognitive ability tests, then it is sensible to evaluate them in view of the same validity theory, which in turn influences the kind of validation procedures carried out (Guion, 1998). In line with the critical aspects of validity measurement purported by Cronbach (1971), PSEd is evaluated in view of test dimensionality and predictive validity. Test dimensionality, which refers to the minimum number of abilities that can describe score differences among examinees (Tate, 2002), may be reflective of construct validity.

3.2 Method

Sample

One hundred and five female examinees and three male examinees were seeking admission to an Education and Child Studies bridging program. The examinees' mean age was 28 years old (*SD*=7.19). All students completed a Bachelor's degree in Education in the Netherlands.

Predictor variable

The PSEd contains application, paraphrase, and inference tasks, which together define comprehension tasks. There were two application tasks in which examinees were supposed to employ a certain theory relevant in the field of Education and Child Studies to explain the case study in question; three paraphrase tasks wherein examinees were asked to clarify theoretical concepts in a research study; and four inference tasks in which examinees were asked to interpret results of an empirical study (see Table 3.1).

Each task included a text to be read and a question relating to the text. The content of the text varied but remained relevant to the field of Education and Child Studies. The tasks were of constructed-response format and took

Table 3.1

Task samples

Type of tasks	Task sample
Application	Provide a concrete solution to the problem described in the case study. Base your solution on the theory you have read.
Paraphrase	Differentiate deep learning from surface learning approach.
Inference	Interpret the results on the table and relate these results to the theoretical framework discussed in the study.

four hours to complete. The choice for a constructed-response format was based on two reasons: the academic work in the bridging program generally involves constructed responses; and according to Scouller (1998), constructed-response format "allows students control over the selection, organization and presentation of their knowledge and understanding" (p. 455).

There were two independent raters who rated each task according to a 4-score level of a holistic scoring rubric: 1=*poor*; 2=*acceptable*; 3=*good*; and 4=*very good*. Holistic scoring entails grading of overall performance on a task (Lane & Stone, 2006). In this case, raters assigned a single score for each task according to the level of proficiency in which a certain task is performed. When the two raters disagreed by more than one score level in a given task, a third rater was asked to rate the task. Every examinee was given a score on each task, and this score was obtained by taking the score given by the two raters when they agreed, taking the highest score given between the two raters when they disagreed by one score level, or taking the score to which the third rater agreed with one of the two raters when the latter disagreed by two score levels (cf. Kolen, 2006; Lane, Liu, Ankenmann, & Stone, 1996). A score level of 2 (*acceptable*) on each task was selected as the cutoff score for a minimally acceptable performance.

Criterion measure

Grade average in the bridging program is the criterion measure in this study. This was calculated using grades in the completed coursework, with grades being based on a 10-point system.

Psychometric analyses

The 4-score level was ordinal and as such confirmatory factor analysis for ordinal data in LISREL was employed to examine the dimensionality of PSEd. In addition, generalizability and decision studies were conducted to evaluate the reliability of test scores and pass/fail decisions, and to identify the number of tasks that can be used to improve reliability. Two raters scored each task, hence the use of the *Examinees* x *Tasks* x *Raters* (*ptr*) design (Shavelson & Webb, 1991; Brennan, 2001). Inter-rater reliability is expressed in terms of the variance accounted for by the *Raters* (*r*), *Examinees* x *Raters* (*pr*), and *Tasks* x *Raters* (*tr*) facets. The EDUG software (2006) program was used to run generalizability and decision studies. Subsequently, regression analysis was carried out to assess the predictive validity of the test on grade average in the bridging program.

3.3 Results

Test dimensionality

Confirmatory factor analysis for ordinal data was conducted to assess the dimensionality of PSEd. Initially, the polychoric correlation matrix and asymptotic covariance matrix were calculated using PRELIS (Jöreskog & Sörbom, 2006). Each of the polychoric correlation (Table 3.2) met the assumption of bivariate normality. Subsequently, the polychoric correlation matrix was used to estimate parameters through the method of diagonally weighted least squares in LISREL (Jöreskog et al., 2006), which is comparable to robust weighted least squares (Flora & Curran, 2004). Since PSEd is defined as primarily assessing performance on comprehension tasks, a one-factor model (Figure 3.1) was hypothesized. The following indices indicated good fit: $\chi^2(27)=22.34$, $p=.72$, RMSEA=0.00, CFI=1.00 and AGFI=0.98. However, the large unique variances of the tasks suggest that in addition to random error, other abilities specific to every task are captured. Because of the small sample size and the small number of tasks in this study, it was not feasible to perform factor analysis for each type of tasks, namely application, paraphrase, and inference tasks.

Table 3.2

Polychoric correlations between tasks

Task	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1) Application 1								
(2) Application 2	.18							
(3) Paraphrase 1	.22	.32						
(4) Paraphrase 2	.26	.27	.46					
(5) Paraphrase 3	.43	.29	.46	.29				
(6) Inference 1	.31	.34	.41	.44	.37			
(7) Inference 2	.21	.18	.50	.49	.37	.49		
(8) Inference 3	.41	.35	.51	.40	.48	.40	.47	
(9) Inference 4	.32	.20	.34	.32	.50	.23	.31	.36

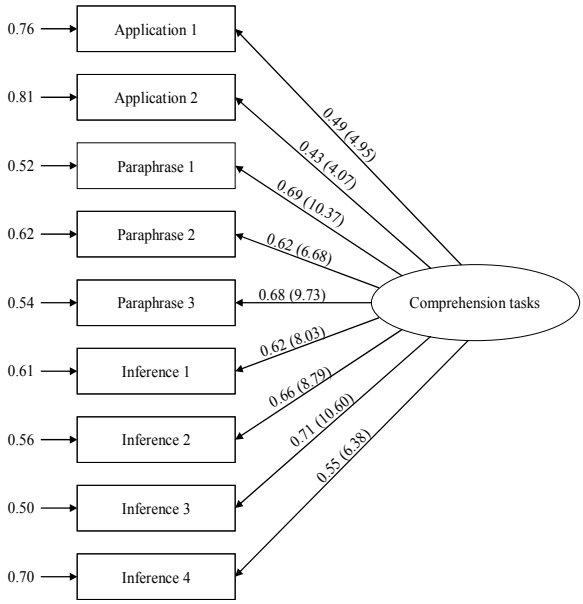


Figure 3.1. Standardized estimates of the hypothesized one-factor model of the Performance Samples on academic tasks in Education and Child Studies (*t*-values in parentheses).

Reliability of test scores

The substantial agreement between raters is reflected in the minute amount of variance accounted for by the *r* facet, and the *pr* and *tr* interaction facets (Table 3.3). The *p* facet indicates differential performance of examinees, while the *t* facet suggests variation in tasks. The largest amount of variance is accounted for by the *pt* interaction facet, which shows that examinees' scores vary across tasks. Some examinees consistently obtained high or low scores across tasks, and other examinees scored high on some tasks and low on other tasks. The *ptr* interaction facet indicates that error variance is minimal.

Table 3.3

Sources of variation with their estimated variance

Source of Variation	*df*	Mean Squares	Estimated Variance Component	Percentage of Total Variance
Examinees (*p*)	107	6.06	.27	25.5
Tasks (*t*)	8	25.06	.11	10.3
Raters (*r*)	1	0.37	.00	0.0
Examinees x Tasks (*pt*)	856	1.24	.58	56.1
Examinees x Raters (*pr*)	107	0.10	.00	0.3
Tasks x Raters (*tr*)	8	0.81	.01	0.7
Examinees x Tasks x Raters (*ptr*)	856	0.07	.07	7.1

The reliability of the test scores is reflected in the dependability coefficient of Φ=.76, which can be considered as adequate at this initial stage of test development and validation (Nunnally & Bernstein, 1994), and taking into account the small number of tasks. This value though is lower than the required reliability of >.90 for high-stakes decisions. On the other hand, the reliability of the pass/fail decisions meets this requirement with a dependability coefficient of $\Phi(\lambda)$=.92. The $\Phi(\lambda)$ coefficient denotes "the accuracy with which a test indicates examinees' distance from the cut score" (Haertel, 2006: p. 100). The cutoff score was set at score level 2 (*acceptable*) in making pass/fail decisions. This cutoff score defines the response criteria for a minimally acceptable performance.

Accordingly, a decision study was carried out to determine the number of tasks necessary to improve reliability. Since the tasks require a constructed-response format, the maximum number of tasks that can eventually be administered is estimated at 20. Increasing the number of tasks to 20 with two raters rating each task provides a dependability coefficient of Φ=.88, which is still somewhat lower than the >.90 requirement. Using the $\Phi(\lambda)$ coefficient instead may ameliorate reliability since PSEd entails pass/fail decisions.

Predicting academic performance

Mean scores on the PSEd were used in regression analysis to examine the predictive validity of the test on grade average in the bridging program. The grand mean score was 3.15 (*SD*=.39). Results showed that PSEd significantly predicted grade average in the bridging program β=.38, $t(62)$=3.22, p=.002 with an explained variance of R^2=.14, $F(1,62)$=10.34, p=.002. The β value of .38 is considered to be high for admission purposes (Kaplan & Sacuzzo, 2005).

3.4 Discussion

This study illustrates the development and validation of PSEd, an admission test designed to assess samples of performance on academic tasks characteristic of those that would eventually be encountered by examinees in an Education and Child Studies bridging program, and thus identify examinees that are most able to perform the academic tasks involved in the program. The test was based on one of Doyle's (1983) categories of academic tasks namely comprehension tasks. Results showed that the test is basically unidimensional. Moreover, the reliability of PSEd scores can be considered adequate considering the small number of tasks involved, though lower than the required reliability of >.90 for high-stakes decisions. Nonetheless, the reliability of the pass/fail decisions meets this requirement. PSEd scores predicted grade average in the bridging program as well. The test explained 14% of variance in grade average in the bridging program which can be considered high for admission purposes (Kaplan & Sacuzzo, 2005).

In view of these results, the use of performance assessments in predicting later academic performance shows potential considering that performance assessments attempt to capture a broader set of abilities that can

be based on the general categories of academic tasks described by Doyle (1983). In this study however, PSEd was limited to comprehension tasks. Whether the other academic tasks described by Doyle (1983) will further improve the amount of variance in the grade average in the bridging program that can be explained by PSEd is yet to be explored.

Sampling performance on academic tasks focuses on the proficiency of a student to perform a task adequately within a relevant domain. This study though did not take into account how samples of performance on academic tasks relate to traditional predictors of academic performance particularly that of cognitive ability tests. This question is yet to be answered but for now the assumption is that samples of performance on academic tasks have incremental value over and above cognitive ability tests. It may be argued that the same underlying cognitive processes are involved in samples of performance on academic tasks as well as in cognitive ability tests. However, in samples of performance on academic tasks, the stimuli are context-specific. As such, students' responses are accentuated. Taking the study of Saxe (as cited in Barab, & Plucker, 2002) on children's arithmetic as an example, it was shown that children selling products in markets provided correct answers to arithmetic problems that take place in the markets 99% of the time. Upon presenting the same arithmetic problems on a math test, the same children got the correct answers only 65% of the time.

The use of performance assessments in high-stakes decisions has been hindered not only by the time and costs it takes to administer them (Ryan, 2006) but also by issues of task specificity, that is, low correlations between task scores (Kane et al., 1999). Low correlations between task scores decrease the internal consistency of the test (Ghiselli, Campbell, & Zedeck, 1981; Oosterveld & Vorst, 2003). If one develops a test with high correlations between tasks or items however, one has a test that is internally consistent, but the predictive power of the test decreases. To maximize prediction, which is the prime objective of admission testing, one has to have low correlations between task scores but high correlations between task scores and the criterion of interest. A test that highly correlates with the criterion captures broader abilities. PSEd has been indicated as basically unidimensional, but the large unique variances of the tasks suggest that in addition to random error, other abilities specific to every task are captured. Performance assessments such as

PSEd tap into broader abilities, and thus may further improve prediction of academic performance.

Using performance assessments for admission purposes may be informative as well. Instructors are informed about students' level of proficiency in relevant tasks at the beginning of an educational program. They are then better able to monitor changes in students' level of proficiency in the course of the curriculum and may accordingly adapt instructional activities beneficial to students' learning progress. As for prospective students, performance assessments allow them to be confronted with relevant tasks that they have to perform if admitted in the educational program of their choice. In this case, they would be better able to decide whether their preferred program approaches their expectations, leading to a better and more committed choice, eventually decreasing dropout rates during the educational program itself. Performance assessments as admission instruments therefore may not only be predictive of later academic performance but also informative for instructors as well as for prospective students.

4 Score comparability and incremental validity of a performance assessment designed for student admission

Submitted for publication

This study examines comparability of scores from three forms of a performance assessment designed for student admission. The incremental validity of the performance assessment forms over and above an academic achievement test is examined as well. There were three cohorts with 108, 171, and 144 students, respectively. Prior to admission to a study program, the students completed the performance assessment consisting of nine comprehension tasks. Score comparability was analyzed using multigroup confirmatory factor analysis. Results showed that the three performance assessment forms demonstrate similar measurement intent. Factor loadings and error variances however, differ across the forms. Subsequently, hierarchical regression analysis showed that the performance assessment forms have significant incremental validity over and above an academic achievement test in predicting later academic performance. In view of these results, the use of performance assessments for student admission purposes is discussed.

4.1 Introduction

Performance assessments are alternative tools used to evaluate student performance. Formally defined, they are measured behaviors and products carried out in conditions similar to those in which the relevant abilities are actually applied (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999). The concept of performance assessments may be appealing to many educational practitioners; however, validation issues have hampered the use of these alternative tools. In the area of construct validation for example, performance assessments are likely to fall short since they are not particularly designed to measure a single construct but a constellation of constructs (see Maclellan, 2004). Furthermore, comparability of scores from and incremental validity of performance assessments are validation issues that have been scarcely addressed through empirical evidence (see also Haertel & Linn, 1996; Elliot & Fuchs, 1997). Comparable scores across assessment or test forms, which are designed to measure the same attribute while the items differ across forms (Kaplan & Sacuzzo, 2009), are essential as they increase proper interpretation of scores. That is, scores are given the same meaning regardless of which test form was taken by an examinee (Muraki, Hombo, & Lee, 2000).

As to incremental validity, performance assessments as alternative academic measures capture not only components of cognitive ability such as numerical reasoning and verbal reasoning, but also abilities such as integration of new information with prior knowledge, sifting through relevant and irrelevant information, and formulating coherent arguments (see Hedlund, Wilt, Nebel, Ashford, & Sternberg, 2006; Lindblom-Ylänne, Lonka, & Leskinen, 1999; Rothstein, Paunonen, Rush, & King, 1994). Performance assessments therefore, may demonstrate significant potential as predictors of academic performance.

In view of these notions, it is of importance to build empirical evidence through validation on the utility of performance assessments so as to guide educational practitioners when employing these tools. The aim of the current study then is to contribute to empirical evidence on the use of performance assessments, particularly in student admission procedures in higher education, by examining the comparability of scores from and

incremental validity of three forms of a performance assessment. The three performance assessment forms are designed to evaluate performances of students on academic tasks that emulate those that are typically encountered in a bridging program aimed at improving academic skills of students before being admitted to a Master's program. The academic tasks included in this study are comprehension tasks which involve applying previously encountered information to new situations, recognizing previously encountered information, or formulating assumptions based on previously encountered information (Doyle, 1983).

Comparability of scores from performance assessments

Establishing comparability of scores from performance assessments that involve few test items, open-ended responses, and ratings by judges can be challenging (see Kolen, 1999). According to Haertel et al. (1996) there are three components that have to be taken into account when comparing scores across performance tasks. These are measurement intent, which pertains to construct-relevant abilities the task intends to measure; ancillary abilities, which are construct-irrelevant abilities necessary for adequate task completion; and error variance, denoting random and unique attributes that influence scores. To illustrate, a task requiring students to interpret results of a research study and relate them to a certain theoretical framework is rated according to the correct interpretation of the results and a coherent synthesis between results and theoretical framework. The task is designed to measure the ability to interpret results and synthesize them with theory. At the same time, the level of familiarity with statistics and with the theoretical framework at hand is an ancillary requirement to adequately perform the task. The research study presented could be a random influence on the performance of the task.

The degree of similarity of measurement intent and error variances across forms of a performance assessment may be examined using multigroup confirmatory factor analysis (CFA). Since CFA does not provide separate estimates of specific variance and measurement error variance (Brown, 2006) however, the degree of similarity of ancillary requirements across assessment forms may be reflected in the factor loadings and error variances. The goal of multigroup CFA is to analyze measurement invariance across groups, that is, whether differences in observed scores across groups indeed reflect differences

in performance (Wicherts, 2007). Accordingly, groups from different populations are compared as to their scores on a certain measure. It is feasible as well to compare scores from test forms taken by students from the same population using multigroup CFA (see also Ameriks, 2009).

The present study examines comparability of scores from three performance assessment forms, taken by students from the same population, using multigroup CFA. A similar factorial model across the three performance assessment forms, that is, the presence of configural invariance, would indicate similar measurement intent across the three forms. Subsequently, adding the constraint of equal factor loadings, i.e. metric invariance, would provide evidence regarding the strength of association between the performance tasks and the hypothesized construct domain across the forms. Further constraining the factorial model to have equal residual variances would test for the comparability of error variances across the forms. The similarity of the ancillary abilities across the forms may be reflected in metric invariance and equal error variances. If tasks have high factor loadings and low error variances that are invariant across assessment forms, ancillary abilities and error variances are comparable; simultaneously, less of these ancillary abilities and error variances are in play.

Incremental validity of performance assessments

Alternative measures designed to predict academic performance should have incremental validity over and above traditional academic predictors to demonstrate their utility. Performance assessments as alternative measures directly reflect criterion performance (see Kane, Crooks, & Cohen, 1999); contrary to traditional academic predictors such as academic achievement tests that focus on prior knowledge (see Sternberg, 1999), thus limiting evaluation of potential performance (see Zysberg, Levy, & Zisberg, 2011) This contrast suggests that performance assessments may have significant additional value in predicting academic performance. In this study, the incremental validity of the three performance assessment forms over and above an academic achievement test is examined.

4.2 Method

Setting

The Netherlands has a two-tier university system: the higher professional and the academic tier. Higher professional colleges focus on practice-oriented education while universities focus on research-oriented education. To facilitate student mobility, most countries within the European Union have decided to implement a common bachelor-master format in their universities similar to that of North American universities. A drawback of this common format is that students with a Bachelor's degree from the higher professional tier are not granted direct admission to an academic Master's program. Instead, these students can be first admitted to a bridging program aimed at improving their academic skills, allowing them to catch up with those students who are granted direct admission. The data obtained for this study include students' scores on the three performance assessment forms administered prior to admission to an Education and Child Studies bridging program.

Sample

Students completed a Bachelor's degree in Education from the higher professional tier in the Netherlands. There were 108, 171, and 144 students in Cohort 1, 2, and 3, respectively. In Cohort 1, there were 105 female and three male students with a mean age of 28 years (*SD*=7.19). In Cohort 2, there were 160 female and 11 male students with a mean age of 26 years (*SD*=5.87). In Cohort 3, there were 136 female and eight male students with a mean age of 28 years (*SD*=7.07). The female-male ratio on this sample population reflects that of the bridging program as well as of the subsequent Master's program wherein far more female than male students are enrolled.

Predictor variables

Academic achievement test. This standardized test includes language, science, and math subjects administered at the end of the secondary education. It is comparable to SAT Subject Tests. Composite scores based on a 10-point scale were used.

Performance assessment. This measure comprised of application, paraphrase, and inference tasks, which together define comprehension tasks

(Doyle, 1983). These tasks reflect those that are typically encountered in the bridging program. There were two application tasks in which students were supposed to use a certain theory relevant in the field of Education and Child Studies to explain the case study in question; three paraphrase tasks wherein students were asked to paraphrase definitions of concepts in a research study; and four inference tasks in which students were asked to draw inferences from results of an empirical study.

Each task included a text to be read and a question relating to the text. The content of the text varied but remained relevant to the field of Education and Child Studies. The tasks were of constructed-response format and took four hours to administer. The choice for a constructed-response format was based on two reasons: the academic work in the bridging program generally involves constructed-response format; and response construction provides students the option to select, organize and present their knowledge and understanding (Scouller, 1998).

The tasks were rated according to a 4-score level of a holistic scoring rubric: 1=*poor*; 2=*acceptable*; 3=*good*; and 4=*very good*. Holistic scoring describes overall task performance (Lane & Stone, 2006). To ensure that raters' scores are consistent with the scoring rubrics, two independent raters rated each task for Cohort 1 and 2. For each task, raters assigned a single score that corresponds to a set of criteria a task response had to meet. In case of rater disagreement by more than one score level in a given task, a third rater was asked to rate the task. Every student obtained a score on each task based on the score given by the two raters when they agreed, the highest score given between the two raters when they disagreed by one score level, or taking the score to which the third rater agreed with one of the two raters when the latter disagreed by two score levels (cf. Kolen, 2006; Lane, Liu, Ankenmann, & Stone, 1996). The average inter-rater reliability of the nine tasks has a weighted kappa value of .85 and .61 for Cohort 1 and 2, respectively. According to Landis and Koch (1977), a kappa statistic in the range of 0.61-0.80 and 0.81-1.00 indicates substantial and almost perfect inter-rater agreement, respectively.

A score level of 2 (*acceptable*) on each task was chosen as the cutoff score for a minimally acceptable performance. The reliability of this cutoff score is denoted by the dependability coefficient of $\Phi(\lambda)=.92$ for Cohort 1, $\Phi(\lambda)=.82$ for Cohort 2, and $\Phi(\lambda)=.70$ for Cohort 3. These values signify "the

accuracy with which a test indicates students' distance from the cut score" (Haertel, 2006: p. 100).

The task difficulty and task discrimination indices for all cohorts are provided in Table 4.1. Difficulty indices are expressed as a ratio of item mean to maximum item score possible (Huynh, Meyer, & Barton, 2000 as cited in Johnson, Penny, & Gordon, 2009). Difficulty indices considered acceptable are between the range of .30 and .90, with indices around .50 highly contributing to the total score variance. As shown in Table 4.1, tasks performed by all cohorts have acceptable difficulty indices. Task discrimination indices were expressed in polyserial correlations, which indicate the association between task score and total score (see also Johnson, Penny, & Gordon, 2009). Tasks with discrimination indices of 0.2 and higher acceptably discriminate between low scoring and high scoring students (Ebel, 1972). Discrimination indices shown in Table 4.1 suggest that students with low task scores tend to get low total scores.

Table 4.1

Difficulty and discrimination indices of performance tasks

Task	Cohort 1		Cohort 2		Cohort 3	
	diff	*dis*	*diff*	*dis*	*diff*	*dis*
Application 1 (Develop a plan)	.58	.65	.31	.40	.52	.59
Application 2 (Connect results to theory)	.73	.69	.49	.53	.62	.64
Paraphrase 1 (Explain concepts)	.75	.70	.67	.54	.56	.69
Paraphrase 2 (Describe research design)	.64	.53	.74	.36	.57	.47
Paraphrase 3 (Formulate goal of research)	.74	.74	.63	.59	.50	.52
Inference 1 (Relate question to design)	.66	.72	.75	.46	.69	.61
Inference 2 (Derive conclusion)	.75	.72	.67	.54	.61	.50
Inference 3 (Interpret tables and graphs)	.88	.57	.61	.58	.64	.43
Inference 4 (Criticize research design)	.73	.77	.63	.38	.46	.37

Note. diff, item difficulty; *dis*, item discrimination.

Criterion measure

Grade average on the completed coursework in the bridging program is the criterion measure in this study. Grades are based on a 10-point system.

4.3 Results

Multigroup CFA was carried out in LISREL (Jöreskog & Sörbom, 2006) to examine score comparability among three performance assessment forms. The method of maximum likelihood estimation was used because it is relatively robust for departures from multivariate normality (Raykov & Marcoulides, 2000) and allows for corrections of standard errors and chi-square statistic for non-normality (Jöreskog, 2005; Millsap & Yun-Tein, 2004). Thresholds were set to be equal for all tasks.

The performance assessment comprised of application, paraphrase, and inference tasks, which together define comprehension tasks (Doyle, 1983). Following this definition, a one-factor model was hypothesized. This one-factor model shows good fit for each cohort as indicated by the fit indices of the single group analyses in Table 4.2. That is, for each cohort, the tasks represent the hypothesized domain of comprehension tasks. The one-factor model was then tested for configural invariance, metric invariance, and equal residual variances across cohorts (Table 4.2). The one-factor model shows configural invariance but not metric invariance. This result indicates that there is similar measurement intent across forms but that the strength of association between measurement intent and tasks, i.e. factor loadings, differ across forms. Table 4.3 shows that the standardized factor loadings in Cohort 1 are larger than in Cohort 2 and Cohort 3. Fit indices of the one-factor model with the additional constraint of equal residual variances show poor fit as well, suggesting that error variances differ across forms.

Predicting academic performance

Means, standard deviations, and intercorrelations between predictors and criterion are shown in Table 4.4. Notably, there is a weak to negative correlation between the predictors. This suggests that the predictors capture different sets of abilities. While the academic achievement test primarily assessed prior knowledge, the performance assessment may have well emulated critical features of later academic performance.

Table 4.2
Measurement invariance across cohorts

Measurement model	S-B χ^2	*df*	*p*-value	ΔS-B χ^2	*df*	RMSEA (90% CI)	CFit	CFI	NNFI
Single group									
Cohort 1 (*n* = 108)	18.23	27	.90			0.00 (0.00-0.03)	0.98	1.00	1.00
Cohort 2 (*n* = 171)	18.24	27	.90			0.00 (0.00-0.03)	0.99	1.00	1.00
Cohort 3 (*n* = 144)	30.78	27	.28			0.03 (0.00-0.08)	0.71	0.98	0.97
Measurement invariance									
Configural invariance	65.82	81	.89			0.00 (0.00-0.02)	1.00	1.00	1.00
Metric invariance	151.18	99	.00	135.83**	18	0.06 (0.04-0.08)	0.17	0.93	0.92
Equal residual variances	322.82	117	.00	172.28**	18	0.11 (0.10-0.13)	0.00	0.71	0.73

Note. N = 423. ΔS-B χ^2, nested χ^2 difference; RMSEA, root mean square error of approximation; 90% CI, 90% confidence interval for RMSEA; Cfit, probability RMSEA $\leq$.05; CFI, comparative fit index; NNFI, non-normed fit index. **p<.01.

Table 4.3
Standardized factor loadings for the configural invariance measurement model across cohorts

Task	Cohort 1			Cohort 2			Cohort 3		
	factor loading	*SE*	*t*-value	factor loading	*SE*	*t*-value	factor loading	*SE*	*t*-value
Application 1 (Develop a plan)	0.63	0.13	4.84	0.21	0.08	2.78	0.56	0.13	4.39
Application 2 (Connect results to theory)	0.59	0.09	6.64	0.34	0.09	3.78	0.49	0.09	5.70
Paraphrase 1 (Explain concepts)	0.87	0.11	7.70	0.54	0.17	3.23	0.53	0.09	6.17
Paraphrase 2 (Describe research design)	0.33	0.09	3.74	0.08	0.11	0.70	0.21	0.07	2.95
Paraphrase 3 (Formulate goal of research)	0.79	0.08	9.27	0.47	0.12	4.02	0.24	0.06	4.31
Inference 1 (Relate question to design)	0.96	0.11	8.70	0.12	0.10	1.23	0.33	0.07	4.81
Inference 2 (Derive conclusion)	0.38	0.07	5.44	0.20	0.08	2.39	0.29	0.12	2.35
Inference 3 (Interpret tables and graphs)	0.38	0.09	4.25	0.32	0.08	4.24	0.22	0.07	3.38
Inference 4 (Criticize research design)	0.85	0.09	9.61	0.13	0.07	1.82	0.14	0.09	1.48

Table 4.4

Means, standard deviations, and intercorrelations of predictors and criterion

Variable	*M*	*SD*	1	2
Cohort 1				
1. Academic achievement test	6.41	0.78		
2. Performance assessment	3.15	0.39	.06	
3. Grade average in the bridging program	7.34	0.62	.16	.38**
Cohort 2				
1. Academic achievement test	6.36	0.62		
2. Performance assessment	2.54	0.34	.03	
3. Grade average in the bridging program	7.02	0.59	.15	.33**
Cohort 3				
1. Academic achievement test	6.47	0.55		
2. Performance assessment	2.36	0.37	.11	
3. Grade average in the bridging program	7.01	0.54	.00	.28**

Note. $N = 264$. Values in parentheses are one-tailed p-values. ** $p<.01$.

Hierarchical regression was employed to examine the incremental validity of the performance assessment over and above the academic achievement test in predicting grade average in the bridging program. Results in Table 4.5 show that, for all cohorts, the performance assessment has significant incremental validity over and above the academic achievement test in predicting grade average in the bridging program. The partial correlations between the predictors and criterion, specifically in Cohort 3, may be lower than what could actually be found. That is, correlations between variables in the sample population tend to be lower than in the total population, and this may be attributed to selection effects (De Gruijter & Van der Kamp, 2008; Sackett & Yang, 2000).

4.4 Discussion

This study examined score comparability and incremental validity of three performance assessment forms designed to assess samples of performance on academic tasks characteristic of those that are encountered by students in an Education and Child Studies bridging program. In using performance assessments for admission purposes, it is crucial to demonstrate comparability of scores from these assessments since score interpretation

Table 4.5
Hierarchical regression analyses predicting grade average in the bridging program

Predictor	Cohort 1				Cohort 2				Cohort 3			
	Model 1		Model 2		Model 1		Model 2		Model 1		Model 2	
	β	*r*	*β*	*r*	*β*	*r*	*β*	*r*	*β*	*r*	*β*	*r*
Academic achievement	.16	.16	.13	.14	.15	.15	.14	.15	.00	.00	-.03	-.03
Performance assessment			.37	.37			.32	.33			.28	.28
R^2	.02		.16		.02		.13		.00		.08	
F	1.52		5.81*		1.94		6.16**		0.00		4.79*	
ΔR^2			.14				.11				.08	
ΔF			9.88**				10.18**				9.58**	

Note. *β*, standardized regression coefficient; *r*, partial correlation. ** $p<.01$. * $p<.05$.

should be consistent across test administrations. Scores from the three forms of performance assessment examined in this study are comparable in as far as they show configural invariance. However, the forms lack metric invariance and equality of error variances. Performance assessment tasks, although designed according to the same specifications, may vary in difficulty as well as in ancillary abilities required by a task (see also Ackerman, 1986; Maclellan, 2004). The varying degrees of these facets may well be reflected in the lack of metric invariance and equality of error variances. Simultaneously, configural invariance suggests that the scores from these three forms reflect similar measurement intent.

This study also showed that the performance assessment forms have incremental validity in predicting academic performance. Performance assessments cover a large space of the construct domain that typifies a given criterion, resulting in construct overrepresentation. Contrary to traditional academic predictors such as admission tests that narrowly measure cognitive ability, leading to construct underrepresentation. Construct overrepresentation does not necessarily have to be a problem in prediction of performance because the criterion of interest may involve the same range of abilities as the performance assessment (Messick, 1993). Performance assessments can thus function as an academic predictor.

In the current study, a one-factor model was fitted that defines the tasks included in the performance assessment. This may seem inconsistent with the notion that performance-based tests tend to assess a constellation of constructs (Maclellan, 2004). However, the considerable unique variances of the tasks in the one-factor model suggest that in addition to random error, other abilities specific to every task are captured. Further, that scores on the performance assessment were more valid than academic achievement test scores in predicting grade average in the bridging program may be partly attributed to temporal proximity. That is, association between a predictor and a criterion is stronger if performance on both variables occurs temporally close. In this case, the time interval between performance assessment and performance in the bridging program was nine months which is much shorter than the time interval between performance on the academic achievement test and performance in the bridging program which was four years. Accordingly, the meta-analytic study of Hulin, Henry, and Noon (1990) on predictive validity coefficients across time showed that the longer the time that has elapsed between prediction of performance and criterion performance itself, the weaker the predictive validity of a variable becomes. In admission procedures then, time as a facet in predictor-criterion relations should be taken into account. Finally, the tasks used in the performance assessment were tailored to those that are performed in the bridging program. On the one hand, this limits the generalizability of the findings of this study. On the other hand, what is required of an adequate performance of academic tasks varies across disciplines such as Educational Sciences or Psychology. If a test adequately represents tasks typical of a given discipline, performances on such specific tasks could contribute to improving prediction of academic performance.

5 Incremental validity of a performance-based test over and above conventional academic predictors

Learning and Individual Differences, *21(2)*, 223-226

As has been presented in the previous chapter, PSEd shows incremental validity over and above an academic achievement test. In the current chapter, the incremental validity of PSEd is further examined. Specifically, the value of PSEd as an academic predictor in addition to grade average in prior education and academic achievement test is analyzed. In view of the continuous development and validation of PSEd, the instrument reported in Chapters 3 and 4 consisted of nine tasks, while that reported in the present chapter comprised of 12 tasks. Further, the data discussed in the present chapter involved a single cohort.

The present chapter focuses on the conceptual distinction between conventional academic predictors and performance-based tests. Conventional academic predictors include grade average in prior education and academic achievement tests. Performance-based tests involve direct measures of a criterion. Subsequently, an empirical study is presented examining the incremental validity of a performance- based test over and above conventional academic predictors. The test consisted of 12 knowledge application and inferential tasks. The data included records of 150 students enrolled in an Education and Child Studies bridging program, a program that links undergraduate study to academic graduate study in the Netherlands. Hierarchical regression analysis showed that the performance-based test has incremental validity in predicting academic performance operationalized as grade average. Accordingly, performance-based tests demonstrate potential as an academic predictor.

5.1 Introduction

The prime objective of admission testing in higher education is to predict academic performance. The most common operationalization of academic performance is grade average, that is, a higher grade average is indicative of academic capability. Being academically capable increases the probability of completing an education and moving on to a desired professional career. This prospect is conceivably one of the reasons why there is a constant attempt to improve prediction of academic performance. Grade average in prior education and academic achievement tests have established their predictive value; hence their use has become conventional. The use of grade average in prior education and academic achievement tests as academic predictors stems from the assumption that prior performance is a good estimate of future performance (see also Guthke & Beckmann, 2003).

Studies investigating the predictive validity of conventional academic predictors show that undergraduate grade point average (GPA) can account for approximately 9-12% of variance in graduate GPA (e.g., Kuncel, Credé, & Thomas, 2007; Kuncel, Hezlett, & Ones, 2001) and that academic achievement tests such as SAT Subject Tests can account for as much as 16% of variance in GPA (e.g., Geiser & Santelices, 2007; Geiser & Studley, 2002). A large percentage of variance in academic performance however, is yet to be explained (Kaplan & Sacuzzo, 2005).

The constant attempt to improve the prediction of academic performance has led to studies that investigate the predictive validity of performance-based tests. These tests, also known as performance assessments, are measured behaviors and products carried out in conditions similar to those in which the relevant abilities are actually applied (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999). Examples of performance-based tests are learning-from-text (LFT) tests (Lindblom-Ylänne, Lonka, & Leskinen, 1996, 1999) that measure critical thinking skills and have been found to be predictive of grades in medical courses, and situational-judgment inventories (SJIs) that measure problem solving skills and have been found to have incremental validity over and above the Graduate Management Admission Test (GMAT; Hedlund, Wilt, Nebel, Ashford, & Sternberg, 2006).

The current study extends existing research on performance-based tests by examining samples of performance on academic tasks as an academic predictor. Academic tasks emulate tasks embedded in the academic work students encounter at a regular basis in as far as the products that arise from these tasks are graded. A test that assesses performance on academic tasks similar to those that students would eventually encounter in the educational program of their choice represents an actual demonstration of academic performance, which is the criterion measure in admission testing. The aim of the current study then is to examine the incremental validity of this performance-based test over and above conventional academic predictors. A conceptual distinction between these measures from the perspective of higher education is first provided.

Performance-based tests and conventional academic predictors

Performance-based tests and conventional academic predictors as grade average in prior education and academic achievement tests are similar in as far as they are cognitive outcome measures (Shavelson & Huang, 2003; Klein, Kuh, Chun, Hamilton, & Shavelson, 2005). What distinguishes performance-based tests from conventional academic predictors is that the latter focus on prior knowledge and developed abilities (see also Sternberg, 1999; Kuncel et al., 2001), while the former as direct measures of the criterion tap into potential performance.

Whereas conventional academic predictors primarily reflect students' prior knowledge, performance-based tests assess students' performance on tasks that are relevant to the academic program they are applying for. If students have encountered these relevant tasks previously, then it is most likely that performance-based tests and conventional academic predictors are highly associated. In tasks that require declarative knowledge for example, such high association could be expected. However, if the relevant tasks are novel to the students, then it can be assumed that a different set of abilities is likely to be captured by performance-based tests. Performance on these relevant tasks may involve a set of abilities that depends not only on prior knowledge but also on abilities such as integration of new information with prior knowledge, sifting through relevant and irrelevant information, and formulating coherent arguments (see also Hedlund et al., 2006; Lindblom-Ylänne et al., 1999;

Rothstein, Paunonen, Rush, & King, 1994). Performance-based tests then become informative tools that can assess abilities involved in potential performance.

The direct link between predictor and criterion such that similar behavior or performance is measured on both has been underrated (Reeve & Hakel, 2002). Performance-based tests can be of additional value in providing direct measures of the criterion. As an example, History students may be required to write a literature review while Psychology students may be required to write a research plan. One can sample students' performance on tasks similar to the required academic work and rate their performance as an alternative to administering a verbal reasoning test to find out the scope of the vocabulary students can use to perform the required academic work. Performance-based tests, as measures of samples of performance, include tasks which stimuli are context-specific and directly similar to criterion tasks.

The current study

In light of the distinctions between performance-based tests and conventional academic predictors, the incremental validity of a performance-based test over and above conventional academic predictors is examined in this study. The purpose of the performance-based test is to assess samples of performance on academic tasks characteristic of those that would eventually be encountered by students in an Education and Child Studies bridging program, and thus identify students who are most able to perform the academic tasks involved in the program. Bridging programs, wherein students are required to pursue preparatory courses before they can enroll in the graduate program of their choice, link undergraduate programs to academic graduate programs in the Netherlands. Bridging programs were established as a result of changes in European higher education to accommodate students from within and outside Europe (Westerheijden et al., 2008). Students who are most able to perform the academic tasks in the bridging programs are likely to cope with the academic work and obtain passing grades, eventually minimizing the dropout rates in these programs and simultaneously increasing the likelihood of students continuing to the graduate program of their choice.

The academic tasks emulated in the performance-based test include knowledge application and inferential tasks, which together define

comprehension tasks (Doyle, 1983). Knowledge application tasks involve applying theoretical concepts in solving case problems, and inferential tasks pertain to deducing new information from previously encountered information. These tasks are commonly embedded in academic work; in this particular study however, performances on these tasks are explicitly assessed as elements of the criterion in a particular domain. Consequently, the use of performance-based tests as direct measures of the criterion can lead to the identification and inclusion of predictors of academic performance specific to some disciplines, such as Psychology or Educational Sciences.

5.2 Method

Sample

Data in this study was derived from the records of students who took the performance-based test and were admitted to the Education and Child Studies bridging program. There were 147 female and three male students with a mean age of 27 years (*SD*=6.50).

Predictor variables

Grade average in prior education. The grade average in the last two years of the students' higher vocational education was calculated. This variable represents performance in advanced courses at the higher vocational education level. The grades were based on a 10-point scale.

Academic achievement test. The standardized national examination at the end of the secondary education involves tests on language, science, and math subjects. This academic achievement test is comparable to SAT Subject Tests that closely reflect school subjects instead of a specific school curriculum. Composite scores on this examination were based on a 10-point scale.

Performance-based test. The test consisted of 12 tasks that require students to apply theoretical concepts in solving case problems, and draw inferences using previously encountered information. In view of the difference in academic work between higher vocational education and the bridging program (Witte, Van der Wende, & Huisman, 2008), these tasks are considered to be fairly novel to the students. The tasks were of constructed-response format and took six hours to administer. The choice for a constructed-response format was based on two reasons: the academic work in the bridging

program generally involves constructed-response format; and response construction provides students the option to select, organize and present their knowledge and understanding (Scouller, 1998). There were two independent raters who rated each task according to a 4-level holistic scoring rubric: 1=*poor*; 2=*acceptable*; 3=*good*; and 4=*very good*. Holistic scoring entails a scoring rubric describing overall performance on the task (Lane & Stone, 2006). Table 5.1 provides task samples and their corresponding scoring rubrics. The inter-rater reliability estimate using weighted kappa is .75, which is within the range of substantial level of inter-rater agreement (Landis & Koch, 1977). The Cronbach's alpha for the 12 tasks is .68, which suggests that the tasks are relatively heterogeneous. Accordingly, a measure with low coefficient alpha is not necessarily a fundamental obstruction to its use if it includes tasks that cover relevant components of the criterion (Schmitt, 1996). To assess dimensionality, categorical principal component analysis was performed. A one-dimensional component solution accounted for 26% of score variance. This dimension involved knowledge application and inferential tasks, which together define comprehension tasks (Doyle, 1983).

Table 5.1

Task samples with corresponding scoring rubrics

Task sample	Scoring rubric
Provide a concrete solution to the problem described in the case study. Base your solution on the theoretical framework described in the text.	The student provides a solution that 4 - correctly applies the theoretical framework described in the text; is coherent, feasible, and can be empirically tested. 3 - correctly applies the theoretical framework described in the text; is coherent, and feasible. 2 - correctly applies the theoretical framework described in the text, and is coherent. 1 - incorrectly applies the theoretical framework described in the text, or is incoherent.

Table 5.1 (*continued*)

Task sample	Scoring rubric
Interpret the results on the tables and relate these results to the theoretical framework discussed in the text.	The student 4 - correctly interprets results; coherently synthesizes results and theoretical framework, and sifts through results relevant to the theoretical framework at hand. 3 - correctly interprets results, and coherently synthesizes results and theoretical framework. 2 - correctly interprets results; coherently synthesizes results and theoretical framework, but uses irrelevant outside information. 1 - incorrectly interprets results, or fails to synthesize results and theoretical framework.

Criterion measure

Grade average in the bridging program is the criterion measure in this study. This was calculated using grades in the completed coursework, with grades being based on a 10-point scale.

5.3 Results

Means, standard deviations and intercorrelations of the predictors and criterion are given in Table 5.2. The only correlation between predictors that reach statistical significance (α=.05) is that of grade average in prior education and academic achievement test. A negative correlation between these two predictors was found. This result could be attributed to the subject contents included in prior education and academic achievement test. In this case, while the academic achievement test consisted of language, science and math subjects, prior education also involved practical courses aimed at developing workplace-related skills of the students. The negative correlation then suggests that academic achievement is weakly negatively associated with practical skills (see also Sternberg, 1999).

To analyze the incremental validity of the performance-based test in predicting later academic performance, hierarchical regression analysis was conducted. Results are shown in Table 5.3. The criterion variable is grade average in the bridging program. The first model included grade average in prior education and academic achievement test score as predictors. The second model included both these variables and the mean score on the performance-based test as predictors. The first model showed that grade average in prior education and academic achievement test score were significant predictors of grade average in the bridging program. The inclusion of the mean score on the performance-based test in the second model showed that this mean test score

Table 5.2

Means, standard deviations, and intercorrelations of predictors and criterion

Variable	*M*	*SD*	1	2	3
1. Grade average in prior education	7.38	0.39			
2. Academic achievement test	6.39	0.66	-.16(.02)*		
3. Performance-based test	2.70	0.40	.12(.07)	.06(.21)	
4. Grade average in the bridging program	7.15	0.62	.25(.00)**	.15(.04)*	.46(.00)**

Note. $N = 150$. Values in parentheses are one-tailed *p*-values. *$p<.05$. **$p<.01$.

Table 5.3

Hierarchical regression analyses of grade average in the bridging program on predictors

Predictor	Model 1		Model 2	
	β	*r*	β	*r*
Grade average in prior education	.29**	.28	.23**	.25
Academic achievement test	.19*	.20	.16*	.18
Performance-based test			.42**	.44
R^2	.10		.27	
F	8.28**		18.09**	
ΔR^2			.17	
ΔF			33.99**	

Note. β, standardized regression coefficient; *r*, partial correlation. *$p<.05$. **$p<.01$.

as well as grade average in prior education and academic achievement test score are significant predictors of grade average in the bridging program. The significant partial correlation between performance-based test score and grade average in the bridging program implies that the former is predictive of the latter when combined with grade average in prior education and academic achievement test score. The performance-based test thus has incremental validity over and above grade average in prior education and academic achievement test, and significantly explained 17% of variance in grade average in the bridging program.

5.4 Discussion

Whereas conventional academic predictors focus on prior knowledge and developed abilities (see also Sternberg, 1999; Kuncel et al., 2001), performance-based tests as direct measures of the criterion tap into potential performance. These distinctions between conventional academic predictors and performance-based tests are highlighted in this study that illustrates the incremental validity of a performance-based test over and above conventional academic predictors in predicting grade average in a bridging program.

Limitations of the presented empirical study on the incremental validity of a performance-based test include the female gender majority that characterized the sample population and the setting in which the study took place. As to female gender majority, review studies have shown decreasing cognitive differences between gender (e.g., Hyde, 1981; Hyde & Lynn, 1988; Voyer, Voyer, & Bryden, 1995; Wilder & Powell, 1989). As such, gender may play a negligible role when it comes to generalizing the study findings to both male and female students. As to the setting, admission practices differ between universities and countries and what could be a valid predictor in one setting does not necessarily have to be a valid predictor in another setting. This study though contributes to empirical support for using supplementary measures, particularly that of performance-based tests, which could be considered upon searching for valid predictors of academic performance. Another limitation is the criterion used in this study. Grade average is a composite of knowledge, skills and abilities employed in academic work. Academic work itself differs in content and difficulty, and hence uniformity of grades is lacking (Wolming, 1999). Future studies on predictor-criterion relations should try to take into

account criterion measures other than grade average that can be considered comparable across settings such as total point requirement, duration of study, and degree attainment.

The use of performance-based tests brings with it problems of construct validity and internal consistency (Kane, Crooks, & Cohen, 1999; Maclellan, 2004). Paradoxically, an internally consistent test may limit predictive validity and a highly predictive test could have low internal consistency (Ghiselli et al., 1981; Oosterveld et al., 2003). With prediction of later academic performance as the prime objective of admission testing, performance-based tests as direct measures of the criterion may contribute to this objective. It is essential as well to be able to define which elements of the criterion are being targeted by performance-based tests.

Using conventional academic predictors alone would likely result in limited predictive validity and may have unfavorable effects on particular groups of students that differ in educational, cultural and socio-economic backgrounds. Performance-based tests are measures that can be used in addition to more conventional measures to improve prediction of academic performance. They serve as tools that assess not only developed abilities but also potential performance of students. In considering potential performance, unfavorable effects such as bias against certain groups are reduced.

Using performance-based tests may be informative as well. Instructors are informed about students' level of proficiency in relevant tasks at the beginning of an educational program. They are then better able to monitor changes in students' level of proficiency in the course of the curriculum and may accordingly adapt instructional activities beneficial to students' learning progress. Performance-based tests also allow students to be confronted with relevant tasks that they have to perform if admitted in the educational program of their choice. In this case, they would be better able to decide whether their preferred program approaches their expectations, leading to a better and more committed choice, eventually decreasing dropout rates during the educational program itself. Performance-based tests as admission measures therefore may not only be predictive of later academic performance but also informative for instructors as well as for students.

6 Discussion

This thesis is about the development and validation of a performance-based test, labeled as Performance Samples on academic tasks in Education and Child Studies (PSEd). PSEd is designed to identify students who are most able to perform the academic tasks involved in an Education and Child Studies bridging program. Many Dutch universities set up bridging programs that aim to prepare students with non-university degrees in the Netherlands for Master's programs at the university level. Some universities set up admission procedures to the bridging programs, primarily because students vary in acquired competencies, and because of the limited resources accessible to the bridging programs. The development and validation of PSEd is part of establishing an admission procedure for the Education and Child Studies bridging program at Leiden University. In the process of developing and validating the PSEd, sources of validity evidence, reliability, and item properties were addressed in accordance with the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999).

Validation outcomes on PSEd

One source of validity evidence is test content, traditionally known as construct validity. The tasks included in the PSEd were developed according to Doyle's (1983) categories of academic tasks. Using confirmatory factor analysis, a one-factor model was hypothesized that pertains to comprehension tasks. This may seem inconsistent with the notion that performance-based tests tend to assess a constellation of constructs (Maclellan, 2004). The one-factor model simply defines the common factor that has an effect on test performance (see Hambleton, Swaminathan, & Rogers, 1991). The large unique variances of the tasks in the one-factor model suggest that the items may actually be defined by a multidimensional factor model. However, fitting a multidimensional factor model was not feasible because of the small sample size from which the data is obtained.

In the process of construct validation, comparable scores on test forms are essential as they increase proper interpretation of scores. Score comparability of three PSEd forms showed that the three forms demonstrate similar measurement intent but that factor loadings and error variances differ

across the forms. Performance assessment tasks, however designed according to the same specifications, may vary in difficulty and construct-irrelevant variance (see also Ackerman, 1986; Maclellan, 2004) that may well be reflected in a lack of metric invariance and equality of error variances. Future validation studies on PSEd should consider analyzing task characteristics and response processes that could influence metric invariance and equality of error variances (see Lane, Wang, & Magone, 1996).

The internal consistency of a test and the association of the test with a criterion are sources of validity evidence that make up two sides of the same coin. Specifically in Chapter 5, results suggest that the tasks included in the PSEd were not highly correlated with each other, i.e. low internal consistency; at the same time, PSEd as a test showed validity in predicting grade average in the bridging program. Accordingly, a measure with low coefficient alpha is not necessarily a fundamental obstruction to its use if it includes tasks that cover relevant components of the criterion (Schmitt, 1996). For future validation studies on PSEd, one way of improving internal consistency is to generate several subtests that each contains highly correlated tasks.

Reliability of PSEd

The *Standards* point out that estimates of reliability obtained from different test theories are not necessarily comparable to each other (p. 32). In classical test theory (CTT), reliability may be expressed in terms of internal consistency, test-retest, or split-half. Generalizability theory expresses reliability in terms of sources of variance; item response theory (IRT) expresses reliability in terms of test information function. While reliability of PSEd is estimated from these different theories, attention is particularly given to the consistency of pass/fail classifications. Performance-based tests used to make competence-based decisions, such as PSEd, usually involve the classification of students into competence or noncompetence (see also Luecht, 2006). As in this thesis, pass/fail classifications are made based on what is considered as a minimally acceptable performance. Tasks included in the PSEd were rated according to a 4-score level: 1=*poor*; 2=*acceptable*; 3=*good*; and 4=*very good*. A score level of 2 (*acceptable*) on each task was selected as the cutoff score for a minimally acceptable performance. Incorrect classifications were minimized by setting the cutoff score below what is considered as good performance, thereby

allowing for measurement error (see Maurer, 2005). As such, the risk of false negatives, i.e. the error of classifying students with proficient skills as not being proficient is reduced. However, the risk of false positives, i.e. the error of classifying students with inadequate skills as proficient is increased.

The degree of classification consistency can be expressed in a dependability coefficient which indicates in how far examinees' test scores can be consistently classified as below or above a cutoff score (Haertel, 2006). Reliability however, is influenced by test length, sample size, and number of scale points (Fitzpatrick & Yen, 2001; Shumate, Surles, Johnson, & Penny, 2007). Given this and the preliminary stage in which PSEd was validated, the dependability coefficients of $\Phi(\lambda)=.92$, $\Phi(\lambda)=.82$, and $\Phi(\lambda)=.70$ reported in Chapter 4 for three cohorts can be considered as satisfactory degrees of classification consistency (see Nunnaly et al., 1994). For admission decisions however, dependability coefficients should have values $\geq .95$. Hence, the reported values indicate a need for further improvement of the reliability of PSEd. Further improvement could include the use of selected-response items, such as multiple-choice questions, with or instead of constructed-response items, such as essay questions. By doing this, more items could be incorporated in the test. Using selected-response items does not necessarily mean loss of information relative to constructed-response items. Some studies show that selected-response items and constructed-response items are highly correlated (see Bridgeman & Morgan, 1996; Hancock, 1994).

Item properties

Using CTT, item difficulty and item discrimination indices for the tasks included in the PSEd were calculated. CTT is especially useful when the sample size is small (Hambleton & Jones, 1993). Results showed that the PSEd tasks were of adequate difficulty and acceptably discriminate between low-scoring and high-scoring students. For a more precise estimation of item parameters however, the application of IRT is recommended, which requires a large sample size. For this dissertation, collecting test data from a large body of students to facilitate the use of IRT in determining the quality of the test items was not feasible. Firstly, there is not a large body of students who seek admission to the program. At this point, self-selection occurs among students who seek admission and who do not. The former group is likely to be more

motivated to do academically well than the latter group (see also Ryan, Ployhart, Gregoras, & Schmit, 1998). Secondly, administering the test and then allowing non-restrictive admission to the program would result in students not seriously taking the test because no consequences are attached to the test results (Nedermeijer, De Gruijter, & Wijdeveld, 2006). Nonetheless, future validation studies on PSEd should find a way to administer the test to a large number of students for a more precise estimation of item as well as ability parameters using IRT. There are several IRT models that could be used to fit test results data, one of which is the graded response model (GRM; Ostini & Nering, 2006; Samejima, 1997). The GRM can be used to model data composed of polytomous scores from constructed-response items (see Chapter 7 for a preliminary attempt at applying GRM to PSEd). Further, one has to take into account the multidimensionality of the tasks included in performance-based tests when fitting IRT models.

Is PSEd a valid test?

As a predictor yes, as a construct less so. This statement epitomizes the tension between construct representation and predictive utility of a test (Borsboom, Mellenbergh, & Van Heerden, 2004), and is reflected in the bandwidth-fidelity trade-off. Bandwidth refers to the breadth of abilities that reflect the criterion of interest and fidelity pertains to the precision in which these abilities can be measured (Hogan & Roberts, 1996). Cognitive ability tests commonly used in student admission procedures, for example, measure with high precision a narrow range of the domain that reflects academic performance as the criterion of interest, that is, these tests are of low bandwidth and high fidelity. On the other hand, performance-based tests such as PSEd measure with modest precision a wide range of the domain that reflects academic performance as the criterion of interest, that is, these tests are of high bandwidth and low fidelity.

One implication of using performance-based tests such as PSEd for student admission purposes is that these tests can have significant incremental value beyond conventional academic predictors. In addition, performance-based tests provide a baseline of students' abilities which can be employed to assess students' learning growth and to evaluate the effectiveness of a curriculum. The use of performance-based tests, however, has been hindered

by the costs and laborious work that take to develop them (Hardy, 1995). Future studies on the development of performance-based tests should examine whether the benefits derived from these tests can offset the costs that are involved in developing them.

Performance-based tests within the Dutch educational context

The issue of implementing admission testing prior to entry to universities is a controversial topic in the Netherlands. The main arguments against admission testing prior to entry to universities, specifically for Dutch students, are that a) students are stratified according to school grades at the beginning of the secondary educational level; b) the national examinations at the end of the secondary education is a form of admission testing as is; and c) economical opportunities for and social involvement of the youth are optimized through participation in higher education. For these reasons, there is scant empirical knowledge within the Dutch context as to the incremental value of potential academic predictors beyond the conventional ones that are in use at the secondary educational level. As to students with an educational background other than Dutch university schooling, admission testing could aid in identifying students whose capacities are at least at par with what could be expected of students with Dutch university schooling.

If admission testing is to be implemented in higher education, for Dutch university students and students with an educational background other than Dutch university schooling alike, it is relevant to identify potential academic predictors that have incremental value beyond conventional academic predictors. This thesis proposed performance-based tests as a potential academic predictor. Specifically, performance samples on academic tasks expand the covered prediction space of academic performance.

7 Appendix: A preliminary attempt at applying the graded response model to PSEd[1]

[1] Dr. Eduardo Cascallar, Dr. Rudy Ligtvoet, Dr. Dimitri Rizopoulos, and Dr. Matthijs Warrens provided comments to an earlier version of this appendix. Particularly helpful were the comments of Dr. Rudy Ligtvoet on the interpretation of obtained parameter estimates relative to model-data fit, and the comments of Dr. Dimitri Rizopoulos on the calculation of the rule of thumb for three-way margins and for pointing out that the margins approach in the ltm package is an indication of model fit rather than a strict statistical test.

Item response theory (IRT) focuses on the premise that observed performance on test items can be explained by a latent ability (Hambleton, Swaminathan, & Rogers, 1991). There are various IRT models that could be applied to explain a given test data, one of which is the graded response model (GRM; Ostini & Nering, 2006; Samejima, 1997). The GRM is suitable for test data comprised of ordered polytomous score categories such as PSEd. In polytomous IRT models, such as the GRM, score categories are separated by category boundaries (Ostini et al., 2006). In the case of PSEd, this means that the 4 score levels are separated by three category boundaries, the boundary between score level 1 and 2, 2 and 3, and 3 and 4, respectively. Category boundaries are used to determine the probability of passing the steps required to obtain a particular score level. A category boundary is represented by a category boundary response function (CBRF), which is characterized by a discrimination parameter and boundary location parameters. For a more detailed explanation of polytomous IRT models, the work of Ostini et al. (2006) can be consulted.

Using the ltm package in R software (R Development Core Team, 2005), GRM was applied to the data on the three forms of PSEd presented in Chapter 4. Item parameters were estimated using the marginal maximum likelihood method (Rizopoulos, 2006). For each of the three PSEd forms, item discrimination parameter a_i, and boundary location parameters b_{ik} are estimated (Table 7.1). Item discrimination may be interpreted according to the qualitative classification proposed by Baker (1985): a_i<0.20, very low discrimination; a_i ranging between 0.21-0.40, low discrimination; a_i ranging between 0.41-0.80, moderate discrimination; a_i>0.80, high discrimination. The location parameter b_i for each of the k category boundaries is indicative of item difficulty. Generally, items in the three PSEd forms have moderate to high discriminating power with the exception of Items 5 and 6 in PSEd Form 2, and Item 9 in PSEd Forms 2 and 3. In addition, Item 6 in PSEd Form 2 required extremely low ability level to obtain a score level of 1. Item 9 in PSEd Form 3 on the other hand required extremely high ability to obtain a score level of 4.

The test information function (TIF) for each PSEd form is shown in Figures 7.1, 7.2, and 7.3, respectively. TIF represents the amount of test information across the ability continuum (Wainer & Thissen,1996). Figures 7.1, 7.2, and 7.3 show that the area under the TIF for ability levels in the interval

(-4,4) amounts to 89.83%, 65.39%, and 73.85%, respectively. PSEd Form 1 yielded more test information than PSEd Form 2 and 3.

As to model fit, the ltm package employs χ^2 goodness-of-fit test to examine if the model fits the data. For each of the three forms of PSEd, some of the χ^2 residuals across three items considered are higher than the rule of thumb of $3.5*i_n*j_n$, wherein n is the number of categories for items i and j, respectively (Rizopoulos, 2006). This suggests that GRM inadequately fits the data.

This appendix is an initial attempt at applying the GRM to PSEd data. Results presented in this appendix are exploratory and should be cautiously interpreted for a number of reasons. First, the sample size on which the results are based is small. The application of IRT models requires large sample sizes to obtain stable parameter estimates (e.g., Hulin, Lissak, & Drasgow, 1982; Sireci, 1991). Second, the interpretation of the obtained parameter values and the TIFs presented in this appendix depend on the adequacy of the GRM as a model to describe the data. Here, it was found that GRM showed inadequate fit to the data. IRT assumes that the test results can be explained by a unidimensional latent variable, that is, all test items should measure one and the same latent ability. Performance-based tests such as PSEd do not necessarily assume a single latent ability (see Lane & Stone, 2006). While a unidimensional model is used to describe the tasks included in PSEd, the large error variances of these tasks suggest that the tasks may in fact be multidimensional, which might explain the inadequate model-data fit. In addition, unidimensional models fitted to tests composed of multidimensional items may lead to a decrease in test information (Luecht & Miller, 1992). Third, IRT is a powerful tool to analyze the quality of test items, simultaneously, "optimizing measurement properties and optimizing predictive properties are not convergent lines of test construction" (Borsboom, Mellenbergh, & Van Heerden, 2004, p. 1067).

Table 7.1

GRM item parameters for three forms of PSEd

Task item	PSED Form 1				PSED Form 2				PSED Form 3			
	a_i	bi_1	bi_2	bi_3	a_i	bi_1	bi_2	bi_3	a_i	bi_1	bi_2	bi_3
Item 1 (Develop a plan)	1.14	-2.32	-0.42	0.36	0.49	-2.59	2.94	10.64	0.89	-1.61	0.08	1.85
Item 2 (Connect results to theory)	1.36	-4.08	-2.33	-0.31	0.81	-2.88	0.33	2.64	1.19	-2.29	0.04	1.58
Item 3 (Explain concepts)	1.34	-0.81	0.18	1.49	0.85	1.82	4.29	6.46	1.25	-0.61	0.59	2.90
Item 4 (Describe research design)	1.68	-1.56	0.03	0.82	0.92	-3.06	-1.55	0.96	1.00	-5.49	-1.02	3.39
Item 5 (Formulate goal of research)	1.50	-1.77	-0.78	0.35	0.38	-9.96	0.33	3.47	1.26	-1.32	0.33	3.95
Item 6 (Relate question to design)	0.86	-3.27	0.03	2.42	0.13	-29.27	-5.38	6.02	0.68	-3.78	1.08	5.96
Item 7 (Derive conclusion)	0.93	-3.91	-1.1	1.3	1.00	-3.57	0.14	2.63	0.67	-2.39	2.49	7.70
Item 8 (Interpret tables and graphs)	1.56	-3.72	-0.65	0.53	0.86	-5.57	0.06	1.84	0.48	-2.68	-0.43	4.01
Item 9 (Criticize research design)	1.76	-1.8	-0.57	0.54	0.37	-7.59	-0.32	7.60	0.25	-2.80	7.89	13.41

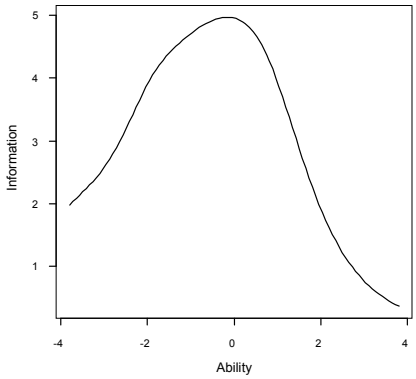


Figure 7.1. Test information function of PSEd Form 1.

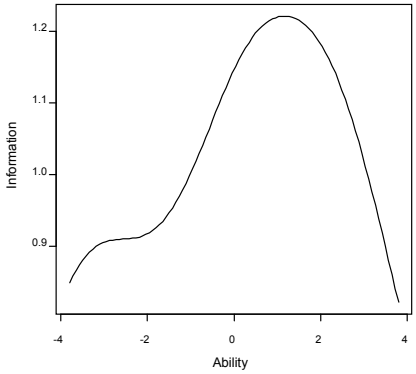


Figure 7.2. Test information function of PSEd Form 2.

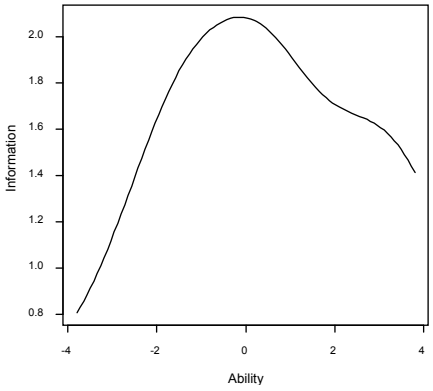


Figure 7.3. Test information function of PSEd Form 3.

References

Abu-Alhija, F.N. (2007). Large-scale testing: Benefits and pitfalls. *Studies of Educational Evaluation*, *33*, 50-68.

Ackerman, P.L. (1986). Individual differences in information processing: An investigation of intellectual abilities and task performance during practice. *Intelligence*, *10*, 101-139.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Ameriks, Y.S. (2009). *Investigating validity across two test forms of the examination for the certificate of proficiency in English (ECPE): A multi-group structural equation modeling approach* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (Publication no. 3348566)

Anderson, J.R., Reder, L.M., & Simon, H.A. (1996). Situated learning and education. *Educational Researcher*, *25(4)*, 5-11.

Baker, F.B. (1985). *The basics of item response theory*. Portsmouth, NH: Heineman.

Barab, S.A., & Plucker, J.A. (2002). Smart people or smart contexts? Cognition, ability, and talent development in an age of situated approaches to knowing and learning. *Educational Psychologist*, *37(3)*, 165-182.

Barnes, L.L.B., & Wise, S.L. (1991). The utility of a modified one-parameter IRT model with small samples. *Applied Measurement in Education*, *4(2)*, 143-157.

Borsboom, D., Mellenbergh, G.J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111(4)*, 1061-1071.

Bredo, E. (1994). Reconstructing educational psychology: Situated cognition and deweyian pragmatism. *Educational Psychologist*, *29(1)*, 23-35.

Brennan, R.L. (2001). *Generalizability theory*. New York: Springer.

Bridgeman, B., & Morgan, R. (1996). Success in college for students with discrepancies between performance on multiple-choice and essay tests. *Journal of Educational Psychology*, *88(2)*, 333-340.

Brown, T.A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: The Guilford Press.

Brown, S.D., Tramayne, S., Hoxha, D., Telander, K., Fan, X., & Lent, R.W. (2008). Social cognitive predictors of college students' academic performance and persistence: A meta-analytic path analysis. *Journal of Vocational Behavior*, *72*, 298-308.

Cheung, M. W. L., & Chan, W. (2005). Meta-analytic structural equation modeling: A two-stage approach. *Psychological Methods*, *10*, 40–64.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233–255.

Costa, P.T., & McCrae, R.R. (1989). *NEO-PI/NEO-FFI Manual supplement*. Odessa, FL: Psychological Assessment Resources.

Costa, P.T., & McCrae, R.R. (1992). *Revised NEO personality inventory (NEO-PI-R) and NEO five-factor inventory (NEO-FFI): Professional manual.* Odessa, FL: Psychological Assessment Resources.

Credé, M. & Kuncel, N.R. (2008). Study habits, skills, and attitudes: The third pillar supporting collegiate academic performance. *Perspectives on Psychological Science*, 3(6), 425-453.

Cronbach, L.J. (1971). Test validation. In R.L. Thorndike (Ed.), *Educational Measurement*, (2nd edition). Washington DC: American Council on Education, 443-507.

De Gruijter, D.N.M., & Van der Kamp, L.J.Th. (2008). *Statistical test theory for the Behavioral Sciences*. Boca Raton, FL: Taylor & Francis Group.

De Weert, E., & Boezerooy, P. (2007). Higher education in the Netherlands. Universiteit Twente: Center for Higher Education Policy Studies.

Doyle, W. (1983). Academic Work. *Review of Educational Research*, *53*, 159-199.

Ebel, R.L. (1972). *Essentials of educational measurement* (1st ed.). New Jersey: Prentice Hall.

EDUG Software, 2006. Switzerland: Swiss Society for Research in Education Working Group.

Elliott, S.N., & Fuchs, L.S. (1997). The utility of curriculum-based measurement and performance assessment as alternatives to traditional intelligence and achievement tests. *School Psychology Review*, *26(2)*, 224-233.

Farsides, T. & Woodfield, R. (2003). Individual differences and undergraduate academic success: the roles of personality, intelligence, and application. *Personality and Individual Differences*, *34*, 1225-1243.

Fitzpatrick, A.R., & Yen, W.M. (2001). The effects of test length and sample size on the reliability and equating of tests composed of constructed-response items. *Applied Measurement in Education*, *14(1)*, 31-57.

Flora, D.B., & Curran, P.J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*, 466-491.

Furnham, A., Moutafi, J., & Chamorro-Premuzic, T. (2005). Personality and Intelligence: Gender, the Big Five, self-estimated and psychometric intelligence. *International Journal of Selection and Assessment*, *13(1)*, 11-24.

Gardner, H. (2003). Three distinct meanings of intelligence. In R.J. Sternberg, J. Lautrey, & T.I. Lubart (Eds.), *Models of intelligence: International perspectives* (pp. 43-54). Washington DC: American Psychological Association.

Gellatly, I.R. (1996). Conscientiousness and task performance: Test of a cognitive process model. *Journal of Applied Psychology*, *81(5)*, 474-482.

Ghiselli, E.E., Campbell, J.P., & Zedeck, S. (1981). *Measurement Theory for the Behavioral Sciences*. New York: W.H. Freeman and Company.

Geiser, S., & Santelices, M.V. (2007). *Validity of high-school grades in predicting student success beyond the freshman year: High-school record vs. standardized tests as indicators of four-year college outcomes* (Research Report No. CSHE.6.07). Retrieved from University of California website: http://cshe.berkeley.edu/publications/docs/ROPS.GEISER._SAT_6.12.07.pdf

Geiser, S., & Studley, R. (2002). UC and the SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California. *Educational Assessment*, *8(1)*, 1-26

Goldberg, L.R., Johnson, J.A., Eber, H.W., Hogan, R., Ashton, M.C., Cloninger, C.R. et al. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*, 84-96.

Govaerts, M.J.B., Van der Vleuten, C.P.M., & Schuwirth, L.W.T. (2002). Optimising the reproducibility of a performance-based assessment test in midwifery education. *Advances in Health Sciences Education*, *7*, 133-145.

Guion, R.M. (1998). Jumping the gun at the starting gate: When fads become trends and trends become traditions. In M.D. Hakel (Ed.), *Beyond multiple choice: Evaluating alternatives to traditional testing for selection*. Mahwah, NJ: Lawrence Erlbaum Associates, 7-15.

Guthke, J., & Beckmann, J.F. (2003). Dynamic assessment with diagnostic programs. In R.J. Sternberg, J. Lautrey, & T.I. Lubart (Eds.), *Models of Intelligence: International perspectives*. Washington, DC: American Psychological Association, 227-242.

Haertel, E.H. (2006). Reliability. In R.L. Brennan (Ed.), *Educational Measurement* (4th edition). Westport, CT: American Council on Education & Praeger Publishers, 65-110.

Haertel, E.H., & Linn, R.L. (1996). Comparability. In Gary W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (NCES-96-802). Washington, DC: National Center for Education Statistics. Retrieved August 30, 2010 from http://nces.ed.gov/pubs/96802.pdf, 59-78.

Hambleton, R.K., & Jones, R.W. (1993). An NCME instructional module on comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice. 12(3)*, 38-47.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. California: Sage Publications, Inc.

Hancock, G.R. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *Journal of Experimental Education*, *62(2)*, 143-157.

Hardy, R.A. (1995). Examining the costs of performance assessment. *Applied Measurement in Education*, *8(2)*, 121-134.

Hedlund, J., Wilt, J.M., Nebel, K.L., Ashford, S.J., & Sternberg, R.J. (2006). Assessing practical intelligence in business school admissions: A supplement to the graduate management admissions test. *Learning and Individual Differences*, *16*, 101-127.

Hogan, J., & Roberts, B.W. (1996). Issues and non-issues in the fidelity-bandwidth trade-off. *Journal of Organizational Behavior*, *17*, 627-637.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Hulin, C.L., Henry, R.A., & Noon, S.L. (1990). Adding a dimension: Time as a factor in the generalizability of predictive relationships. *Psychological Bulletin*, *107(3)*, 328-340.

Hulin, C.L., Lissak, R.I., & Drasgow, F. (1982). *Recovery of two- and three-parameter logistic item characteristic curves: A Monte-Carlo study*. Applied Psychological Measurement, 6(3), 249-260.

Hunsley, J., & Meyer, G.J. (2003). The incremental validity of psychological testing and assessment: Conceptual, methodological, and statistical issues. Psychological Assessment, 15(4), 446-455.

Hyde, J.S. (1981). How large are cognitive gender differences? A meta-analysis using ω^2 and d. *American Psychologist*, *36*, 892-901.

Hyde, J.S., & Linn, M.C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, *104*, 53-69.

John, O., Donahue, E., & Kentle, R. (1991). *The "Big Five" Inventory – Versions 4a and 54*. Technical Report, Institute of Personality Assessment and Research, Berkeley, CA: University of California, Berkeley.

Johnson, R.L., Penny, J.A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks*. New York, NY: The Guilford Press.

Jöreskog, K. G., & Sörbom, D. (1996). LISREL 8: User reference guide [computer software manual]. Chicago, IL: Scientific Software International.

Jöreskog, K.G., & Sörbom, D. (2006). LISREL 8. Scientific Software International, Inc.

Judge, T.A. & Ilies, R. (2002). Relationship of personality to performance motivation: A meta-analytic review. *Journal of Applied Psychology*, *87(4)*, 797-807.

Kaiser, F. & De Weert, E. (1994) Access-policies and Mass Higher Education: A comparative analysis of the use of policy-instruments in seven countries. In: L. Goedegebuure & F. van Vught (Eds.), *Comparative Policy Studies in Higher Education*. Utrecht: LEMMA.

Kane, M., Crooks T., & Cohen A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18*, 5-17.

Kaplan, R.M., & Sacuzzo, D.P. (2005). Psychological testing: Principles, applications, and issues (6th ed.). Belmont, CA: Thomson Wadsworth.

Klein, S.P., Kuh, G.D., Chun, M., Hamilton, L., & Shavelson, R. (2005). An approach to measuring cognitive outcomes across higher education institutions. *Research in Higher Education*, *46*, 251-276.

Kolen, M.J. (1999). Threats to score comparability with applications to performance assessments and computerized adaptive tests. *Educational Assessment*, *6(2)*, 73-96.

Kolen, M.J. (2006). Scaling and Norming. In R.L. Brennan (Ed.), *Educational Measurement* (4th edition). Westport, CT: American Council on Education & Praeger Publishers, 155-186.

Kuncel, N.R., Crede, M., & Thomas, L.L. (2007). A meta-analysis of the predictive validity of the Graduate Management Admission Test (GMAT) and undergraduate grade point average (UGPA) for graduate student academic performance. *Academy of Management Learning and Education*, *6*, 51-68.

Kuncel, N.R., Hezlett, S.A., & Ones, D.S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin*, *127*, 162-181.

Kuncel, N.R., Hezlett, S.A., & Ones, D.S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all?. *Journal of Personality and Social Psychology*, *86*, 148-161.

Landis, J.R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159-174.

Lane, S., Liu, M., Ankenmann, R.D., & Stone, C.A. (1996). Generalizability and validity of a mathematics performance assessment. *Journal of Educational Measurement*, *33(1)*, 71-92.

Lane, S., & Stone, C.A. (2006). Performance assessment. In R.L. Brennan (Ed.), *Educational Measurement* (4th edition). Westport, CT: American Council on Education & Praeger Publishers, 387-431.

Lane, S., Wang, N., & Magone, M. (1996). Gender-related differential item functioning on a middle-school mathematics performance assessment. *Educational Measurement: Issues and Practice*, *15*, 21-27.

Lievens, F., Buyse, T., & Sackett, P.R. (2005). The operational validity of a video-based situational judgment test for medical college admissions: Illustrating the importance of matching predictor and criterion construct domains. *Journal of Applied Psychology*, *90*, 442-452.

Lievens, F., & Coetsier, P. (2002). Situational tests in student selection: An examination of predictive validity, adverse impact and construct validity. *International Journal of Selection and Assessment*, *10(4)*, 245-257.

Lindblom-Ylänne, S., Lonka, K., & Leskinen, E. (1996). Selecting students for medical school: What predicts success during basic science studies? A cognitive approach. *Higher Education*, *31*, 507-527.

Lindblom-Ylänne, S., Lonka, K., & Leskinen, E. (1999). On the predictive value of entry-level skills for successful studying in medical school. *Higher Education*, *37*, 239-258.

Lounsbury, J.W., Sundstrom, E., Loveland, J.M., & Gibson, L.W. (2003). Intelligence, "Big Five" personality traits, and work drive as predictors of course grade. *Personality and Individual Differences*, *35*, 1231-1239.

Lubinski, D. (2004). Introduction to the special section on cognitive abilities: 100 years after Spearman's (1904) "'General Intelligence,' objectively determined and measured". *Journal of Personality and Social Psychology*, *86(1)*, 96-111.

Luecht, R.M. (2006). Designing tests for pass-fail decisions using item response theory. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of Test Development*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc., 575-596.

Luecht, R.M., & Miller, T.R. (1992). Unidimensional calibrations and interpretations of composite traits for multidimensional tests. *Applied Psychological Measurement*, *16(3)*, 279-293.

Maclellan, E. (2004). How convincing is alternative assessment for use in higher education. *Assessment & Evaluation in Higher Education*, *29(3)*, 311-321.

Maurer, T.J. (2005). Distinguishing cutoff from critical scores in personnel testing. *Consulting Psychology Journal: Practice and Research*, *57(2)*, 153-162.

McCloy, R.A., Campbell, J.P., & Cudeck, R. (1994). A confirmatory test of a model of performance determinants. *Journal of Applied Psychology*, *79(4)*, 493-505.

Messick, S. (1993). Test validation. In R.L. Linn (Ed.), *Educational Measurement*, (3rd edition). Washington DC: American Council on Education & National Council on Measurement in Education, 13-103.

Miller, D.M., & Linn, R.L. (2000). Validation of performance-based assessments. *Applied Psychological Measurement*, *24(4)*, 367-378.

Millsap, R.E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39(3)*, 479-515.

Mislevy, R.J. (1994). Evidence and inference in educational assessment. *Psychometrika*, *59(4)*, 439-483.

Mislevy, R.J. (1996). Test theory reconceived. *Journal of Educational Measurement*, *33(4)*, 379-416.

Muraki, E., Hombo, C.M., & Lee, Y. (2000). Equating and linking performance assessments. *Applied Psychological Measurement*, *24(4)*, 325-337.

Nedermeijer, J., De Gruijter, D., & Wijdeveld, P. (2006). *Project "Experimenten met selectie"* [A project on investigating student selection]. Leiden, the Netherlands: Leiden University, Interfacultair Centrum voor Lerarenopleiding, Onderwijsontwikkeling en Nascholing.

Nunnally, J.C. & Bernstein, I.H. (1994). *Psychometric Theory* (3rd Edition). New York: McGraw-Hill, Inc.

O'Connor, M.C. & Paunonen, S.V. (2007). Big Five personality predictors of post-secondary academic performance. *Personality and Individual Differences*, *43*, 971-990.

Oosterveld, P. & Vorst, H.C.M. (2003). *Testconstructie en testonderzoek* [Test construction and test research]. Amsterdam: Universiteit van Amsterdam.

Ostini, R., & Nering, M.L. (2006). *Polytomous item response theory models*. Thousand Oaks, California: Sage Publications, Inc.

Parshall, C.G., Kromrey, J.D., Chason, W.M., & Yi, Q. (1997). Evaluation of parameter estimation under modified IRT models and small samples. Paper presented at the Annual Meeting of the Psychometric Society, Gatlinburg, TN. (ERIC Document Reproduction Service No. ED421535) Retrieved August 4, 2011, from EBSCOhost ERIC database.

Powers, D.E. & Kaufman, J.C. (2004). Do standardized tests penalize deep-thinking, creative, or conscientious students? Some personality correlates of Graduate Record Examinations test scores. *Intelligence*, *32*, 145-153.

Premack, S.L. & Hunter, J.E. (1988). Individual unionization decisions. *Psychological Bulletin*, *103(2)*, 223-234.

R Development Core Team (2005). R: A language and environment for statistical computing, reference index version 2.13.0. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

Raykov, T., & Marcoulides, G.A. (2000). *A first course in structural equation modeling*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

Reeve, C.L., & Hakel, M.D. (2002). Asking the right questions about *g*. *Human Performance*, *15*, 47-74.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17.5*. Retrieved from http://www.jstatsoft.org/v17/i05

Robbins, S.B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin*, *130(2)*, 261-288.

Rothstein, M.G., Paunonen, S.V., Rush, J.C., & King, G.A. (1994). Personality and cognitive ability predictors of performance in graduate business school. *Journal of Educational Psychology*, *86(4)*, 516-530.

Ryan, T.G. (2006). Performance assessment: critics, criticism, and controversy. *International Journal of Testing*, *6(1)*, 97-104.

Ryan, A.M., Ployhart, R.E., Gregoras, G.J., & Schmit, M.J. (1998). Test preparation programs in selection contexts: Self selection and program effectiveness. *Personnel Psychology*, *51*, 599-621

Sackett, P.R., & Yang, H. (2000). Correction for range restriction. *Journal of Applied Psychology*, *85(1)*, 112-118.

Samejima, F. (1997). Graded response model. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory*. New York, NY: Springer, 85-100.

Schmidt, F.L. & Hunter, J.E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124(2)*, 262-274.

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8(4)*, 350-353.

Schoonheim-Klein, M., Muijtens, A., Habets, L., Manogue, M., Van der Vleuten, C., Hoogstraten, J., & Van der Velden, U. (2008). On the reliability of a dental OSCE, using SEM: effect of different days. *European Journal of Dental Education*, *12*, 131-137.

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education*, *35*, 453-472.

Shavelson, R.J., & Huang, L. (2003). Responding responsibly to the frenzy to assess learning in higher education. *Change*, *35*, 10-19.

Shavelson, R.J. & Webb, N.M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Shumate, S.R., Surles, J., Johnson, R.L., & Penny, J. (2007). The effects of the number of scale points and non-normality on the generalizability coefficient: A Monte-Carlo study. *Applied Measurement in Education*, *20(4)*, 357-376.

Sireci, S.G. (1991). "Sample-independent" item parameters? An investigation of the stability of IRT item parameters estimated from small data sets. *Paper presented at the annual meeting of the Northeastern Educational Research Association*. Ellenville, NY.

Smolkowski, K. (2004). Effects of multicollinearity on completed models. Retrieved from http://www.ori.org/~keiths/Files/Methods/Multicollinearity.html

Snow, R.E. (1994). Abilities in academic tasks. In R.J. Sternberg & R.K. Wagner (Eds.), *Mind in Context*. New York, NY: Cambridge University Press, 3-37.

Sternberg, R.J. (1999). Intelligence as developing expertise. *Contemporary Educational Psychology*, *24*, 359-375.

Tanilon, J., Segers, M., Vedder, P., & Tillema, H. (2009). Development and validation of an admission test designed to assess samples of performance on academic tasks. *Studies in Educational Evaluation*, *35*, 168-173.

Tanilon, J., Vedder, P., Segers, M. (2011). *Examining score comparability and incremental validity of a performance assessment designed for student admission.* Submitted for publication.

Tanilon, J., Vedder, P., Segers, M., Tillema, H. (2011). Incremental validity of a performance-based test over and above conventional academic predictors. *Learning and Individual Differences*, *21(2)*, 223-226.

Tanilon, J., Vedder, P., Segers, M., & Van Geel, M. (2011). *Examining relations between academic predictors in higher education: An overview using meta-analytic path analysis.* Submitted for publication.

Tate, R. (2002). Test dimensionality. In G. Tindal & T.M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy and implementation*. Mahwah, NJ: Lawrence Erlbaum Associates, 181-212.

The College Board. (1999). Toward a taxonomy of the admissions decision-making process: A public document based on the first and second college board conferences on admissions models (Document No. 215672). New York: College Board Publications.

Trapmann, S., Hell, B., Hirn, J.W., & Schuler, H. (2007). Meta-analysis of the relationship between the Big Five and academic success at university. *Zeitschrift für Psychologie*, *215(2)*, 132-151.

Tzelgov, J., & Henik, A. (1991). Suppression situations in psychological research: Definitions, implications, and applications. *Psychological Bulletin, 109,* 524–536.

Valsiner, J. & Leung, M. (1994). From intelligence to knowledge construction: a sociogenetic process approach. In R.J. Sternberg & R.K. Wagner (Eds.), *Mind in Context*. New York, NY: Cambridge University Press, 202-217.

Van der Haar, S., & van Lakerveld, J. (2004). *Instapbeoordeling HBO-WO master: Rapportage en handleiding voor de ontwikkeling* [Transition from higher vocational education to academic graduate program: Reporting and developing guidelines]. Leiden, The Netherlands: Leiden University, Platform Opleiding, Onderwijs en Organisatie.

Viswesvaran, C., & Ones, D. S. (1995). Theory testing: Combining psychometric meta-analysis and structural equation modeling. *Personnel Psychology*, *48*, 865–885.

Voyer, D., Voyer, S., & Bryden, M.P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, *117*, 250-270.

Wagerman, S.A. & Funder, D.C. (2007). Acquaintance reports of personality and academic achievement: A case of conscientiousness. *Journal of Research in Personality*, *41*, 221-229.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability. *Educational Measurement: Issues and Practice*, *15(1)*, 22-29.

Westerheijden, D.F., Cremonini, L., Kolster, R., Kottmann, A., Redder, L., Soo, M., Vossensteyn, H., & De Weert, E. (2008). *New degrees in the Netherlands: Evaluation of the Bachelor-Master structure and accreditation in Dutch higher education*. Twente, The Netherlands: University of Twente, Center for Higher Education Policy Studies.

Whitney, D.R. (1989). Educational admissions and placement. In R.L. Linn (Ed.), *Educational Measurement* (3rd edition). London: Collier MacMillan, 515-525.

Wicherts, J.M. (2007). *Group differences in intelligence test performance* (Doctoral dissertation). Retrieved from http://dare.uva.nl/document/44999?fid=44999

Wilder, G.Z., & Powell, K. (1989). *Sex differences in test performance: A survey of the literature* (College Board Report No. 89-3). New York: College Entrance Examination Board.

Witte, J., Van der Wende, M., & Huisman, J. (2008). Blurring boundaries: how the Bologna process changes the relationship between university and non-university higher education in Germany, the Netherlands and France. *Studies in Higher Education*, *33*, 217-231.

Wolming, S. (1999). Validity issues in higher education selection: A Swedish example. *Studies of Educational Evaluation*, *25*, 335-351.

Zwick, R. (2006). Higher education admissions testing. In R.L. Brennan (Ed.), *Educational Measurement* (4th edition). Westport, CT: American Council on Education & Praeger Publishers, 647-679.

Zysberg, L., Levy, A., & Zisberg, A. (2011). Emotional intelligence in applicant selection for care-related academic programs. *Journal of Psychoeducational Assessment*, *29(1)*, 27-38.

Samenvatting

Conventionele academische predictoren zoals gemiddelde cijfers in de vooropleiding en cognitieve tests zijn bewezen goede voorspellers van academische prestaties (Kuncel, Hezlett, & Ones, 2001; Kuncel, Hezlett, & Ones, 2004). Een groot deel van de variantie in academische prestaties is echter nog onverklaard (Kaplan & Sacuzzo, 2005). Dit proefschrift gaat na of een performance-based test, namelijk de Performance Samples on academic tasks in Education and Child Studies (PSEd), toegevoegde waarde heeft naast dergelijke conventionele predictoren van academische prestaties. Scores op cognitieve tests worden beschouwd als een indicatie van algemene intelligentie, een vrij stabiele psychologische eigenschap die verschilt tussen individuen en grotendeels onafhankelijk is van contextuele invloeden (Barab & Plucker, 2002; Gardner, 2003; Snow, 1994). Bij performance-based tests wordt de contextafhankelijkheid juist als uitgangspunt genomen. Scores op performance-based tests dienen in het onderhavige geval aan te geven wat de bekwaamheid van een student is in het uitvoeren van taken die kenmerkend zijn voor een te volgen opleiding.

Om de mobiliteit van studenten te vergroten zijn Europese universiteiten overgegaan op een bachelor-mastersysteem. Ondanks de toename in uniformiteit in naamgeving tussen Europese landen is de vergelijkbaarheid van de opleidingen tussen die landen nog steeds beperkt. Dit betekent dat een bachelordiploma Pedagogische Wetenschappen behaald in land A niet zomaar toegang geeft tot een masteropleiding Pedagogische Wetenschappen in land B. Binnen Nederland is er sprake van een vergelijkbaar aansluitingsprobleem dat samenhangt met het onderscheid tussen het hoger beroepsonderwijs (HBO) en het wetenschappelijk onderwijs (WO). Een bachelordiploma voor een opleiding met een sterk pedagogische component in het HBO geeft niet vanzelfsprekend toegang tot een universitaire masteropleiding Pedagogische Wetenschappen. Om de overstap vanuit een buitenlandse opleiding of vanuit een Nederlandse HBO-bachelor naar een Nederlandse universitaire masteropleiding mogelijk te maken hebben veel Nederlandse universiteiten voor bepaalde masteropleidingen zogenaamde schakelprogramma's ontwikkeld: een verkort en versneld inhaalprogramma met een omvang van een half tot een heel jaar, wat in studiepunten uitgedrukt

tussen 30 en 60 ECTS zijn. Veel schakelprogramma's zijn selectief: alleen die studenten worden toegelaten waarvan aannemelijk is dat zij het programma met succes zullen doorlopen. De opleiding Pedagogische Wetenschappen van Universiteit Leiden heeft hiertoe een toelatingsexamen ontwikkeld, de hiervoor reeds genoemde PSEd. Dit proefschrift gaat over de ontwikkeling en de validatie van dit toelatingsexamen.

In hoofdstuk 2 van dit proefschrift is een meta-analytische procedure gebruikt om relaties tussen conventionele academische predictoren en academische prestaties in kaart te brengen. Er werd een passend model gevonden dat op basis van o.a. cognitieve toetsen, cijfers in de vooropleiding en motivatie 29% van de verschillen in academische prestaties van studenten kon verklaren. Alhoewel dit model aangeeft dat een substantieel deel van de academische prestaties van studenten al verklaard kan worden, wordt in dit hoofdstuk beargumenteerd dat alternatieve predictoren, zoals performance-based tests, nodig zijn om de academische prestaties van studenten nog beter te kunnen voorspellen.

In hoofdstuk 3 zijn de dimensionaliteit, de betrouwbaarheid en de predictieve validiteit van de PSEd onderzocht. Er is gevonden dat de test unidimensioneel is, maar een relatief lage betrouwbaarheid heeft. De test blijkt een significante voorspeller van academische prestaties. Deze bevindingen worden bediscussieerd in het licht van de 'bandwith-fidelity tradeoff': Naarmate binnen een test meer heterogeen wordt gemeten zal deze test mogelijk meer variantie kunnen verklaren in het criterium, maar tegelijkertijd zal de betrouwbaarheid lager worden. De gevonden unidimensionaliteit betekent in dit geval niet zozeer dat de test sterk homogeen in materiaal is, maar wel dat er één factor (comprehension) is die antwoorden op een heterogene set items reguleert.

In hoofdstuk 4 is bekeken in hoeverre toetsscores tussen alternatieve versies van de PSEd vergelijkbaar zijn. Omdat PSEd als toelatingsexamen wordt gebruikt is het niet mogelijk bij de verschillende afnames identieke versies te gebruiken. Daarom is getracht voor elke afname een alternatieve versie te ontwikkelen, waarbij er zorg voor wordt gedragen dat de meetintentie tussen de versies vergelijkbaar is. Uit een confirmatieve factoranalyse op drie versies van de PSEd die in drie achtereenvolgende jaren zijn gebruikt bleek dat een unidimensioneel model een passende beschrijving levert van de PSEd. Dit

ondersteunt het idee dat de PSEd bij alle drie de jaargangen hetzelfde construct heeft gemeten, alhoewel regressiegewichten en errorvariantie tussen de jaargangen verschilden. Tevens bleek dat alle drie de versies van de PSEd significante voorspellers waren van latere academische prestaties. Het is dus mogelijk alternatieve versies van de PSEd te construeren die gelijk zijn in meetintentie en voorspellend zijn voor academische prestaties.

In hoofdstuk 5 is de meerwaarde van de PSEd naast traditionele academische predictoren bij het voorspellen van toekomstige academische prestaties verder onderzocht. Academische prestaties als criterium werden geoperationaliseerd als het gemiddelde cijfer voor gemaakte tentamens tijdens de opleiding. Predictieve validiteit is het belangrijkste aspect van een toelatingsexamen, en nieuwe toetsen moeten variantie verklaren in het criterium boven al bestaande toetsen (Hunsley & Meyer, 2003) Uit een hierachische regressie bleek dat de PSEd voorspellend was voor academische prestaties, ook als gecontroleerd wordt voor de cijfers uit HBO vooropleiding en het eindexamen op de middelbare school. De PSEd blijkt een toegevoegde waarde te hebben naast deze meer traditionele voorspellers van academische prestaties.

In dit proefschrift is een appendix geschreven op basis van de data in hoofdstuk 4. In dit appendix werd exploratief item respons theorie (IRT) toegepast. De voordelen van IRT zijn dat de vaardigheid van studenten onafhankelijk van de moeilijkheidsgraad van de toets geschat kan worden, dat itemparameters onafhankelijk van de steekproef geschat kunnen worden, en dat test items gekoppeld kunnen worden aan het vaardigheidsniveau van de studenten (Hambleton & Jones, 1993; Hambleton, Swaminathan, & Rogers, 1991). De data zijn geanalyseerd met het 'graded response model' (Ostini & Nering, 2006; Samejima, 1997), een IRT model dat gebruikt kan worden bij toetsen met geordende scorecategorieën, zoals de PSEd. De analyse gaf aan dat het model niet fit met de data, wat zou kunnen betekenen dat de PSEd, zoals ook in hoofdstuk 3 is beschreven, door de heterogene set items de voor IRT belangrijke aanname van unidimensionaliteit schendt. Bovendien is een steekproefomvang van 500 of meer personen wenselijk voor een robuuste analyse met IRT (Hulin, Lissak, & Drasgow, 1982; Sireci, 1991). Deze aantallen waren niet haalbaar in onderhavige studie. Derhalve kunnen aan de IRT

analyses geen sterke conclusies worden verbonden. In toekomstig onderzoek zal de PSEd verder gevalideerd moeten worden aan de hand van IRT analyses.

Naast het nut van performance-based tests als predictor van studiesucces kunnen deze tests docenten ook informeren over de beginsituatie van studenten, zodat het onderwijsprogramma van studenten daarop kan worden afgestemd en de groei in relevante leerprestaties kan worden gevolgd. Dit kan belangrijke informatie opleveren voor de verbetering van de kwaliteit van het onderwijs. Het gebruik van toelatingsexamens door universiteiten is een controversieel onderwerp in Nederland. Het belangrijkste argument tegen toelatingsexamens is dat de selectie van studenten al plaatsvindt bij de aanvang en de afsluiting van het voortgezet onderwijs. Daarnaast is er in Nederland beperkt empirisch onderzoek uitgevoerd met betrekking tot de toegevoegde waarde van academische predictoren anders dan die welke al worden gebruikt. Echter, voor studenten die hun vooropleiding niet in Nederland hebben gevolgd, of voor studenten die geen vooropleiding hebben gevolgd die rechtstreeks toegang geeft tot een vervolgopleiding, kan een toelatingsexamen gebruikt worden om te identificeren of zij vergelijkbare capaciteiten hebben als studenten met een Nederlandse opleidingsachtergrond die wel direct zijn toegelaten. Indien vooropleidinggegevens onvoldoende basis bieden voor toelating wordt het belangrijk om goede alternatieven te vinden, of goede aanvullende gegevens. Op basis van dit proefschrift blijkt dat performance-based tests een dergelijk alternatief bieden om studenten te selecteren.

Acknowledgment

Een promotietraject is net als een pelgrimstocht. De hele dag wandelen om je bestemming te bereiken, met blaren op de voeten maar mooie uitzichten om je heen is zeker de moeite waard. Net als wekenlang werken aan een artikel, dat vervolgens geaccepteerd wordt voor publicatie. Tijdens een pelgrimstocht zijn er mensen die de tocht bijzonder maken, zoals een oude man in een plaatselijke dorpskantine die een kopje warme thee en een glas ijs serveerde toen ik vroeg om té con hielo (mijn Spaanse vertaling van ijsthee). Tijdens mijn promotietraject zijn er ook mensen die het traject prettig hebben gemaakt en die mensen wil ik bedanken.

Het Onderwijsbureau Pedagogiek, voor hun medewerking tijdens de dataverzameling.

Studentassistenten, voor hun inzet tijdens de dataverwerking.

Vrienden voor de steun die ze aan mij hebben gegeven. (Friends for being supportive, and for staying in touch, no matter where in the world they are.)

Familie en schoonfamilie die er voor zorgen dat ik met beide benen op de grond sta. (My family for keeping me grounded.)

Mijn man, voor zijn waardevolle suggesties en commentaren. Een lieve echtgenoot en een voorbeeldige vader. Iedere dag met jou is een wonder. Mahal na mahal kita.

To my parents, for instilling in me the value of education.

Curriculum Vitae

After completing secondary education at the Sacred Heart Academy and obtaining her Bachelor's degree in Psychology at the Colegio de San Juan de Letran in the Philippines, Jenny Tanilon went on to get her Master's degree in Developmental Psychology at the University of Amsterdam in the Netherlands. Her thesis was on task-switching as explained by the cognitive theories of inhibition and binding. Having been exposed to fundamental research during her Master's studies, she opted to conduct applied research as a PhD. The project which she worked on at Leiden University was on the development and validation of an admission test that draws on research regarding abilities from an individual-context interaction perspective. Her research interests include cognitive measures, cognitive development, and meta-analytic methods.